The Value of Privacy: Evidence from Online Borrowers

Huan Tang *

December 2019

[click for the updated version]

Abstract

This paper studies the value of privacy, for individuals, using data from large-scale field experiments that vary disclosure requirements for loan applicants and loan terms on an online peer-to-peer lending platform in China. I find that loan applicants attach positive value to personal data: Lower disclosure requirements significantly increase the rate at which applications are completed. I quantify the monetary value of personal data—and the welfare effect of various disclosure policies—by developing a structural model that links individuals' disclosure, borrowing, and repayment decisions. Using detailed application-level data, I estimate that social network ID and employer contact are valued at 230 RMB (i.e., $33, or 70% of the average daily salary in China); for successful borrowers, this accounts for 8% of the average net present value of a loan. Requiring answers to these application questions reduces borrower welfare by 13% and costs the platform $0.50 in expected revenue per applicant.

JEL Classification: D14, D18, G23

*HEC Paris, 1 rue de la libération, 78350 Jouy-en-Josas, France; email: huan.tang@hec.edu. I am grateful to my committee members Johan Hombert (advisor), Jean-Edouard Colliard, Denis Gromb, and Boris Vallée. For useful comments I also thank Will Cong (discussant) as well as participants at FIRS 2019, HEC Paris PhD workshop, and HEC Paris brownbag. This research received financial support from the GREGHEC (Groupement de Recherche et d'Etudes en Gestion à HEC Paris) and the French National Research Agency (l'Agence Nationale de la Recherche) through Project F-STAR (ANR-17-CE26-0007-01). All errors are my own.

1 Introduction

Personal data is an essential element of business in the digital era. Online and mobile technologies have revolutionized the collection, storage, and processing of huge amounts of data through the use of smart forms, digital footprints, cookies, and so forth. Collecting personal data allows businesses to provide tailored services and improve customer experience, but it also poses the risk of privacy intrusion and data breach—of which the recent Cambridge Analytica scandal is but one alarming example.[1] As a result, digitization has been associated with increased privacy concerns on the part of consumers (Goldfarb and Tucker, 2012) and policy makers.

The regulatory framework governing how personal data are gathered and used has evolved in response to those developments. In 2018, both the European Union and California passed major privacy protection regulations: the EU General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), each of which demanded significantly more transparency and accountability in the processing of personal data. At the core of those reforms is the notion that privacy is valuable. Indeed, these regulations improve privacy protection at the expense of significant compliance costs for businesses.[2] Statistical agencies similarly face the dual mandate to publish accurate statistics while protecting respondent privacy (Abowd and Schmutte, 2019). Since customers have become increasingly aware of how their data are used, businesses may also need to assess the value of privacy before devising their data elicitation incentives. But do consumers actually value privacy? And, if they do, how much and why? Despite the importance of this issue, data limitations have prevented previous studies from offering much guidance.

In this paper, I investigate whether and how much online borrowers value the privacy

[1] In 2018, the personal data of up to 87 million Facebook users were reported to have been compromised and shared with a third-party political consulting firm that harvested the data to provide targeted political or commercial content. Shortly after reports of the incident, Facebook CEO Mark Zuckerberg agreed to testify before Congress about the scandal. The Senate Judiciary Committee also sent requests to the CEOs of Alphabet Inc. (Google's parent company) and Twitter Inc. for the hearing.

[2] According to a survey conducted for the audit, tax, and consulting network RSM, 92% of businesses in Europe are not prepared for GDPR and many are cutting back in other areas—including product innovation (23%) and international expansion (22%)—in response to increasing compliance costs. See https://www.prnewswire.com/news-releases/92-of-european-businesses-are-unprepared-for-gdpr-658099083.html for details.

of personal data. For this purpose I use large-scale field experiments that vary disclosure requirements and loan terms for loan applicants on a Chinese peer-to-peer (P2P) lending platform. I first provide reduced-form evidence that individuals attach positive value to privacy. Then I further quantify the value of privacy by using application-level data to build and estimate a structural model for individuals' disclosure, borrowing, and repayment decisions. This model also allows for an assessment of borrower welfare and platform profit under different disclosure policies.

The data come from two randomized controlled trials (RCTs) conducted on a major Chinese P2P lending platform.[3] Chinese online lending markets provide an ideal real-life context for studying the value of privacy. One reason is that there is no official credit score system and there are limited resources available to verify borrower credentials and documents, so Chinese P2P platforms commonly ask for nonstandard personal information from borrowers—for example, applicants' phone contact list and online shopping account credentials.[4] This setting is appropriate also because the disclosure of personal data is a pre-condition for loans, which means that online borrowers face a trade-off between privacy and access to credit. I exploit this trade-off to gauge the monetary value of privacy.

Participants in the "disclosure" RCT are individuals who began filling out fresh online applications. Upon arrival on the platform, they are presented with a two-page application questionnaire that includes about 20 questions. Applicants observe loan terms *after* they complete the questionnaire. The randomized element is the set of questions. Loan applicants in the control group are required to answer all questions, whereas other applicants are divided into four treatment groups. In the first three treatment groups, one of the following items is removed: social network ID (QQ ID, similar to a WhatsApp ID), marital status, or employer contact (landline phone number). In the fourth treatment group, all

[3]These two RCTs were designed and implemented by the platform in the first quarter of 2018, independently of this study. I collected the data from the platform following completion of the RCTs.

[4]There are two reasons why P2P platforms collect nonstandard information. First, unlike brick-and-mortar banks, online P2P platforms have limited resources to collect and verify personal documents such as proof of income and employment; instead they rely on digital information, which can be collected via an automated process, to screen borrowers. Second, using such alternative data gives online platforms a technological advantage when underwriting loans to traditionally underserved borrowers. According to Brusa et al. (2019), online lending services greatly improve access to credit for low-income households and small entrepreneurs.

of these three items are removed. Overall, the four treatment groups include 216,881 applicants; the control group includes 53,507 applicants. It is critical that, when applicants do not complete the entire application process, I can observe the stage at which they stop. Hence I am able to evaluate the causal impact of different disclosure requirements on the completion of applications.

From this disclosure RCT, I establish that lower disclosure requirements increase application completion rates. In particular: removing the social network ID and employer contact questions increases completion rates (which otherwise average 31 percentage points) by 3.4% and 2.0%, respectively. In contrast, removing the question on marital status has no effect on completion rates.

There may be several reasons why online borrowers are reluctant to disclose personal data. First, applicants may have incentives to hide information that reveals lower creditworthiness. A prediction based on this mechanism would be that applicants in the control group who choose disclosure—when completing questionnaires that include additional items—are less risky. However, I find no discernible differences in ex ante (e.g., loan grade) or ex post (e.g., delinquency) risk measures across the five groups. This outcome rules out an adverse selection channel.[5]

Second, individuals may be less likely to disclose personal data because of the effort required to fill out the questionnaire. I argue against two versions of this channel. First, I rule out that exertion leads to lower completion rates by using that the "social network ID" question appears on page 1 of the questionnaire. Relative to the treatment group for which only this question was removed, the control group exerts more effort to complete page 1. So conditional on reaching page 2, this channel would predict lower page 2 completion rates for the control group. Yet I find the opposite result—namely, higher page 2 completion rates for the control group—which is more consistent with the privacy channel than with an "effort" channel. Indeed, the additional social network ID question on page 1 should have selected control group applicants with lower sensitivity to privacy, who are more likely to complete page 2.

I also rule out that applicants have difficulty retrieving information. To test this hypo-

[5]This test follows the approach in Einav et al. (2010).

thesis, I split the sample by employment status. Currently unemployed applicants may have more difficulty retrieving their last "employer contact", so the inclusion of that item should lower completion rates even more among such applicants. Instead, I find the effect of this question to be entirely concentrated among employed applicants. This result is consistent with the privacy channel, since the privacy content of an employer contact is arguably higher for applicants who are currently employed.

It may be that individuals have an intrinsic preference for privacy. Because preferences usually relate directly to individual demographic characteristics, we should observe heterogeneity in the treatment effects. Consistent with this mechanism and with Goldfarb and Tucker (2012), who show that women and older individuals are more concerned with privacy issues, I find that the effect of removing the social network ID item on the completion rate is greater for women and for applicants above age 50.

Having established that applicants indeed value the privacy of personal data, I turn to assessing the monetary value of that privacy. Doing so requires gauging how much a forgone loan is worth to applicants. Toward that end, I exploit the "loan" RCT—implemented independently of the disclosure RCT—that randomizes loan size and origination fees for 48,375 individuals who completed the application process.[6] The control group consists of 34,366 applicants who receive regular loans that average of 3,770 RMB ($540; all reported dollar values are USD) with a 29% origination fee and a 11% annual interest rate. Two treatment groups (consisting of 7,032 and 6,977 applicants, respectively) were offered loans twice as large as regular loans, and the second treatment group also received a fee reduction averaging 900 RMB. Observing whether applicants accept or reject their loan offer allows me to assess their valuation of the loan.

It is hardly surprising that borrowers prefer larger loans and lower origination fees. In the first treatment group, doubling the loan size increases take-up rates (the probability of accepting the loan offer) by 6.5 percentage points, or 11% relative to the control group's take-up rate of 57.6%. In the second treatment group, a 900 RMB ($135) fee reduction increases take-up rates by an additional 5.5 percentage points.

[6]Participants in the loan RCT must answer all items in the questionnaire, including social network ID, marital status, and employer contact.

One can obtain a rough estimate of the monetary value of privacy of personal data using a back-of-the-envelope calculation that combines the marginal effects from the two RCTs. Removing the social network ID and employer contact questions leads to a 1.3–percentage point increase in the completion rate, and reducing the origination fee by 900 RMB results in a 5.4–percentage point increase in the take-up rate. Thus employer contact and QQ combined are worth 211 RMB (i.e., $(900/5.4)\times 1.3$); this amounts to \$30 in nominal value, or \$58 when adjusted for purchasing power parity (PPP). The amount is not trivial: it is equal to 60% of the average daily salary (\$48) in China in 2018.[7]

Yet this back-of-the-envelope calculation is subject to several limitations. First, selection may occur as applicants go through the application process. For example, individuals who agree to disclose information may derive higher utility from the loan than those who do not, which implies different price sensitivities for applicants reaching different stages of the application. Second, the calculation does not allow for assessing the effect of counterfactual disclosure policies on borrower welfare and firm profit.

To overcome these limitations, I build and estimate a structural model of consumer demand that incorporates individuals' disclosure, borrowing, and repayment choices. The model starts with a standard consumer theory framework in which heterogeneous consumers choose whether or not to disclose personal data based on their expectations about loan terms. Because applicants do not observe actual loan terms when making the disclosure decisions, I assume rational expectations: applicants predict loan size and fees up to a zero-mean error. Applicants who opt to disclose then observe the actual loan terms and decide whether to accept them. Those who accept receive the loan amount, and at maturity they choose either to repay or to default. I derive a set of linear estimating equations that capture disclosure, borrowing, and repayment decisions; I then estimate the model using the simulated maximum likelihood method. I use the *disclosure* RCT to estimate parameters associated with the disclosure decision and use *both* RCTs to estimate parameters associated with the borrowing and repayment decisions.

I first recover individuals' preference parameters for disclosure requirements. I estim-

[7] According to the National Bureau of Statistics of China, the average monthly salary was 6,872 RMB (\$982) in 2018.

ate that requiring an applicant's social network ID on page 1 has the same effect on that page's completion probability as does reducing the expected loan size by 145 RMB (keeping repayment fixed). Similarly, the effect of the employer contact question on page 2 is equivalent to a 85-RMB decrease in expected loan size. The combined value, 230 RMB, is 9% larger than the back-of-the-envelope calculation. The reason is that the structural model accounts for selection: compared to the initial pool, applicants reaching page 2 are those who derive more utility from the loan and are thus less reluctant to disclose personal data. Without accounting for this effect, the reduced-form analysis underestimates the value of the employer contact item on page 2.

In addition, I find that applicants are much more sensitive to loan size than to repayment amount; this finding indicates that they are liquidity constrained. For the initial pool of applicants, I estimate that a 1,000-RMB increase in the loan amount increases the completion probability of page 1 by 20.3% whereas a 1,000-RMB reduction in total repayment, spread over 12 months, increases that completion probability by only 13.8%. These numbers implies an annual discount factor of 0.48, which is consistent with magnitudes reported in the literature.[8] Applicants proceeding to page 2 of the questionnaire and the borrowing stage have similar discount rates.

One way to interpret the value of the privacy of personal data is to compare it to the value of loans. Using the estimated sensitivities to loan amount and repayments, I calculate that successful applicants value the loans at 2,939 RMB, on average. In comparison, the combined value of social network ID and employer contact is 230 RMB, or 7.8% of the average loan value for such borrowers.

Using the demand estimates, I compute borrower utility in monetary terms for the five groups in the disclosure RCT. I find that removing the social network ID and employer contact questions increases borrower welfare by 13.4% in total. This welfare improvement comes from two margins: 7.5% is from the intensive margin (i.e., borrowers enjoy higher utility on average because of lower disclosure requirements); 5.9% comes from the extensive margin (i.e., borrowers who would otherwise not have completed the questionnaire

[8] Adams et al. (2009) report an annual discount factor of 0.21 for subprime borrowers in the auto loan market. The average repayment ratio in my data is 0.87, which is comparable to a ratio of 0.41 in their data.

now proceed to the borrowing stage).

In the last part of this paper, I evaluate platform profit under different disclosure requirements and quantify the cost of excessive data collection. This evaluation is closely linked to individuals' valuations for privacy because the latter affect their willingness to borrow and therefore also affect loan origination volume and platform profit. For this purpose, I consider a counterfactual case where a full set of information is collected and then compute the revenues—as a function of loan demand and delinquency rate—under the current and counterfactual disclosure requirements. Comparing the two cases, I find that collecting additional information on social network ID and employer contact reduces expected per-applicant revenue by 10% (from \$4.88 to \$4.40). I find also that the counterfactual disclosure requirement has little effect on the cost of lending. It follows that relaxing disclosure requirements would benefit not only borrowers but also the platform and investors.

Related literature This study contributes to three strands of research, of which the first concerns the value of privacy. A few studies in the law and marketing literature have investigated this question in the context of Internet-based marketing and offline transactions (Tsai et al., 2011; Jentzsch et al., 2012; Preibusch et al., 2013).[9] Acquisti et al. (2013) discuss the difference between the willingness to pay for a more privacy-protective offer and the willingness to accept a less privacy-protective offer. Whereas extant research is mostly survey based or relies on relatively small samples and lab experiments, I exploit two large field experiments that involve 315,785 participants on an online lending platform. Online lending is an especially suitable context to study privacy because individuals face real-world privacy–service trade-offs (and not merely hypothetical choices). This difference may account for the small or zero value for privacy reported by some previous studies.

My paper also complements the burgeoning FinTech (financial technology) literature. Previous work has focused primarily on the benefits—such as expanded access to credit (Tang, 2019), improved information efficiency (Cong and Xiao, 2019; Vallée and Zeng,

[9]For a thorough review of this literature, see Acquisti et al. (2016).

2019), and faster services (Fuster et al., 2019)—of FinTech innovations. Only a few studies have investigated the challenges caused by those innovations; examples include price discrimination (Fuster et al., 2018; Bartlett et al., 2019) and the increased number of illegal activities facilitated by bitcoin transactions (Foley et al., 2019). I complement this literature by studying a particular cost of financial innovation: privacy intrusion. This issue, which is especially severe in developing economies, is driven by rising smartphone penetration, rapid growth in e-commerce, increasing social media penetration, and the emergence of "Internet of things" devices. In the context of FinTech lending, the risk of privacy intrusion is exacerbated by the the lack of a unified official credit system as well as a low coverage of bank. To the best of my knowledge, this paper provides the first empirical evidence on the cost, to both consumers and to FinTech platforms, of *excessive* data collection.

Finally, I contribute to a new literature on privacy regulation. Among theoretical work, Acemoglu et al. (2019) and Bergemann et al. (2019) argue that one reason why privacy regulation may improve social surplus is because of externalities in data sharing. Ichihashi (2019) shows that a stricter privacy policy does not necessarily improve consumer welfare. Previous empirical work assessing the impact of such regulation has focused on the real costs of privacy regulation and the inhibition of data flows (Miller and Tucker, 2009; Campbell et al., 2015). In contrast, this paper provides a setting that can be generalized to quantify the welfare effects of privacy protection regulation.

The rest of the this study proceeds as follows. Section 2 describes the institutional setting, the experimental design, and the data, after which Section 3 presents the reduced-form evidence. I develop a structural model in Section 4 and report the estimation results in Section 5. The platform's profit maximization problem is presented in Section 6. I conclude in Section 7 with a brief summary and suggestions for future research.

2 Empirical Setting and Data

2.1 Institutional background

Chinese P2P platforms function similarly to their US P2P counterparts. Potential borrowers must create an account and provide personal information to qualify for a loan; the

information is then analyzed and an internal credit score is assigned.

However, a major difference between US and Chinese P2P platforms is the input for credit scoring models. In the United States, a significant part of the information about borrowers is extracted directly from the Fair Isaac Corporation (FICO). However, there are no official credit scores in China. So in order to assess credit risk, Chinese platforms rely on the personal data provided by applicants and data from other sources.

Besides the standard information required by traditional banks—such as national ID, proof of employment, and income—the platform collects a considerable amount of non-standard information, some of which is easily accessible only through Internet-based interactions. Examples are borrowers' QQ ID (QQ is the second-most popular instant message software in China, after WeChat), mobile contact list, phone call records, and mobile device type as well as access to the applicant's Taobao account (the largest online commercial platform in China). From the perspective of online applicants, this information is all arguably private and personal.[10]

Given all the collected information, the platform assigns to borrowers an internal loan grade (i.e., credit score). That score, combined with the fund supply, determines whether the borrower qualifies for a loan and also affects the interest rate, the maximum loan amount, and the loan's maturity. Although the payments of interest and principal go to the investors on the platforms, each platform itself charges borrowers an origination fee. The platform in this study offers installment loans: borrowers receive the loan amount in full at the beginning and then pay back interest and the principal at a monthly frequency.

After loan listings are posted, potential investors can invest in the loans in several ways. First, they can choose to invest in individual loans with a minimum contribution of 50 RMB ($9). The platform also offers quantitative tools enabling investors to automate their decisions as well as financial products with pre-set investment horizon and fixed returns. When using the latter, investors pay a management fee and simply delegate investment decisions to the platform. A listing is funded if and only if bids cover the amount requested.

[10]Both QQ and WeChat are owned by the same company (Tecent). WeChat was developed primarily for mobile users and QQ for desktop users. Each software boasts a penetration rate exceeding 70%. Also, QQ and Taobao have (respectively) the second- and third-most number of mobile users.

Similarly to other Chinese P2P platforms, the platform in this study employs protection schemes to limit investors' losses. One common practice is requiring that borrowers make contributions to a "quality assurance fund"—that is, in addition to the basic origination fees and payments of loan principal and interest—where the contribution amount depends on their credit score. In case of delinquency or default, this fund pays the delinquent installments of principal to investors. Because the platform is the residual claimant on the quality assurance fund, it has an incentive to assess borrower creditworthiness carefully and to minimize default risk.

Thus the origination fee consists of two parts. The first part is a basic origination fee, which borrowers pay directly to the platform over the first 1–3 months of the loan; the second part, the quality assurance contribution, is paid monthly until maturity. Both fees depend on loan grades. The quality assurance fee is increasing in borrower risk, but the origination fee is non-monotonic in risk because of a 36% regulatory cap on the annualized cost of borrowing.[11]

The platform

I use data from a major Chinese P2P lending platform. This platform provides small and short-term installment loans with the following features. First, borrowers are offered a credit limit; in practice, however, almost all borrowers take the full amount. Second, loan maturity ranges from 3 months to 24 months. That said, more than 90% of the loans in my sample have a 12-month maturity. Finally, most loans reflect the 36% maximum cost of borrowing.[12] The average loan on the platform is for 3,770 RMB ($540), has a 12-month maturity, and features a 36% cost of borrowing. Despite this high borrowing cost, the delinquency rate (i.e., the delinquent amount divided by the loan principal) is only about 15%.

To assess whether the platform is representative of the overall P2P industry in China, Table B1 (in Appendix B) compares the characteristics of loans originated on the platform, loans in the whole P2P industry, and credit card loans. There is no evident differ-

[11] The cost of borrowing is defined as $\left(\frac{\text{Total repayment}}{\text{Principal}} - 1\right) \cdot \frac{12}{\text{Maturity}}$. The regulatory cap on the cost of borrowing restricts the basic origination fees that platforms can charge, which are the highest for medium-risk borrowers.

[12] Since Chinese P2P platforms mostly target borrowers who are underserved by banks (Liao et al., 2017), the annualized cost of borrowing—including the interest rate and fees—is about 36%.

ence between loans offered on this platform and in the P2P industry—except for loan size. The average loan size on our focal platform (3,770 RMB) is much lower than the industry average of 18,000 RMB. But overall, the platform offers loans comparable to those offered by other platforms. In comparison with credit card loans, loans on this platform and in the P2P overall industry are significantly smaller, costlier, and riskier.

2.2 The RCTs

I use data from two randomized controlled trials, the *disclosure RCT* and the *loan RCT*, both of which were conducted on the focal platform.

2.2.1 Disclosure RCT

The platform designed the disclosure RCT to investigate applicants' sensitivity to the set of information that it collects; this trial took place from April through June 2018. Over that three-month period, a small fraction of applicants who visited the platform were randomly selected into the RCT on a daily basis. In total, the disclosure RCT involved 270,388 applicants.

These applicants are randomly divided into five groups: one control group and four treatment groups. All applicants are given a two-page application questionnaire, which includes some 20 questions. The randomized element is the particular set of questions asked. Applicants in the control group are required to answer all questions; other applicants are divided into four treatment groups. For three of these treatment groups, one of three items—the social network ID (QQ ID), marital status, or employer contact (the employer's landline number)—is removed from the questionnaire. For the fourth treatment group, all three items are removed. Hereafter, the four treatment groups are accordingly labeled "no QQ", "no marriage", "no landline", and "delete all":

1. "no QQ" group—applicants are not required to disclose their QQ number;
2. "no marriage" group—applicants are not required to disclose their marital status;
3. "no landline" group—applicants are not required to disclose their employer's landline number;

4. "delete all" group—all three of these questions are removed;

5. control group—applicants are required to answer all questions.

These questions follow the order illustrated in Figure 1, where "Page 1" and "Page 2" refer to the pages of the questionnaire. Social network ID and martial status questions appear on the first page, and employer contact appears on the second. The information required of applicants on page 1 is (in this order) mobile phone number, national ID, social network ID, education level, and marriage status. Page 2 asks for the applicant's profession, employer name, employer contact, and two emergency contacts.

[[INSERT Figure 1 about Here]]

2.2.2 Loan RCT

The second RCT took place from January 2018 through March 2018 independently of the disclosure RCT. It was designed to examine borrowers' sensitivity to origination fees and loan size. Among all qualified borrowers, the platform randomly selected 48,375 borrowers into this trial.

Selected borrowers are randomly divided into three groups: one control group (which accounts for $5/7$ of the loan RCT sample) and two treatment groups, labeled "hi-credit" ($1/7$ of the sample) and "extreme" (another $1/7$). Borrowers in the control group are offered regular loans, while the two other groups receive the following treatment (see Figure 2):

1. "hi-credit" group—applicants receive an offer whose loan size is twice large as a regular loan;

2. "extreme" group—applicants receive an offer whose loan size is twice large as a regular loan *and* the total origination fee is reduced by half.

[[INSERT Figure 2 about Here]]

The data collected from the two RCTs contain three sets of information. The first is applicants' choices and the timestamps of each decision. Those choices include whether

an applicant completes each page of the questionnaire, whether she takes up the loan after observing its terms, and whether the repayment is made in each month. The second type of information involves the internal loan grade and loan characteristics: loan size, repayment amount, and maturity. Third, there is also information on such borrower demographics as age, gender, location, and brand of mobile devices.

2.2.3 Summary statistics

To examine the quality of randomization, in Figure 3 I plot applicant characteristics for the disclosure RCT's five groups . The distribution of gender, age, mobile device, and education is nearly identical across groups.

[[INSERT Figure 3 about Here]]

We can also use these summary statistics to characterize P2P applicants more generally. First, the majority of them are young males. Panel (a) of the figure shows that 78.6% of applicants are male, which is in line with the industry-level gender ratio (78%). Panel (b) shows that the applicant base is dominated by individuals younger than 30 (49%), followed by those aged between 30 and 39 (30%) and those between 40 and 49 (17%). Only 5% of applicants are more than 50 years old. Although the ratio of young to mid-aged (below 40) borrowers is similar to that in the US P2P market, most US borrowers are between the ages of 30 and 39 (rather than below age 30). Finally, applicants use more than 700 varieties of mobile phone models, with 79% of them using one of four brands: Apple, Vivo, Oppo, and Huawei. In the empirical analysis, I use mobile phone brand as a proxy for income (Bertrand and Kamenica, 2018; Berg et al., 2019).[13] Individuals using Apple products presumably earn more than do those who use the other brands, which target consumers in the mid- to low-tier segments.

Figure 4 illustrates the geographical distribution of applicants. The 31 provinces in mainland China are sorted by the quintile of the penetration of P2P lending rates, defined as the number of applications per 100 Internet users. Besides Guangdong province, which

[13] According to Counterpoint Research Market Monitor (a market research firm), Huawei, Oppo, Vivo, Xiaomi, and Apple were the top five smartphone brands in China in 2017; their market shares were (respectively) 19%, 18%, 17%, 12%, and 11%.

hosts several tech companies, the three other provinces in the highest quintile (Yunnan, Guizhou, and Gansu) are less developed areas in the southwest and northwest of China. This finding is consistent with P2P platforms targeting areas with low banking penetration.

[[INSERT Figure 4 about Here]]

3 Reduced-form Analysis

In this section, I present the reduced-form evidence from the disclosure and loan RCTs. See Figure 5 for a schematic illustration of the applicant decision process. The disclosure RCT allows me to identify the causal effect of different disclosure requirements on the completion rates of applicants, and the loan RCT explores the causal effect of loan size and fees on the probability of accepting the loan offer. The disclosure and take-up decisions reveal how individuals value personal data and loans, respectively. Using these quantities, I provide a back-of-the-envelope calculation for the monetary value of privacy.

3.1 Disclosure RCT

Here I report the empirical results from the disclosure RCT, which establishes that applicants indeed attach positive value to the privacy of personal data.

I first show how applicants respond to different disclosure requirements and describe the heterogeneous effects in terms of demographic characteristics. Then I test for whether the treatment effects could be driven by something other than privacy concerns—for instance, an incentive to hide negative information, the effort required to complete an application, and applicants' difficulty retrieving required information.

[[INSERT Figure 5 about Here]]

3.1.1 Do applicants value privacy?

If applicants value privacy, then we would expect those facing stricter disclosure requirements to be less likely to complete the application process. Table 1 reports applicant completion rates for the five groups in the disclosure RCT. In the upper part ("a.") of this

table, the dependent variable is an indicator for whether or not the applicant completed the page 1 questions; this panel reports the treatment effects of removing the social network ID item and the marriage question. In the lower part ("b."), the dependent variable is an indicator for the completion of page 2 and the reported values are the treatment effect of omitting the employer landline question.

[[INSERT Table 1 about Here]]

According to the first row of Panel (a), removing the social network ID (QQ) item raises the page 1 completion rate by 2.97 percentage points, which is equivalent to 6% of the base completion rate in the control group (48.60

The first row in Panel (b) of Table 1 gives the treatment effect of removing the employer contact (landline number) question on the completion rates for page 1 and page 2. The point estimate is 0.61 percentage points, which represents 2.0% of the base completion rate in the control group (30.76%) and is significant at the 1% level. The second row shows that removing all three items (QQ, marriage, and landline) increases the overall completion rate from 30.76% to 32.04%. This 1.28–percentage-point difference amounts to 4.2% of the base completion rate in the control group.

So far, I have established two results from the disclosure RCT. First, individuals assign positive value to privacy—that is, because looser disclosure requirements lead to higher completion rates. Second, the marginal value of the privacy of personal data seems to be constant: the treatment effect of the QQ and marriage questions combined is equal to the sum of the two individual treatment effects. In Appendix B, Figures A1 and A2 provide additional evidence that the marginal effects of the QQ and marriage items do not depend on the questionnaire's particular set of questions.

3.1.2 Why do individuals value privacy?

There may be several reasons why online borrowers are reluctant to disclose personal data. In this section, I explore the motivation for privacy concerns.

Intrinsic value of privacy First, individuals may have an intrinsic preference for privacy, which often depends on their demographic characteristics. For example, Goldfarb

and Tucker (2012) show that women and older people are less likely (than are men and younger people, respectively) to reveal personal information when participating in an online marketing research survey. I therefore investigate the heterogeneity in applicants' sensitivity to disclosure requirements and provides evidence for this explanation.

Table 2 reports the differential treatment effects of the social network ID and employer contact questions by gender, age, income, and education level. Columns [1]–[4] (resp., columns [5]–[8]) focus on the treatment effects of the QQ (resp., employer contact) question.

[[INSERT Table 2 about Here]]

The coefficients for "female" in columns [1] and [4] reveal that women are, on average, more likely to complete the application process—perhaps because they are typically more diligent. In Table B2 I show that, indeed, female applicants complete the application 10% faster than do their male counterparts. Yet even more interesting is that the coefficient for the interaction term ("female" × "treated") is positive and implies that the effect of removing the QQ question is twice is great for women as it is for men. Removing the employer contact question does not affect men and women differentially.

Column [2] reports the coefficients for the "old" dummy variable (set to 1 only if the applicant is older than 30, the sample's median age) and for its interaction with the treatment dummy. Compared to applicants in their 20s, those above 30 are less likely to complete the application process. More importantly, the treatment effect of removing the QQ question is due entirely to applicants who are more than 30 years old. There is @>optional...similarly no differential effect of removing the employer contact question on completion rates across different age groups.

Finally, I sort applicants by income (proxied by the brand of their mobile devices) and education level. I find that applicants with iOS mobile devices and bachelor's (or higher) degrees are more likely to complete the applications, but there is no differential treatment effect across income or education groups.

Incentive to hide information Applicants may also have an incentive to hide information that indicates lower creditworthiness. One major friction in microloan markets is

information asymmetry. FinTech platforms enjoy advantages relative to traditional banks because they have the infrastructures to collect and process extensive information, including the digital footprints of individual applicants. This advantage may lead to adverse selection with respect to the disclosure requirements. Thus riskier borrowers may be more reluctant to disclose information that either reveals their true creditworthiness or facilitates the platform?s punishment in the case of default. If so, then applicants who completed their applications in the four treatment groups (i.e., the groups with lower disclosure requirements) should be riskier.

To test this hypothesis, I compare various risk measures across the five groups; see Table 3 for the results. In columns [1] and [2], I use two ex ante risk measures as dependent variables: the risk score assigned by the platform and an indicator for whether the applicant qualifies for a loan. The risk score takes values from 1 to 5, where 1 is the best grade. Also, in columns [3] and [4] I compare the loan amount and total cost of loans across groups. In column [5], I use the *fraction of payments*—calculated as the amount of payments made divided by the total amount owed—as the ex post measure for risk. If laxer disclosure requirements do lead to adverse selection, then we should expect borrowers in the "no QQ", "no landline" and "delete all" groups to be riskier. However, this hypothesis is not supported by the data. Among the 74,762 applicants who completed the application process, those in the control group have an average loan grade of 4.1 (according to the "Constant" row of column [1]) and there is no significant difference across the five groups. Similar results are obtained with respect to pre-approval. Column [2] shows that the average pre-approval rate is 43%. There is no difference across groups, except that the "no QQ" group has a 1% lower approval rate. Still, the "no QQ" group is not significantly different from other groups in terms of any other risk measure. Conditional on pre-approval, the loan amount and costs are also the same. Using the ex post risk measure yields strongly similar results. According to column [5], the five groups exhibit comparable repayment ratios: an average of 85%.

[[INSERT Table 3 about Here]]

Taken together, the absence of differences in the risk profiles of applicants in the five

groups suggests that the observed negative response to disclosure requirements is *not* explained by adverse selection.

Exertion and memory Third, individuals may be reluctant to provide personal data because of the effort required to fill in the questionnaire. There are two versions of this channel. First, I rule out that longer questionnaires cause fatigue and so lead to lower completion rates. For that purpose, I compare the completion rate of page 2—conditional on page 1 being completed—between the group without the QQ question on page 1 (the "no QQ" group) and the control group. If exertion is driving the difference in the completion rates then applicants in the control group should exhibit a lower page 2 completion rate than do those in the "no QQ" group, since the former applicants answered more questions on page 1. Yet if a preference for privacy is driving the difference then the opposite should hold, because applicants in the control group who completed page 1 should be less sensitive about privacy than those in the "no QQ" group (who have not been required to disclose as much information). The results are reported in Table 4(a), which shows that the control group's conditional completion rate of page 2 is 1.52 percentage points higher than that of the "no QQ" group.

[[INSERT Table 4 about Here]]

I also establish that the results are not explained by borrowers not having QQ accounts. The reason is that, if this were the only mechanism at play, then there would be no differences in the conditional completion rate of page 2—that is, unless borrowers who don't have QQ accounts also care more about privacy.

Second, I rule out that applicants have difficulties retrieving information, including employer contact information. In order to test this channel, I group applicants by their employment status. When an applicant is unemployed, information on her previous employer is required. If the memory channel is the true one then removal of the employer contact question should have a greater effect on unemployed applicants, who may well have more difficulty retrieving this information. Yet if the treatment effect is driven by privacy concerns then the opposite should be true, since the privacy content of a current

employer contact (for employed applicants) is presumably higher than is that of an unemployed applicant's previous employer. Results in Table 4(b) support the privacy channel. Among a total of 42,339 borrowers, for which I observe employment status, who are in either the "no landline" group or the control group, 51% are employed. I find that the treatment effect is entirely concentrated among employed borrowers. This result is inconsistent with the exertion channel , which would entail a significant effect for both groups and a larger effect for the unemployed applicants.

3.2 Loan RCT

Having established that applicants indeed value the privacy of personal data, I turn now to assessing the monetary value of privacy. Doing so requires gauging how much a forgone loan is worth to applicants. The loan RCT serves this purpose. Here I present the empirical evidence that individuals value larger loans and lower fees. Moreover, delinquency rates increase (resp., decline) with larger loan sizes (resp., lower fees).

3.2.1 How do loan size and fees affect the take-up decision?

The exogenous variation in loan size and fees allows me to estimate the marginal effects of these factors on the take-up decision. I estimate the treatment effect of loan size on take up by comparing applicants who receive a larger loan size offer ("hi-credit" treatment group) to applicants who receive the regular offer (control group). The treatment effect of fees is estimated from the comparison between applicants receiving both a fee reduction and a larger loan size offer ("extreme" treatment group) and applicants receiving only the larger loan size offer ("hi-credit" group).

Table 5 shows the pairwise difference in take-up rates among the three groups. Three findings are notable. First, doubling the loan size, or increasing it to 7,464 RMB ($1100), boosts the take-up rate by 6.46 percentage points; this amounts to 11% of the baseline take-up rate in the control group (57.61%). Second, reducing the origination fees by half, or by 900 RMB ($135) on average, results in a 5.44–percentage point increase in the take-up rate; this is equivalent to 8.5% of the baseline take-up rate in the "hi-credit" group. Third, the combined effect of doubled loan size and fee reduction is 11.89 percentage points, which

is equal to the sum of the two treatment effects. The implication is that the marginal effect of loan size and origination fees is additively separable. I use this finding to motivate the structural model's functional form assumptions.

[[INSERT Table 5 about Here]]

In short, these results show that borrowers prefer larger loans and lower origination fees.

3.2.2 How do loan size and fees affect repayment?

The loan RCT also allows me to study the impact of loan size and fees on repayment. In Table 6, I compare the probability of a loan being paid in full—as well as the fraction of payments (as defined in Section 3.1.2)—across groups while controlling for loan grade, week fixed effects, and city-tier (i.e. a grade used by the platform to assess a city?s overall economic developments) fixed effects. I find that: (i) a larger loan size is associated with a lower probability of full repayment ($t = 1.42$) and with a lower fraction of payments; and (ii) reducing the fee by half increases the probability of full repayment by 2% and the fraction of payments by 2%. The dynamics of delinquency rates over the 12-month repayment period is plotted in Panel (b) of Figure 6. The delinquency rates increase over time, but to a less extent for the "extreme" group.

[[INSERT Table 6 about Here]][[INSERT Figure 6 about Here]]

These results show that an increase in loan size or in fees reduces both the likelihood of repayment and the fraction of payments. Loan terms could affect loan performance through two channels. A better loan offer may induce a less risky applicant to borrow (the selection effect) while also reducing the repayment amount (the direct effect). Hence the observed negative relationship is a combination of these two effects. I take this into account—and distinguish between the two effects—when modeling borrowers' repayment decisions.

3.3 The monetary value of privacy

One can derive a rough estimate of the monetary value of privacy of personal data via a back-of-the-envelope calculation that combines the marginal effects from the two RCTs. The idea is as follows. Loan demand depends both on the disclosure requirement and on origination fees. More specifically, loan demand is high in the case of lax disclosure requirements and/or low origination fees. So when individuals are required to provide more information, the resulting decrease in loan demand can be offset by a reduction in origination fees. This calculation recovers applicants' willingness to accept to relinquish personal data.

The results presented so far document that removing the social network ID and employer contact questions leads to a 1.28–percentage point increase in the completion rate and that reducing the origination fee by 900 RMB results in a 5.44–percentage point increase in the take-up rate. Thus QQ and employer contact combined are worth 211 RMB (or $900/5.44) \times 1.28$), which amounts to \$30 in nominal value or \$58 when adjusted for PPP. This is a nontrivial amount: it is equal to 60% the average daily salary (\$48) in China in 2018.

As before, however, such back-of-the-envelope calculations are subject to some limitations. First, selection may occur as applicants proceed through the application process. For example, individuals who agree to disclose information may derive higher utility from the loan than those who do not, which would imply different price sensitivities for applicants reaching different stages of the application. Second, the reduced-form evidence that I have presented does not allow for a *quantitative* assessment of how counterfactual disclosure policies affect borrower welfare and firm profit.

4 A Model of Borrower Disclosure

This section develops a structural model of consumer demand that incorporates individuals' disclosure, borrowing, and repayment choices, based on the model in Einav et al. (2012). A structural model is necessary for two reasons. First, it accounts for the dynamics of individuals' decisions and the resulting selection problem as explained above. Second,

the structural estimates allow me to quantify the impact of several counterfactual disclosure policies on consumer welfare.

The model consists of three stages. At the *disclosure* stage ($t = 0$), the platform requires the applicant to disclose an amount Q of personal information and the applicant decides whether or not to do so. If he discloses, then the platform offers a loan of size L and repayment R.[14] At the *take-up* stage ($t = 1/2$), the applicant decides whether or not to take the loan; if he does, he receives L. At the *repayment* stage ($t = 1$), the borrower decides whether to repay R or to default. Individuals derive utility from the loan yet experience disutility from disclosing personal information; All applicant borrowers are assumed to maximize their individual total expected discounted utility. Because the disclosure and take-up stages occur over a short period of time, I assume that there is no discounting between these two stages. The discount factor between take-up and repayment is β. The individual has income y_0 in the disclosure/take-up period and income y_1 in the repayment period. The income process is stochastic.

Disclosure stage ($t = 0$) In the disclosure stage, the applicant's expected utility from answering a set of questions Q—given her expected loan size L, expected repayment amount R, and initial income y_0—is:

$$V_0(y_0, \theta_0) = \begin{cases} -g(\theta_0; Q) + \mathbb{E}V_{\frac{1}{2}}(y_0; Q, L, R) & \text{if she discloses } Q, \\ u(y_0, 0) + \beta \mathbb{E}u(y_1) & \text{otherwise.} \end{cases} \tag{1}$$

In this expression, $-g(Q)$ is the disutility from disclosure and $g(.)$ reflects the innate preference for privacy. The term $\mathbb{E}V_{1/2}(y_0; L, R)$ represents expected utility in the *take-up* stage, where the expectations are taken with respect to loan size L and repayment R (which are not revealed until that subsequent stage).

[14] Because more than 90% of the loans in my sample have a 12-month maturity, I do not model this dimension of the loan contract.

Take-up stage ($t = 1/2$) After disclosing information, the applicant observes loan terms L and R and decides whether or not to take up the loan. His expected utility is:

$$V_{\frac{1}{2}}(y_0; Q, L, R) = \begin{cases} u(y_0, L) + \beta \mathbb{E} V_1(y_1; Q, R) & \text{if she takes-up the loan,} \\ u(y_0, 0) + \beta \mathbb{E} u(y_1) & \text{otherwise,} \end{cases} \tag{2}$$

here $\mathbb{E}V_1(y_1; R)$ is expected utility from the (subsequent) *repayment* stage. At this point, the only uncertainty for the borrower concerns his future income y_1. If we compare $V_{1/2}$ with V_0, it is clear that—once the disclosure cost is sunk—there is no reason for the borrower not to take the loan (unless the offer's terms are too onerous compared to the expected loan terms). The take-up rate evident from the data is 55%, which indicates that individuals do not perfectly forecast L and R.

Repayment stage ($t = 1$) In the last stage, the borrower observes her income y_1 and decides whether to default. Her utility is:

$$V_1(y_1; Q, R) = \begin{cases} u(y_1, 0) - \psi(Q) & \text{if she defaults,} \\ u(y_1, R) & \text{otherwise,} \end{cases} \tag{3}$$

here $\psi(\cdot)$ is the default cost. So when defaulting, the borrower retains her personal income y_1 but incurs a default cost. This cost may come in various forms—for example, being unable to borrow from the platform in the future and the stress related to debt collection—and may depend on the amount Q of information revealed in the first stage.

4.1 Econometric specification

I now link the model to the data. I observe the characteristics X' of applicants, the amount Q of information required by the platform, and the loan offers $\{L, R\}$ as well as borrowers' decisions about whether or not to disclose personal information ($D \in \{0, 1\}$), to take up the loan ($T \in \{0, 1\}$), and to repay the loan in full ($F \in \{0, 1\}$).

To parameterize the model, I make the following assumptions. First, inspired by the reduced-form evidence, I assume that the marginal utility for privacy is constant. In other words, the marginal disutility from disclosing an additional piece of information does not

depend on the amount of information already disclosed. Thus, for example, the effect of removing the social network question is the same regardless of whether or not applicants must disclose their marital status (and vice versa). The second assumption is that the utility derived from loan, and repayment is additively separable—which is plausible given that the effects of loan size and repayment amount can be linearly summed up (see Section 3.2). Note that utility is also assumed to be linear in loan size and repayment amount. This assumption is not unrealistic given that the loan size is relatively small. Third, following Einav et al. (2012), I approximate borrower types—the logarithm of income y_0—using a linear combination of observable characteristics and an unobservable term that incorporates covariates into the model. Finally, I assume that income follows a first-order autoregressive process. Note that these assumptions imply that applicants are heterogeneous in their income and beliefs about the loan terms, but not in their discount factors and valuations for privacy. Formally, the utility functions and borrower types are parameterized as follows:

$$\begin{aligned}
&\text{disutility from disclosure} : g(\theta_0; Q) = \sum_{q \in \{QQ, marriage, landline\}} \theta_{0,q} \mathbb{1}_q; \\
&\text{initial income} : y_0 = X'\xi_y + \nu_y; \\
&\text{income process} : y_1 = \rho y_0 + \nu_f; \\
&\text{default penalty} : \psi(Q) = \sum_{q \in \{QQ, marriage, landline\}} \theta_{1,q} \mathbb{1}_q.
\end{aligned}$$

After applying these assumptions to the model, we can express individuals' indirect utilities as functions of contract characteristics $\{Q, R, L\}$ and of observable and unobservable personal characteristics $\{X', \nu_y, \nu_q, \nu_f\}$. In addition, I assume that $\{\nu_y, \nu_q, \nu_f\}$ are drawn from a multivariate normal distribution $N(0, W)$ and are drawn independently of $\{X', Q, R, L\}$. We can now link borrower decisions to these covariates and unobservables.

Repayment decision The borrower's utility from making a full repayment, net of utility from default, can be written as

$$V_F \approxeq \theta_y y_1 + \theta_R R + \sum_q \theta_{1,q} \mathbb{1}_q = X' \underbrace{\theta_{\mathcal{X}}}_{\theta_y \rho \xi_y} + \theta_{\mathcal{R}} R + \sum_q \theta_{1,q} \mathbb{1}_q + \underbrace{\varepsilon_F}_{\theta_y \rho \nu_y + \nu_f}.$$

Here ε_F is the unobservable that affects the repayment decision; it includes both the unobservable part, from the econometrician's perspective, of the initial income ν_y and the unknown future income shock ν_f. The term $\theta_{\mathcal{R}}$ measures the direct effect of the repayment amount on default choices, and $\theta_{1,q}$ captures the effect of additional information on repayment probabilities. It is optimal for the borrower to repay in full if and only if $V_F > 0$; therefore, $F = \mathbb{1}\{\varepsilon_F > -X'\theta_{\mathcal{X}} - \theta_{\mathcal{R}}R - \sum_q \theta_{1,q}\mathbb{1}_q\}$.

Because I observe not only whether the loan is paid in full but also the repayment length, I include the latter continuous variable in the model. In particular, it is plausible that borrowers who derive higher utility from a full repayment have a longer repayment length, or a higher fraction of payments (defined in Section 3.1.2). Hence I let "frac" denote the observed fraction of payments made and then refine the preceding equation as follows:

$$frac = \begin{cases} frac^* = X' \underbrace{\alpha_{\mathcal{X}}}_{\alpha_y \rho \xi_y} + \alpha_{\mathcal{R}} R + \sum_q \alpha_{1,q}\mathbb{1}_q + \underbrace{\varepsilon_F}_{\alpha\rho\nu_y + \nu_f} & \text{if } frac^* < 1, \\ 1 & \text{if } frac^* \geqslant 1. \end{cases} \tag{4}$$

The first case applies to borrowers with zero or partial payments (i.e., $F = 0$); the second case applies to borrowers who repaid the loan in full ($F = 1$). The implicit assumption of Equation (4) is that the observed fraction of payments follows a normal distribution that is right-censored at one.

Take-up decision An individual's utility from taking up the loan, net of his outside option, is

$$V_T \approxeq \beta_y y_0 + beta_L L + \beta_R R = X' \underbrace{\beta_{\mathcal{X}}}_{\beta_y \xi_y} + \beta_{\mathcal{L}} L + \beta_{\mathcal{R}} R + \underbrace{\varepsilon_T}_{\beta\nu_y}, \tag{5}$$

where $\beta_{\mathcal{L}}$ and $\beta_{\mathcal{R}}$ are the marginal utility of (respectively) the loan size and repayment amount for disclosing borrowers. Hence the take-up decision can be written as $T = \mathbb{1}\{\varepsilon_T > -X'\beta_{\mathcal{X}} - \beta_{\mathcal{L}}L - \beta_{\mathcal{R}}R\}$.

Disclosure decision The applicants' utility from disclosing personal information in response to the page 1 and page 2 questions, net of her outside option, can be specified

as follows:

$$\begin{aligned}
\text{page 1: } V_{D1} &\cong -\sum_q \theta_{1,q}\mathbb{1}_q + \gamma_{1,y}y_0 + \gamma_{1,\mathcal{L}}E[L] + \gamma_{1,\mathcal{R}}E[R] \\
&= -\sum_q \theta_{1,q}\mathbb{1}_q + X'\underbrace{\gamma_{1,\mathcal{X}}}_{\gamma_{1,y}\xi_y} + \gamma_{1,\mathcal{L}}E[L] + \gamma_{1,\mathcal{R}}E[R] + \underbrace{\varepsilon_{1,D}}_{\gamma_{1,y}\nu_y}; \\
\text{page 2: } V_{D2} &\cong -\sum_q \theta_{2,q}\mathbb{1}_q + \gamma_{2,y}y_0 + \gamma_{2,\mathcal{L}}E[L] + \gamma_{2,\mathcal{R}}E[R] \\
&= -\sum_q \theta_{2,q}\mathbb{1}_q + X'\underbrace{\gamma_{2,\mathcal{X}}}_{\gamma_{2,y}\xi_y} + \gamma_{2,\mathcal{L}}E[L] + \gamma_{2,\mathcal{R}}E[R] + \underbrace{\varepsilon_{2,D}}_{\gamma_{2,y}\nu_y}.
\end{aligned} \tag{6}$$

Here I separately consider page 1 and page 2 disclosure decisions ($D1$ and $D2$, respectively) to account for the selection that occurs between page 1 and page 2. That is, the initial pool of applicants who complete page 1 may be less sensitive to loan terms—than those who also complete page 2. I therefore allow $\gamma_{\mathcal{L}}$ and $\gamma_{\mathcal{R}}$ to differ in the two expressions displayed above. Yet because I assume a homogeneous discount factor, I expect to observe a similar ratio between $\gamma_{\mathcal{L}}$, and $\gamma_{\mathcal{R}}$ in these two expressions.

The parameters θ_{QQ}, θ_{marital}, and θ_{landline} also have different interpretations for page 1 versus page 2 choices. In particular: $\theta_{1,\text{QQ}}$ and $\theta_{1,\text{marital}}$ capture the applicant's disutility from disclosing (respectively) his QQ ID and marital status, while $\theta_{2,\text{QQ}}$ and $\theta_{2,\text{marital}}$ capture the differential willingness of applicants who have disclosed (respectively) their QQ number and marital status to complete page 2. It is optimal for the borrower to disclose if and only if $V_D > 0$. Equivalently, $D = \mathbb{1}\{\varepsilon_D > \sum \theta_q \mathbb{1}_q - X'\gamma_{\mathcal{X}} - \gamma_{\mathcal{L}}E[L] - \gamma_{\mathcal{R}}E[R]\}$ for $D \in \{D_1, D_2\}$.

Recall that borrowers make disclosure decisions based on their expectations about loan size and repayment amount. I further assume that borrowers have rational expectations in this sense: each borrower i can predict the loan size $\hat{L}_i$ and cost of borrowing $\hat{R}_i$ up to a zero-mean error with respective precision parameters σ_R and σ_L (Equation 7). To construct $\hat{R}_i$ and $\hat{L}_i$, I follow two steps. First, I regress actual loan terms R and L on a set of observables using the subsample of borrowers who have completed the application (but excluding borrowers in the "hi-credit" and "extreme" groups). I then use the linear relationship to predict $\hat{R}_i$ and $\hat{L}_i$ for all applicants. By construction, the expectations $\hat{R}_i$ and $\hat{L}_i$

have the same means as the actual R and L:

$$\begin{aligned} E[R_i] &= \hat{R}_i + \Delta_R \text{ with } \Delta_R \sim N(0, \sigma_R), \\ E[L_i] &= \hat{L}_i + \Delta_L \text{ with } \Delta_L \sim N(0, \sigma_L). \end{aligned} \tag{7}$$

While the previous back-of-the-envelope calculation is subject to the selection problem, this structural model addresses this concern by allowing disclosing and nondisclosing borrowers to have different marginal utilities for the loan size and for the origination fee. Thus $\gamma_{\mathcal{L}}$ and $\gamma_{\mathcal{R}}$ can differ from $\beta_{\mathcal{L}}$ and $\beta_{\mathcal{R}}$. The discrepancy between $\gamma_{\mathcal{R}}$ and $\beta_{\mathcal{R}}$, or between $\gamma_{\mathcal{L}}$ and $\beta_{\mathcal{L}}$, measures the extent of the selection issue. From the applicant's perspective, the monetary value of her QQ ID, marital status, and employer contact are given by (respectively) $\theta_{\text{QQ}}/\gamma_{1,L}$, $\theta_{\text{marital}}/\gamma_{1,L}$, and $\theta_{\text{employer}}/\gamma_{2,L}$.

Stochastic assumptions To close the model, I specify a stochastic structure for the unobservables, ε_F, ε_T, and ε_D, where $\varepsilon_D = (\varepsilon_{D1}, \varepsilon_{D2})$. From earlier assumptions, $(\varepsilon_F, \varepsilon_T, \varepsilon_D)$ is also normally distributed with mean zero and covariance matrix Σ:

$$\begin{pmatrix} \varepsilon_F \\ \varepsilon_T \\ \varepsilon_D \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_F^2 & \rho_{\mathcal{F}T}\sigma_F\sigma_T & \rho_{FD}\sigma_F\sigma_D \\ \rho_{TF}\sigma_T\sigma_F & 1 & \rho_{TD}\sigma_T\sigma_D \\ \rho_{DF}\sigma_D\sigma_F & \rho_{DT}\sigma_D\sigma_T & \Sigma_D^2 \end{pmatrix} \right],$$

with

$$\left(\Sigma_D^2\right) \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{D1}^2 & \rho_{\mathcal{D}1D2}\sigma_{D1}\sigma_{D2} \\ \rho_{\mathcal{D}1D2}\sigma_{D1}\sigma_{D2} & \sigma_{D2}^2 \end{pmatrix} \right].$$

There are a few points worth mentioning. First, and without loss of generality, the variance of ε_T is normalized to 1. Second, the correlation parameters ρ_{TF}, ρ_{DF} and ρ_{TD} have important economic meaning: they characterize the selection issue in the dynamic choices. For example, $\rho_{TF} > 0$ implies that—for a given loan offer—borrowers with the same observable characteristics, who are more likely to accept are also more likely to default. Finally, the variance parameters capture the prominence, in borrower decisions, of unobserved characteristics relative to observed characteristics.

4.2 Estimation

The individual's optimization problem is represented by the demand system (4), (5) and (6). I use these equations to compute the likelihood of observing each individual choice before combining the set of probabilities into a full likelihood function.

Choice probabilities I start with the probability of observing disclosure outcomes. For both page 1 and page 2, borrower i discloses if and only if $\varepsilon_{D,i} > -\bar{V}_{D,i}$ for $D \in \{D1, D2\}$, where $\bar{V}_{D,i} = -\sum \theta_{j,q} \mathbb{1}_{q,i} + X_i' \gamma_{j,\mathcal{X}} + \gamma_{j,\mathcal{L}} E[L_i] + \gamma_{j,\mathcal{R}} E[R_i], j \in \{1, 2\}$. The probability of disclosure, given $E[L_i]$ and $E[R_i]$, is then

$$P_{D_i=1} = \Pr(\varepsilon_{D,i} > -\bar{V}_{D,i}) = \Phi\left(\frac{\bar{V}_{D,i}}{\sigma_D}\right),$$

where Φ denotes the standard normal cumulative density function. Yet because the econometrician cannot observe the beliefs $E[R_i]$ and $E[L_i]$ of borrowers, I must integrate this probability over the distributions of those beliefs, which are given by Equation 7. This procedure yields an expression for the likelihood of disclosure:

$$P_{D_i=1} = \iint \Phi\left(\frac{\bar{V}_{D,i}}{\sigma_D}\right) d\Delta_R \, d\Delta_L.$$

Second, the probability of observing a borrower taking up the loan, given the disclosure decision and $\varepsilon_{D,i}$, is

$$\begin{aligned} P_{T_i=1|D_i=1} &= \Pr(\varepsilon_{T,i} > -X_i'\beta_{\mathcal{X}} - \beta_{\mathcal{L}} L_i - \beta_{\mathcal{R}} R_i | \varepsilon_{D,i}) \\ &= F_{\varepsilon_T|\varepsilon_D}(X_i'\beta_{\mathcal{X}} + \beta_{\mathcal{L}} L_i + \beta_{\mathcal{R}} R_i), \end{aligned}$$

where F is a conditional cumulative normal distribution with mean $\mu_{\varepsilon_T|\varepsilon_D}$ and variance $\sigma^2_{\varepsilon_T|\varepsilon_D}$. These moments can be computed in a straightforward way from the covariance matrix Σ. For borrowers who choose not to disclose, I do not observe the take-up decision; hence these borrowers' take-up choices do not enter the likelihood function.

Calculation of this likelihood is complicated by the fact that ε_D is not observed. As a result, I cannot directly calculate the moments of the conditional distribution $\varepsilon_T|\varepsilon_D$ and, instead, must integrate over all the ε_D that result in an observed disclosure decision. Put $\bar{V}_T = X'\beta_{\mathcal{X}} + \beta_{\mathcal{L}} L + \beta_{\mathcal{R}} R$. Then the likelihood of observing a borrower taking up the

loan—after disclosing his personal information—is

$$P_{T_i=1|D_i=1} = \int_{\bar{V}_{D,i}}^{\infty} F_{\varepsilon_T|\varepsilon_D}(\bar{V}_{T,i}) f(\varepsilon_D)\, d\varepsilon_D.$$

All integrals in the likelihood function are computed by simulation. For the integral displayed here, I simulate values of ε_D from the distribution $f(\varepsilon_D)$ in the region given by the limits of the integral; after using those simulated values to compute likelihoods for each take-up outcome, I then average these likelihoods across simulations.

Finally, I compute the probability of observing full payments and partial payments. The probability of full payments is expressed formally as:

$$\begin{aligned} P_{\text{frac}_i=1|T_i=1,D_i=1} &= \Pr\left(\varepsilon_{F,i} > 1 - X_i'\alpha_{\mathcal{X}} - \alpha_{\mathcal{R}} R_i - \sum_q \theta_{1,q}\mathbb{1}_q | \varepsilon_D, \varepsilon_T\right) \\ &= \int_{\bar{V}_{D1,i}}^{\infty}\int_{\bar{V}_{T,i}}^{\infty} F_{\varepsilon_F|\varepsilon_D,\varepsilon_T}\left(1 - X_i'\alpha_{\mathcal{X}} - \alpha_{\mathcal{R}} R_i - \sum \alpha_{1,q}\mathbb{1}_{q,i}\right) f(\varepsilon_D, \varepsilon_T)\, d\varepsilon_T\, d\varepsilon_D; \end{aligned}$$

the probability of observing a partial payment $\text{frac}_i < 1$ is

$$\begin{aligned} P_{\text{frac}_i|T_i=1,D_i=1} &= \Pr\left(\varepsilon_{F,i} = \mathit{frac}_i - X_i'\alpha_{\mathcal{X}} - \alpha_{\mathcal{R}} R_i - \sum_q \theta_{1,q}\mathbb{1}_q | \varepsilon_D, \varepsilon_T\right) \\ &= \int_{\bar{V}_{D1,i}}^{\infty}\int_{\bar{V}_{T,i}}^{\infty} f_{\varepsilon_F|\varepsilon_D,\varepsilon_T}\left(\mathit{frac}_i - X_i'\alpha_{\mathcal{X}} - \alpha_{\mathcal{R}} R_i - \sum \alpha_{1,q}\mathbb{1}_q\right) f(\varepsilon_D, \varepsilon_T)\, d\varepsilon_T\, d\varepsilon_D, \end{aligned}$$

where $f_{\varepsilon_F|\varepsilon_D,\varepsilon_T}$ is the probability distribution function of ε_F conditional on ε_D and ε_T. For these integrals, I simulate values of ε_D and ε_T from the joint distribution $f(\varepsilon_D, \varepsilon_T)$ in the region given by the limits of the integral; I then use these simulated values to compute likelihoods for each default outcome and average those likelihoods across simulations.

Log-likelihood function The log-likelihood function $\log \mathcal{L}(D_i, T_i, \mathit{frac}_i | X_i', Q_i, L_i, R_i)$ combines the disclosure, take-up, and default probabilities . The outcome for any borrower must be one of the following types: I_1, no disclosure; I_2, disclosure but no take-up; I_3, disclosure, take-up, and full payment; or I_4, disclosure, take-up, and partial or zero payment.

The log-likelihood function for all observed outcomes, $\log \mathcal{L}$, is

$$\begin{aligned}\log \mathcal{L} = & \sum_{i \in I_1} \log(1 - P_{D_i=1}) \\ & + \sum_{i \in I_2} [\log(P_{D_i=1}) + \log(1 - P_{T_i=1|D_i=1})] \\ & + \sum_{i \in I_3} [\log(P_{D_i=1}) + \log(P_{T_i=1|D_i=1}) + \log(P_{\text{frac}_i=1|D_i=1,T_i=1})] \\ & + \sum_{i \in I_4} [\log(P_{D_i=1}) + \log(P_{T_i=1|D_i=1}) + \log(P_{\text{frac}_i|D_i=1,T_i=1})].\end{aligned}$$

There are six sets of parameters to be estimated: $\alpha = \{\alpha_{\mathcal{X}}, \alpha_{\mathcal{R}}, \alpha_q\}$; $\beta = \{\beta_{\mathcal{X}}, \beta_{\mathcal{L}}, \beta_{\mathcal{R}}\}$; $\gamma = \{\gamma_{\mathcal{X}}, \gamma_{\mathcal{L}}, \gamma_{\mathcal{R}}\}$; $\theta = \{\theta_{\text{QQ}}, \theta_{\text{marriage}}, \theta_{\text{landline}}\}$; $\sigma = \{\sigma_R, \sigma_L\}$, and Σ. I use simulated maximum likelihood to obtain the estimates. That is, $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\sigma}, \hat{\theta}, \hat{\Sigma}\} = arg \ \ max \log \mathcal{L}$. Standard errors are obtained using a bootstrap method.

5 Demand Estimation

In this section, I present results from estimating the demand system (4), (5) and (6). Estimation of the disclosure choices is based on data from the disclosure RCT, whereas estimation of the take-up and repayment decisions uses data from the disclosure RCT and the loan RCT both.

Now consider how variation in the data identifies the key parameters. The quantities of interest are (i) the sensitivity of disclosure decisions to the disclosure requirement and loan terms and (ii) the sensitivity of the take-up and repayment decisions to loan terms. First, the exogenous variation in disclosure requirements identifies the sensitivity of disclosure decisions to the disclosure requirement. Second, the expected loan size and repayment amount are constructed by exploiting only time variation in the platform pricing scheme but also borrower characteristics. By including borrower characteristics in the demand model, I use the remaining time variation in expected loan terms to identify the disclosure decisions in response to loan size and repayment amount. The time variation in actual loan terms also identifies how take-up and repayment decisions respond to loan terms. Note that the loan RCT's exogenous variation in loan size and repayment provides an additional source of variation that can be used to identify the sensitivity of take-up and repayment

decisions to loan terms.

5.1 Estimation results

Table 7 reports summary statistics for the data along with averages predicted by the model; this table shows that the model is able to fit the data's key moments well. Figure 7 plots the "fraction of payments" distribution in the data versus that predicted by the model. It is clear that the model also does well in matching the distribution of repayment length. This result strongly suggests that the distributional assumption imposed by the model (i.e., truncated normal for the repayment length) is a plausible one.

[[INSERT Table 7 about Here]]

Table 8 presents the estimated marginal effects of disclosure requirements and loan terms on page 1 disclosure, page 2 disclosure, take-up decisions, and fraction of payments. The first two columns report the marginal effects of the variables on the probability of completing page 1 and page 2, respectively, along with the associated t-statistics. Requiring the QQ number on page 1 reduces the completion probability of that page by 2.88 percentage points, and requiring the employer contact on page 2 reduces the completion probability of that page by 0.7 percentage points; these values are consistent with the reduced-form evidence. Moreover, applicants are sensitive to expected loan size and repayment amount. Increasing the expected loan size by 1,000 RMB, while holding repayment fixed, increases the probability of page 1 and page 2 completion rates by (respectively) 20.3 and 8.3 percentage points. A 1000-RMB reduction in the repayment amount increases those respective probabilities by 13.87 and 5.40 percentage points.

[[INSERT Table 8 about Here]]

Combining the sensitivity to loan terms and the sensitivity to privacy, I find that social network ID and employer contact are valued at 145 RMB ($0.0288/0.2028\times1000$) and 85 RMB ($0.007/0.0826\times1000$), respectively. Their combined value, 230 RMB, or \$33, accounts for 70% of the daily average salary in China in 2018. This value is also 9% larger than the back-of-the-envelope calculation (211 RMB)—a gap that would be even larger if

the latter uses the present value of the fee reduction. The reason for this gap is intuitive: the structural model accounts for the selection effect and also for borrowers discounting the reduction in future fee repayment. Applicants who at least begin page 2 derive higher utility from the loan than do those who fail to complete page 1. This result implies that applicants who make it to page 2 are less sensitive to private questions than are those in the initial pool of applicants. It follows that the simple ratio of privacy sensitivity to price sensitivity underestimates the value of employer contact, which is not requested until page 2.

I derive the implied discount factor for applicants, in various stages, using their sensitivity to loan size and repayment amount. A loan's present value is defined as $V(\text{loan}) = L - \sum_{\tau=1}^{12} \delta^{\tau} \times R/12$; in the take-up stage, for example, the estimated sensitivity to L in comparison with the sensitivity to R is $-\hat{\beta}_{\mathcal{R}}/\hat{\beta}_{\mathcal{L}}$. Hence the discount factor δ of borrowers reaching the take-up stage satisfies $\sum_{\tau=1}^{12}(\delta^{\tau}/12) = -\hat{\beta}_{\mathcal{R}}/\hat{\beta}_{\mathcal{L}}$. A similar calculation applies to applicants in the disclosure stage. Using this formula, I calculate an annual discount factor of 0.48, or an annual discount rate of 76.6%, for applicants starting page 1 of the questionnaire (see Table 9a). Consistent with the model's assumption on a homogeneous discount rate, applicants proceeding to subsequent stages (page 2 and the take-up stage) have similar discount factors, with respective averages of 0.43 and 0.45.

[[INSERT Table 9 about Here]]

It is useful to interpret the magnitude of the value of personal data as a share of the value of loans. Using the estimated sensitivities to loan amount and repayments, I calculate the value of loans for applicants in various stages. As reported in part b of Table 9a, applicants who complete page 1, those who complete page 2, and successful borrowers value loans at (on average) 1,087 RMB, 1,197 RMB, and 2,939 RMB, respectively, where the respective average (expected) loan sizes are 1,418 RMB, 1,536 RMB, and 3,770 RMB. The gap in the valuations for loans on page 1 and page2 reflects the selection problem. In comparison, the combined value of social network ID and employer contact is 255 RMB, accounting for 7.8% of the total loan value for successful applicants.

The estimates of take-up and repayment behavior are reported in (respectively) columns

[3] and [4] of Table 8. For individuals reaching these two stages, I observe the internal loan grade—which, not surprisingly, is a strong predictor of take-up probability and expected repayment. Relative to low-risk borrowers, the take-up probabilities are 7.47 and 8.22 percentage points higher for medium-risk and high-risk borrowers, respectively. I also estimate that a 1,000-RMB increase in required repayment amount reduces the expected fraction of payments by 2.9 percentage points. In comparison with low-risk borrowers, the fraction of payments is 4.75 and 8.29 percentage points lower for (respectively) medium and high-risk borrowers. In line with the reduced-form evidence, disclosure requirements have no statistically significant effect on repayment behavior: the expected fraction of payments of borrowers who disclose their social network ID and employer contact is similar to that of those who are not asked to disclose these items. This also suggests that the default cost is independent of whether or not applicants have disclosed the three items.

The lower portion of Table 8 reports the estimated variances and covariances of the unobserved individual characteristics. I find no evidence for adverse selection; that is, the correlation between the unobservable drivers of take-up and repayment decisions is insignificant. Similarly, there is no evidence for the adverse selection with respect to the disclosure requirement. Although the unobservable drivers of page 1 and page 2 disclosure decisions are negatively correlated with those of the take-up decision, the magnitudes are quite small: $\rho_{D1,T} = -0.021$ and $\rho_{D2,T} = -0.016$.

5.2 Borrower welfare

The demand estimates also allow me to assess the welfare effects of different disclosure requirements. In particular, applicants' (expected) utility can be expressed in monetary terms using the utility derived from 1000 RMB, which equals the point estimates of $\gamma_{\mathcal{R}}$ and $\beta_{\mathcal{R}}$ in the disclosure and take-up stages, respectively.

I therefore compute the welfare measure for applicants facing the five different disclosure requirements in the disclosure RCT. I calculate that removing the social network ID question and the employer contact question increases expected average applicant welfare by 13.4%—from \$31.7 to \$35.0 (see Table 9b). This welfare improvement stems from two margins: 7.4% is from the *intensive* margin (i.e., successful applicants enjoy higher utility,

on average, thanks to lower disclosure requirements); the average welfare for these applicants increases from \$470 to \$505; Another 6% is from the *extensive* margin (i.e., borrowers who would have exited in the disclosure stage now proceed to the borrowing stage).

6 Platform Profit

This section addresses the costs of disclosure requirements for the platform. Strict disclosure requirements lead some potential borrowers to drop out of the application process, reducing loan demand and thereby the platform's revenue. I assess the implied loss of revenue while abstracting from the potential platform benefits from collecting information.[15]

To quantify the cost of excessive data collection for the platform and investors, I assess their revenue and profit under different disclosure policies. This exercise is fairly intuitive: the demand estimates allow me to predict the probability of disclosure and take-up—in addition to the fraction of payments—as functions of observables $\{X', Q, L, R\}$. The two probabilities yield the demand for loans, and the payment fraction is useful for calculating (conditional on loan origination) the expected profit per loan. The only other component of the platform's profit is the cost of lending, which I estimate next.

6.1 Revenue accounting

To estimate the platform's cost of lending, I assume that cost to be a constant share c of its revenue. This cost may reflect expenses due to customer acquisition, daily operations, and/or debt collection. The platform's revenue—conditional on loan origination—can be written as

$$\pi^P = \sum_{t=1}^{\hat{T}} f_i \cdot L \cdot \frac{1}{\kappa}\left(1 - \frac{1}{1+\kappa}\right)^t;$$

[15] Additional information on applicants may help the platform improve risk scoring, build a more comprehensive customer database for future marketing use, and so forth. Yet as shown in previous sections, omitting the social network ID and employer contact questions changes neither the internal credit scores nor the loan performance. Hence there may be limited benefits on the risk-scoring dimension.

here κ is the platform's monthly discount factor, f is the monthly percentage fee, and $\hat{T}$ is the repayment length ($\hat{T} = 12$ corresponds to full payment).[16] The repayment length can be calculated as $\hat{T}_i = [\text{frac}_i \times 12]$, or the integer part of $\text{frac}_i \times 12$. While *investors?* revenue can be written as

$$\pi^I = -L + \sum_{t=1}^{\hat{T}} \frac{\frac{1}{\kappa}\left(1 - \frac{1}{1+\kappa}\right)^t}{\frac{1}{r}\left(1 - \frac{1}{1+r}\right)^t},$$

where r is the loan's monthly interest rate. I also assume that investors have the same discount factor (κ) as does the platform itself.

The expected profit from a given applicant depends on the loan demand, the revenue conditional on origination, and the lending cost. For an applicant characterized by $\{X_i', \varepsilon_i\}$ and who faces disclosure requirements Q_i and is presented with a loan offer $\{L_i, R_i\}$, the expected profit for the platform ($j = P$) or investors ($j = I$) is:

$$\begin{aligned}\Pi^j(X_i', Q_i, L_i, R_i, \varepsilon_i; c) = {} & P_{D_i=1}(X_i', Q_i, L_i, R_i, \varepsilon_i) \\ & \times P_{T_i=1|D_i=1}(X_i', Q_i, L_i, R_i, \varepsilon_i) \\ & \times \mathbb{E}[(1-c)\pi^j(X_i, R_i, \varepsilon_i)].\end{aligned}$$

I simulate ε_i to obtain the predicted probabilities of disclosure and take-up and the predicted fraction of payments and use these quantities to calculate the expected revenue from loan origination.

6.2 Cost estimation and counterfactual analysis

To estimate the cost of lending, I start by assuming that the firm's current loan offers are profit maximizing. That is to say, the cost of lending "rationalizes" the observed offers. The platform could, in principal, use unlimited sets of alternative loan offers—in other words, varying the loan amount (at both the extensive and intensive margins), fees, and interest rates. In practice, the platform mainly optimizes the loan amount while setting the sum of interest rates and fees at the regulatory cap of 36% to cover default losses. I therefore focus on a single dimension of pricing, the loan amount, in the profit maximization problem.

[16] Following Einav et al. (2012), I assume an 8% internal annual discount rate. The use of any value in the range from 8% to 12% has little effect on the results.

The variable cost c is such that, under the current Q_0, observed L is profit maximizing:

$$\sum_i \Pi^P(X_i, Q_{0,i}, L_i, R_i, \varepsilon_i; c) \geqslant \sum_i \Pi^P(X_i, Q_{0,i}, (1-a)L_i, R_i, \varepsilon; c) \quad \forall a \neq 0.$$

This revealed preference condition allows for the estimation of c.[17] In contrast to a change in the level of L, a percentage change a in L "respects" the original loan amount in the sense that the resulting *change* in loan amount is larger when the original amount is larger. This approach is also consistent with the platform's actual practice: it determines the optimal loan amount by varying the loan amount proportionately. I remark that a percentage change also does not change the loan amount at the extensive margin.

Panel (a) of Table 10 reports the estimates of lending costs. For the offered loan amount to be preferable to any other uniform percentage change in the loan amount, the unobserved lending cost must be (on average) \$0.11 per dollar originated, or \$59 per loan. I also report the cost separately for each group in the disclosure RCT. The average cost is nearly the same across all groups, which implies that disclosure requirements have little effect on the cost of lending. This outcome is not surprising when one considers that, as shown in Section 3, disclosure requirements do not affect the borrower risk pool. In addition, I calculate that the revenue per loan is \$69 for the platform and \$74 for the "combined" lending entity (i.e.,the platform and its investors). The cost and revenue estimates imply that the average net profit per loan is \$9 and that the net profit margin is about 1.5%. This value is consistent with what the platform reports: its profit from first-time applicants is close to zero.

[[INSERT Table 10 about Here]]

With the demand estimates and the cost matrix in hand, I now turn to the counterfactual analysis. Here I consider the case in which applicant responses are required to all three questions (viz., social network ID, marital status, and employer contact). Hence the change in revenue from the "delete all" group indicates the cost of excess data collection for the platform and investors. Formally, the effect—on the expected per-applicant

[17] Here I assume that investor profit is *not* included in the platform's profit maximization problem.

revenue—of collecting additional data items can be expressed as follows:

$$\frac{1}{N}\left(\sum_{i=1}^{N}\Pi^{j}(X_i, Q^{\text{delete all}}, L_i, R_i, \varepsilon_i) - \sum_{i=1}^{N}\Pi^{j}(X_i, Q^{\text{counterfactual}}, L_i, R_i, \varepsilon_i)\right);$$

here $Q^{\text{delete all}}$ denotes the disclosure requirements for the "delete all" group and $Q^{\text{counterfactual}}$ is the counterfactual set of disclosure requirements, which correspond to the control group's disclosure requirements.

The results on counterfactual revenues are presented in Panel (b) of Table 10. The first two data columns report the lending revenue—under the actual disclosure requirements—for (respectively) the platform and the combined lending entity; the third and fourth columns give the counterfactual revenue, and the last two columns report the percentage change in that revenue. According to the fourth column, if the platform imposes a uniform counterfactual disclosure requirement on the five groups, then under this disclosure requirement, the groupwise profit is not statistically different from each other, which confirms the quality of the RCT's randomization. Furthermore, when comparing the groupwise profit in the actual and counterfactual cases, I find that requiring answers to the social network ID and employer contact questions reduces the platform's per-applicant expected revenue from \$4.48 to \$4.00, a decline of 11.8%. The expected revenue for the combined lending entity falls from \$4.88 to \$4.40, or by 10%. I obtain similar numbers when comparing the actual revenues for the "delete all" group and the control group.

These results suggest that excessive data collection can entail substantial costs for the platform by reducing demand from privacy-sensitive consumers.

7 Conclusion

This paper investigates whether and how much online applicants value the privacy of personal information. Using data from two large-scale field experiments conducted on an online peer-to-peer lending platform, I show that privacy is valuable: the social networking ID and employer contact together are valued by online applicants at 230 RMB (\$32), accounting for 8% of the value of a forgone loan. Collecting such information reduces total borrower welfare by 13% and costs the platform \$0.50 (or 11.8%) in expected revenue per applicant.

The study contributes to an ongoing debate on privacy protection regulations such as GDPR and CCPA, which trade off greater consumer privacy protection against the cost of supplying privacy (e.g., compliance costs). This research provides a foundation for a thorough understanding of the welfare effects of privacy protection policies. I focus on the demand side of privacy and offer an accurate estimate of the benefits from protecting consumer privacy. Research on privacy provision by firms (e.g., Ramadorai et al., 2019) and its associated costs would further advance our understanding of the overall impact of regulating privacy protection.

References

Abowd, John M, and Ian M Schmutte, 2019, An economic analysis of privacy protection and statistical accuracy as social choices, *American Economic Review* 109, 171–202.

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar, 2019, Too much data: Prices and inefficiencies in data markets, Technical report, National Bureau of Economic Research.

Acquisti, Alessandro, Leslie K John, and George Loewenstein, 2013, What is privacy worth?, *The Journal of Legal Studies* 42, 249–274.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman, 2016, The economics of privacy, *Journal of Economic Literature* 54, 442–92.

Adams, William, Liran Einav, and Jonathan Levin, 2009, Liquidity constraints and imperfect information in subprime lending, *American Economic Review* 99, 49–84.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace, 2019, Consumer-lending discrimination in the FinTech era, Technical report, National Bureau of Economic Research.

Berg, Tobias, Valentin Burg, Ana Gombović, and Manju Puri, 2019, On the rise of FinTechs–credit scoring using digital footprints, *Michael J. Brennan Irish Finance Working Paper Series Research Paper* .

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan, 2019, The economics of social data .

Bertrand, Marianne, and Emir Kamenica, 2018, Coming apart? Cultural distances in the united states over time, Technical report, National Bureau of Economic Research.

Brusa, Francesca, Xueming Luo, and Zheng Fang, 2019, The power of non-monetary incentive: Experimental evidence from P2P lending in China, *Available at SSRN 3405902*.

Campbell, James, Avi Goldfarb, and Catherine Tucker, 2015, Privacy regulation and market structure, *Journal of Economics & Management Strategy* 24, 47–73.

Cong, Lin William, and Yizhou Xiao, 2019, Information cascades and threshold implementation, *University of Chicago, Becker Friedman Institute for Economics Working Paper* .

Einav, Liran, Amy Finkelstein, and Mark R Cullen, 2010, Estimating welfare in insurance markets using variation in prices, *The Quarterly Journal of Economics* 125, 877–921.

Einav, Liran, Mark Jenkins, and Jonathan Levin, 2012, Contract pricing in consumer credit markets, *Econometrica* 80, 1387–1432.

Foley, Sean, Jonathan R Karlsen, and Tālis J Putniņš, 2019, Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies?, *The Review of Financial Studies* 32, 1798–1853.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, 2018, Predictably unequal? the effects of machine learning on credit markets .

Fuster, Andreas, Matthew Plosser, Philipp Schnabl, and James Vickery, 2019, The role of technology in mortgage lending, *The Review of Financial Studies* 32, 1854–1899.

Goldfarb, Avi, and Catherine Tucker, 2012, Shifts in privacy concerns, *American Economic Review* 102, 349–53.

Ichihashi, Shota, 2019, Dynamic privacy choices, *Available at SSRN 3472151* .

Jentzsch, Nicola, Sören Preibusch, and Andreas Harasser, 2012, Study on monetising privacy: An economic model for pricing personal information, *ENISA, Feb* 1, 1.

Liao, Li, Zhengwei Wang, Hongyu Xiang, and Xiaoyan Zhang, 2017, P2P lending in China: An overview, Technical report.

Miller, Amalia R, and Catherine Tucker, 2009, Privacy protection and technology diffusion: The case of electronic medical records, *Management Science* 55, 1077–1093.

Preibusch, Sören, Dorothea Kübler, and Alastair R Beresford, 2013, Price versus privacy: An experiment into the competitive advantage of collecting less personal information, *Electronic Commerce Research* 13, 423–455.

Ramadorai, Tarun, Ansgar Walther, and Antoine Uettwiller, 2019, The market for data privacy .

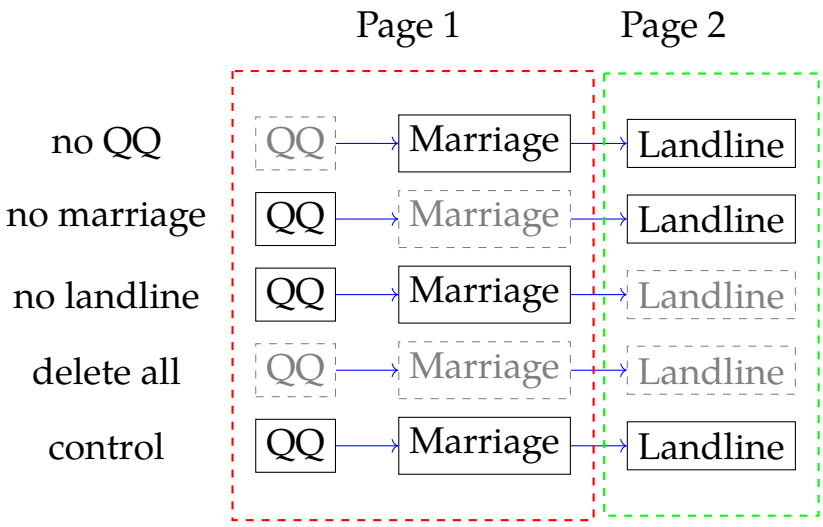
Tang, Huan, 2019, Peer-to-peer lenders versus banks: substitutes or complements?, *The Review of Financial Studies* 32, 1900–1938.

Tsai, Janice Y, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti, 2011, The effect of online privacy information on purchasing behavior: An experimental study, *Information Systems Research* 22, 254–268.

Vallée, Boris, and Yao Zeng, 2019, Marketplace lending: A new banking paradigm?, *The Review of Financial Studies* 32, 1939–1982.

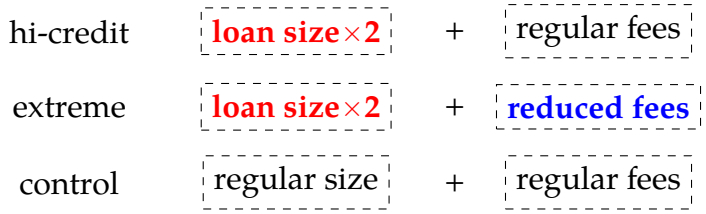
Figures and Tables

Figure 1: The Disclosure RCT



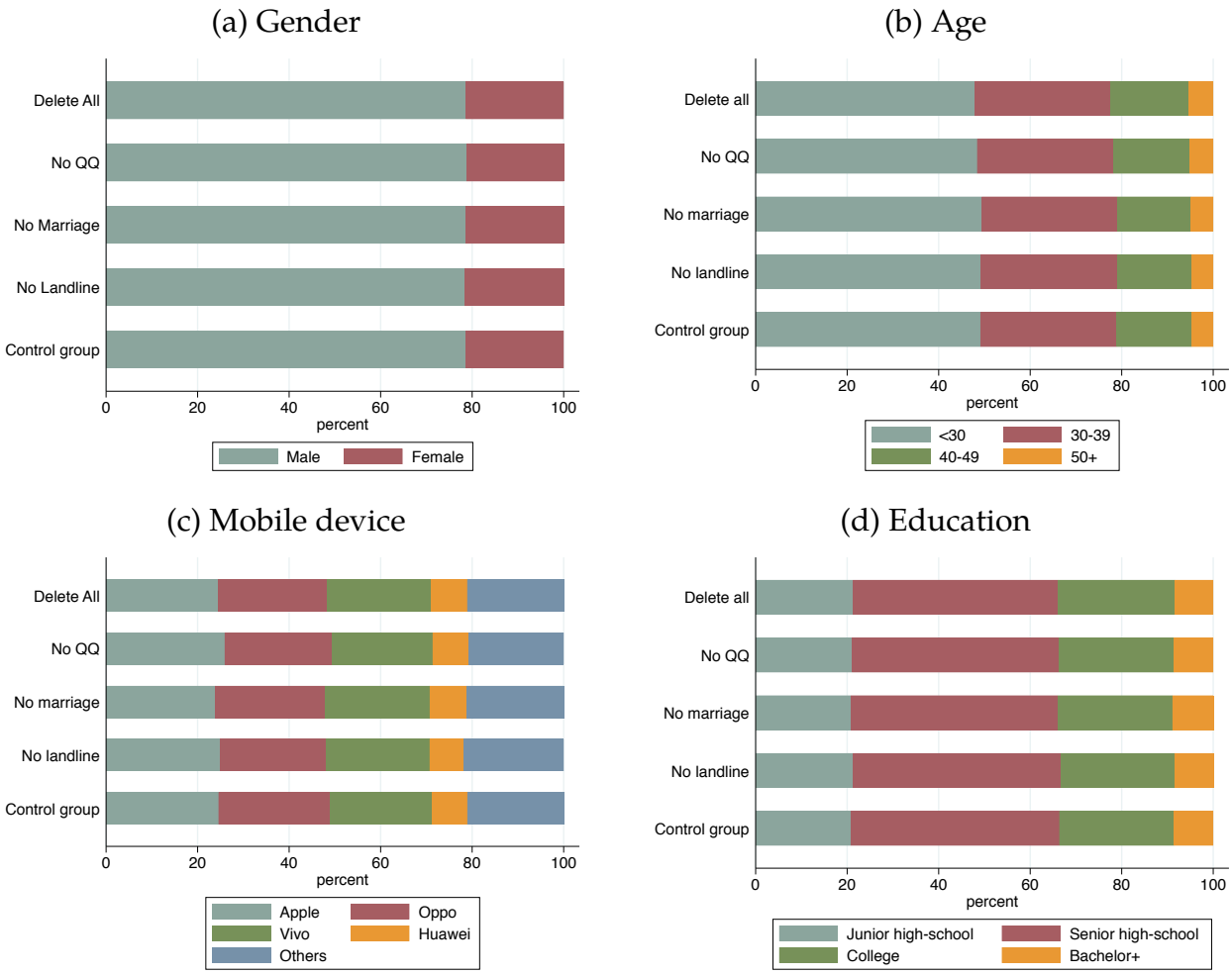
Notes: This figure illustrates the treatments received by the 4 treatment groups in the disclosure RCT. The 4 groups are referred to as "no QQ", "no marriage", "no landline", and "delete all" groups, respectively. Page 1 and Page 2 correspond to the questionnaire page, with QQ and marriage questions appearing on the first and landline question appearing on the second. The complete list of questions on page 1 is: mobile phone number, national ID, QQ, education level, and marriage status in the respective order. The complete list of questions on page 2 includes: profession, employer name, employer contact, and two emergency contacts. Applicants in the control group are required to answer all questions. In each of the 4 treatment groups, one or multiple questions are removed from the application questionnaire. For example, the questionnaire for the "no QQ" group does not include the QQ question. The treatments received by the "no marriage" and "no landline" groups are defined similarly. In the "delete all" group, all the three questions are removed from the questionnaire.

Figure 2: The Loan RCT



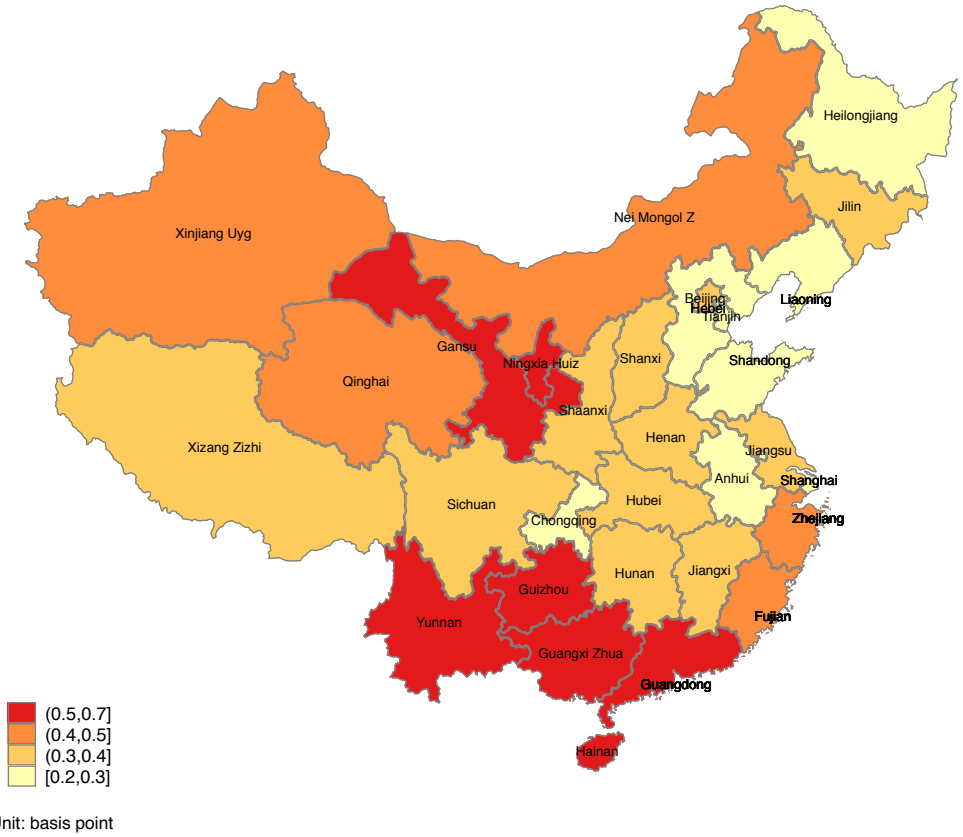
Notes: This figure illustrates the treatments received by the 2 treatment groups in the loan RCT. Borrowers who completed the application are randomly sorted into three groups: one control group ($5/7$ of the RCT sample) and two treatment groups, which are referred to as the "hi-credit" ($1/7$ of the RCT sample) and the "extreme" ($1/7$ of the RCT sample) groups, respectively. Applicants in the control group receive regular loan offers. Applicants in both treatment groups are offered loans twice as large as the regular loans, while applicants in the "extreme" treatment group receive an additional 50% fee reduction.

Figure 3: Demographics by Treatment Groups



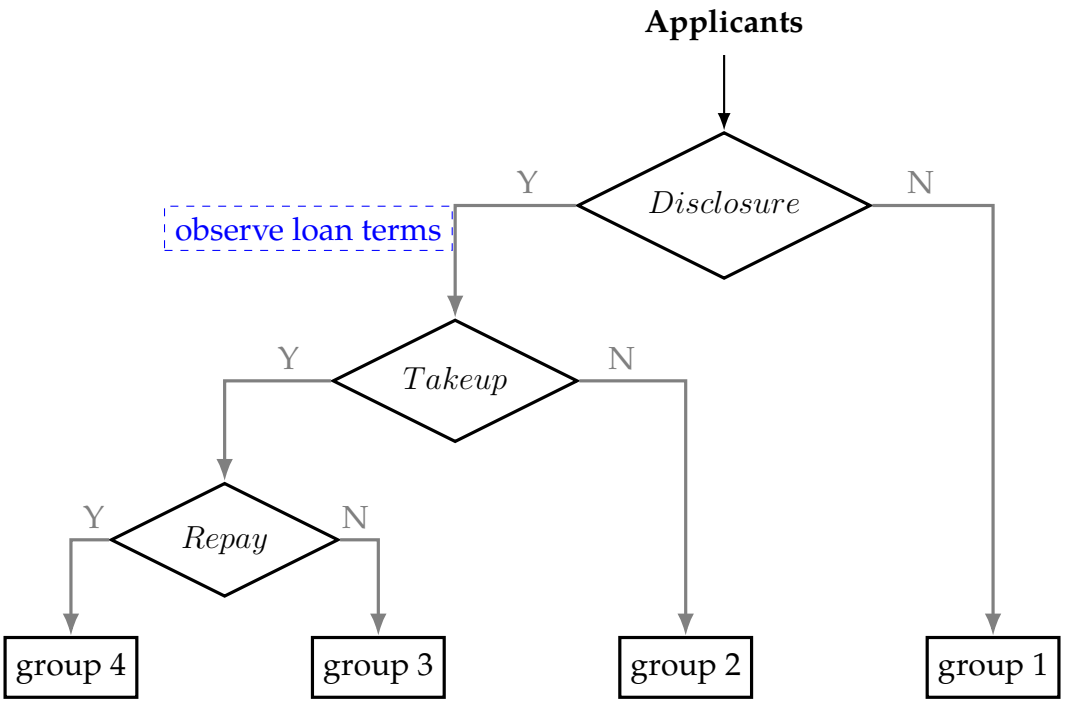
Notes: This figure presents the distribution of gender, age, mobile device brand and education level for the five treatment groups in the disclosure RCT. The y-axis signifies the percentage of applicants in each sub-category. Panel a plots the percentage share of male and female; Panel b plots the percentage share of the four age groups: less than 30, 30-39, 40-49, greater than or equal to 50; Panel c plots the percentage share of applicants using the following four major mobile brands: Apple, Vivo, Oppo, and Huawei; Panel d plots the share of applicants with the following education levels: junior high school and equivalent, senior high school and equivalent, three-year college or technical schools, and bachelor or above.

Figure 4: Geographical Location of Applicants



Notes: The figure presents the P2P penetration rate (in basis points), defined as the total number of applications on the platform per 100 internet users, across geographies in mainland China. The 31 provinces where this platform has lending activities are sorted by the P2P penetration rate into four quartiles.

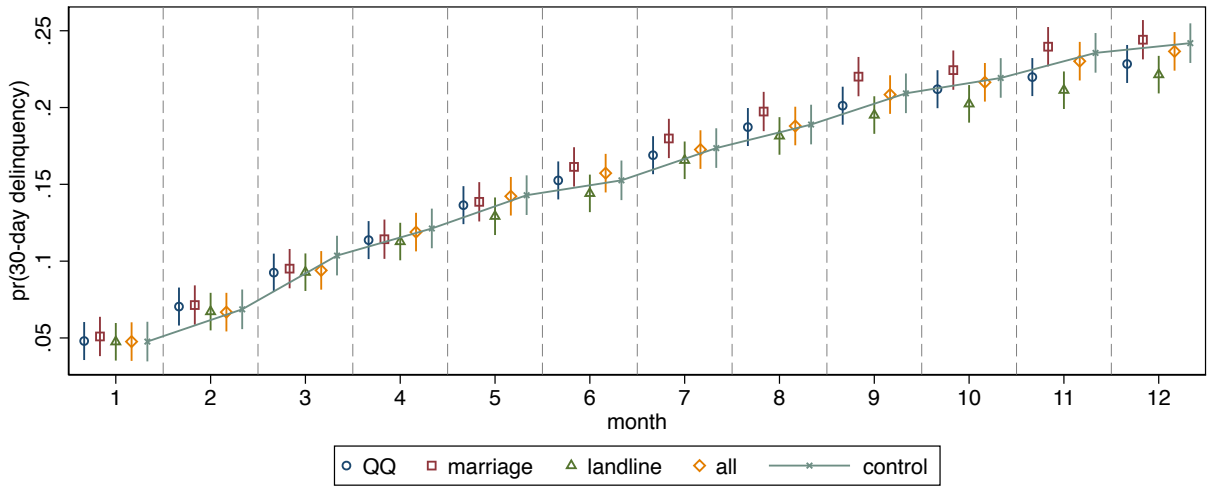
Figure 5: Applicants' Decision Process



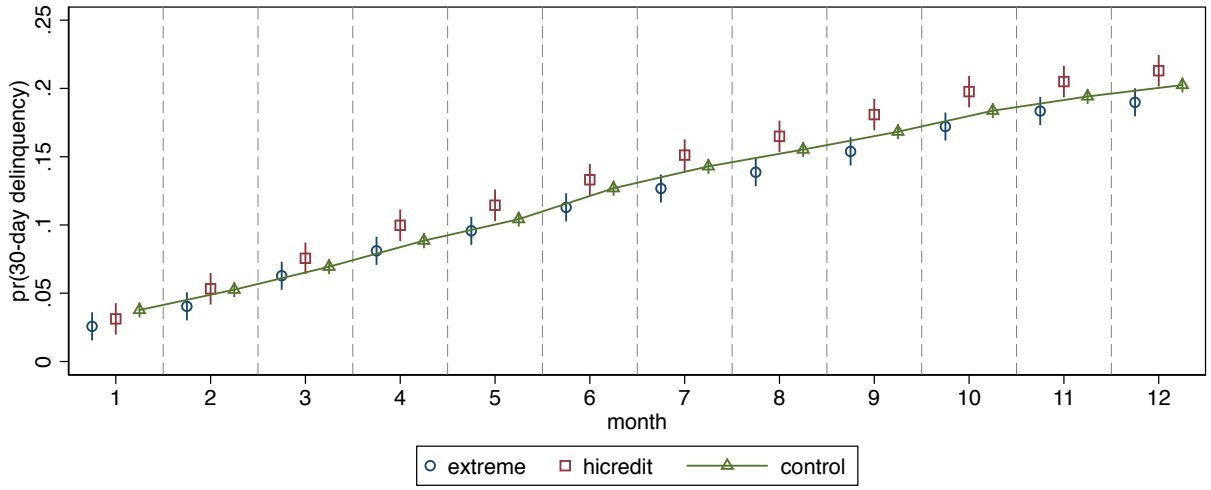
Notes: The figure illustrates the decision process of applicants. There are three stages in the model. In the first stage, applicants choose whether to disclose information before observing the loan offer. After completing the loan application, applicants observe the loan offer and then decide whether to accept it or not in the second stage. In the last stage, borrowers choose between repayment and default.

Figure 6: Delinquency Rates over Time

(a) Disclosure RCT

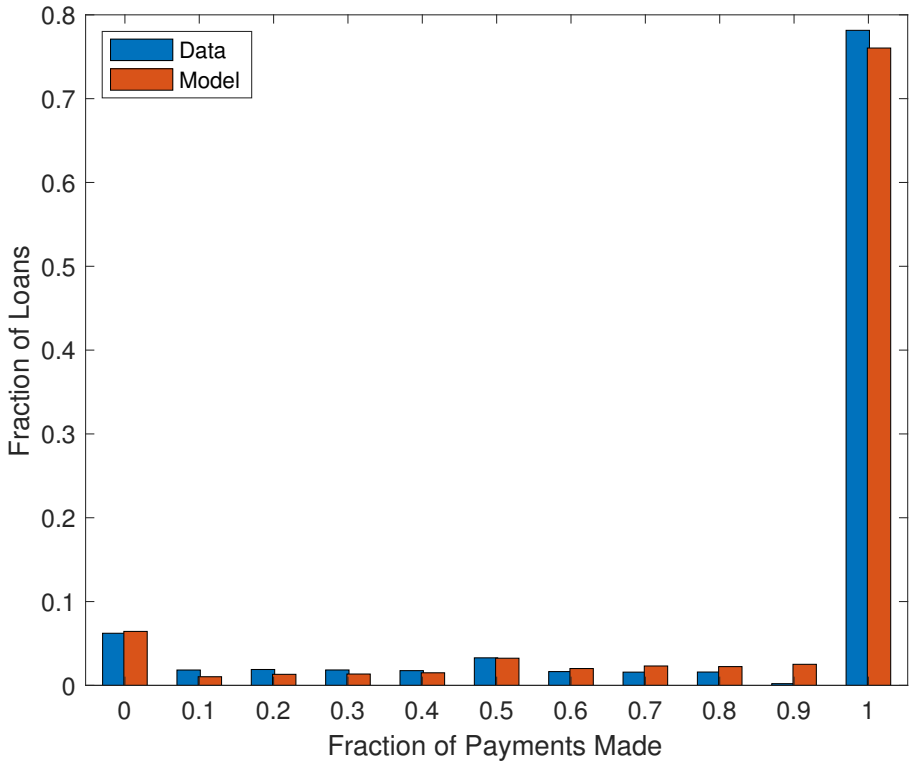


(b) Loan RCT



Notes: This figure shows the delinquency rates (percentage of delinquent loans) over the 12-month life cycle of loans. Subfigure (a) presents the delinquency rates of loans for the five groups in the disclosure RCT, while subfigure (b) shows loan delinquency rates for the three groups in the loan RCT.

Figure 7: Model Fit: Fraction of Payments Made



Notes: This figure shows the model fit for fraction of payments made. The x-axis signifies the payments made as a fraction of the loan principal. The y-axis indicates the fraction of loans in each of the ten bins.

Table 1: The Treatment Effects of Disclosure Requirements on Completion Rates

			Treated		Control	
	Diff.	t-stat	completion%	N	completion%	N
	[1]	[2]	[3]	[4]	[5]	[6]
			a. Page 1			
no QQ	2.97	[9.79]	51.57	54,874	48.60	53,507
no marriage	0.30	[0.97]	48.89	53,257	48.60	53,507
no QQ & marriage	3.17	[10.43]	51.76	54,883	48.60	53,507
			b. Page 2			
no landline	0.61	[2.17]	31.37	53,867	30.76	53,507
delete all	1.28	[4.55]	32.04	54,883	30.76	53,507

Notes: This table presents the treatment effects of lower disclosure requirements on application completion rates. The outcome variables are the completion rates of page 1 (panel a) and page 2 (panel b), respectively. Column 1 shows the differences in completion rates between the treatment groups and the control group. Column 2 shows the t-statistics of the treatment effect. Columns 3 and 5 display the average completion rates for the treatment and control groups, respectively. Columns 4 and 6 present the number of observations in the treatment and control groups, respectively.

Table 2: Heterogeneity in the Treatment Effects of Disclosure Requirements

Treatment	**no QQ**				**no Landline**			
	complete page 1				complete page 2			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
treatment	0.01**	0.01	0.01*	0.01	0.01	0.01	0.00	0.00
	(2.56)	(1.42)	(1.78)	(1.51)	(1.50)	(1.35)	(0.74)	(0.90)
female	0.04***				0.04***			
	(6.90)				(5.79)			
treatment × female	0.02**				-0.00			
	(2.15)				(-0.28)			
old		-0.02***				-0.07***		
		(-4.22)				(-13.65)		
treatment × old		0.01*				-0.00		
		(1.93)				(-0.35)		
iOS			0.11***				0.15***	
			(13.37)				(13.35)	
treatment × iOS			-0.01				0.00	
			(-0.56)				(0.38))	
bachelor				0.01**				0.11**
				(2.19)				(18.78)
treatment × bachelor				0.00				0.00
				(0.60)				(0.22)
Constant	0.72***	0.74***	0.60***	0.73***	0.46***	0.51***	0.43***	0.43***
	(272.16)	(220.41)	(149.87)	(253.01)	(152.66)	(132.78)	(102.92)	(132.24)
Observations	71,956	71,956	39,013	72,671	69,986	69,413	38,103	70,478
R^2	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00

Notes: This table presents the heterogeneity in the treatment effects. Columns 1-4 present the treatment effect of removing the social network ID question and use page-1 completion indicator as the dependent variable. Columns 4-8 present the treatment effect of removing the employer contact question and use page-2 completion indicator as the dependent variable. In each of the four columns, the treatment dummy is interacted with the following four variables: a female indicator, an indicator for whether the applicant is above 30 (the median age in the sample), an iOS device indicator, and an indicator for whether the applicant has a bachelor or above degree, respectively.

Table 3: Testing the Adverse Selection Channel

	grade	pre-approval	loan size	fee	fraction of payments
	[1]	[2]	[3]	[4]	[5]
deleteall	0.02	-0.01	21.30	0.01	0.00
	(1.50)	(-1.64)	(0.38)	(0.12)	(0.35)
QQ	-0.00	-0.01**	29.22	-0.09	0.01
	(-0.13)	(-2.14)	(0.52)	(-1.00)	(0.85)
marriage	-0.02	-0.00	39.15	-0.01	-0.00
	(-1.58)	(-0.72)	(0.69)	(-0.15)	(-0.23)
landline	0.00	-0.00	16.51	-0.12	0.01
	(0.28)	(-0.18)	(0.29)	(-1.34)	(1.51)
Constant	4.12***	0.43***	3797.65***	28.58***	0.85***
	(447.21)	(110.89)	(94.63)	(469.01)	(157.56)
Observations	73,051	73,051	31,006	31,006	15,532
R^2	0.00	0.00	0.00	0.00	0.00

Notes: This table compares the risk profile of applicants in the five groups in the disclosure RCT, provided that applicants have completed the applications. The dependent variables in columns 1-5 are loan grades assigned by the platform, an indicator for loan pre-approval, loan size, origination fees, and fraction of payments made at maturity, respectively.

Table 4: Testing the Exertion and Memory Channel

(a) QQ

		no QQ		Control	
Diff.	t-stat	completion%	N	completion%	N
		a. conditional completion page 2			
-1.52	[-3.61]	59.96	28,104	61.48	25,821

(b) Landline

	complete page 2	
	employed	unemployed
	[1]	[2]
landline	0.02***	-0.01
	(2.75)	(-1.43)
old	-0.06***	-0.07***
	(-9.09)	(-10.05)
female	0.04***	0.06***
	(4.76)	(7.78)
bachelor	0.10***	0.07***
	(14.89)	(9.36)
Constant	0.53***	0.63***
	(79.61)	(103.32)
Observations	21,477	20,862
R^2	0.02	0.01

Notes: This table examines whether applicants completion rates can be explained by exertion or memory. In panel a, I compare the conditional completion rate of page 2 for the control group and the "no QQ" group. Panel b tests the memory channel by showing that the treatment effect of removing landline on completion rates is entirely driven by employed borrowers.

Table 5: The Treatment Effects of Loan Size and Fees on Take-up Rates

			Treated		Control	
	Diff.	t-stat	take-up%	N	take-up%	N
	[1]	[2]	[3]	[4]	[5]	[6]
hi-credit	6.46	[10.03]	64.06	7,032	57.61	34,366
fee reduction	5.44	[6.84]	69.50	6,977	64.06	7,032
hi-credit & fee reduction	11.89	[18.53]	69.50	6,977	57.61	34,366

Notes: This table presents treatment effects of larger loan size and fee reduction on take-up rates. The three rows in column 1 are the differences in take-up rates between the "hi-credit" and control groups, the "extreme" and the "hi-credit" groups, and the "extreme" and the control groups, respectively. Column 2 shows the t-statistics of the treatment effects. Columns 3 and 5 show the take-up rates for the two corresponding groups, respectively. Columns 4 and 6 show the number of observations in the two corresponding groups, respectively.

Table 6: The Treatment Effects of Loan Size and Fees on Repayment

	full payment		**payment fraction**	
	[1]	[2]	[3]	[4]
hi-credit	-0.01		-0.01*	
	(-1.42)		(-1.71)	
extreme		0.02**		0.02***
		(2.29)		(3.40)
grade=2	-0.03	-0.21	-0.10	-0.12
	(-0.15)	(-0.75)	(-0.83)	(-0.64)
grade=3	-0.06	-0.23	-0.13	-0.14
	(-0.35)	(-0.81)	(-1.08)	(-0.76)
grade=4	-0.07	-0.26	-0.14	-0.15
	(-0.40)	(-0.91)	(-1.16)	(-0.84)
grade=5	-0.08	-0.30	-0.15	-0.18
	(-0.46)	(-1.06)	(-1.30)	(-0.98)
Observations	16,773	6,763	16,773	6,763
R^2	0.01	0.01	0.01	0.01

Notes: This tables presents the effects of loan size and origination fees on repayment. The dependent variable is an indicator for full payment in columns 1-2, and is the fraction of payments made at maturity in columns 3-4. Columns 1 & 3 contrast the "extreme" group to the "hi-credit" group, while columns 2 & 4 compare the"hi-credit" group to the control group.

Table 7: Model Fit

Group	D-Page 1		D-Page 2		Take up		Payments	
	Data	Model	Data	Model	Data	Model	Data	Model
Disclosure experiment								
delete all	0.5164	0.5134	0.6171	0.6167	0.5438	0.5552	0.8449	0.8581
no QQ	0.5142	0.5121	0.6144	0.6129	0.5281	0.5398	0.8544	0.8593
no marriage	0.4875	0.4856	0.6278	0.6259	0.5463	0.5567	0.8430	0.8466
no landline	0.4865	0.4835	0.6408	0.6391	0.5553	0.5658	0.8592	0.8635
control	0.4845	0.4834	0.6308	0.6286	0.5408	0.5515	0.8447	0.8537
loan experiment								
extreme					0.6794	0.6857	0.9017	0.8861
hi-credit					0.6209	0.6310	0.8823	0.8858
control					0.5572	0.5687	0.8866	0.8881

Notes: This table shows model fits. Raw data moments are computed directly from the estimation sample. Demand model moments are computed based on the demand model and the parameter estimates in Table 8. The page-1 disclosure probability is computed using data on all applicants. The page-2 disclosure probability is conditional on the completion of page 1. Take-up probability is conditional on completing page 2. Last, the fraction of payments made is computed using data on originated loans only.

Table 8: Demand Estimation: Marginal Effects

	Disclosure page 1 [1]	t	**Disclosure page 2** [2]	t	**Take-up** [3]	t	**Repayment** [4]	t
Information required								
QQ	-0.0288	(-14.40)	0.0158	(6.33)	0.0034	(2.29)	-0.0025	(-0.31)
marriage	-0.0023	(-1.22)	0.0033	(1.46)	-0.0010	(-0.58)	0.0060	(1.54)
landline			-0.0070	(-3.49)	-0.0031	(-3.49)	-0.0083	(-1.50)
(Expected) Loan terms								
loan (000s)	0.2028	(16.03)	0.0826	(9.01)	0.0643	(9.70)		
repayment (000s)	-0.1387	(-14.53)	-0.0540	(-7.63)	-0.0427	(-7.32)	-0.0029	(-1.82)
Borrower risk category								
high risk					0.0822	(21.58)	-0.0829	(-6.87)
medium risk					0.0747	(24.95)	-0.0475	(-5.04)
low risk					omitted		omitted	
Fixed effects	Gender, Age, City tier, Education, Customer acquisition channel							

Variance Covariance Matrix

	ε_{D1}		ε_{D2}		ε_T		ε_F	
ε_{D1} (disclosure page 1)	0.9760	(43.76)						
ε_{D2} (disclosure page 2)	0.0223	(0.17)	0.9880	(47.94)				
ε_T (take-up)	-0.0210	(-2.80)	-0.0161	(-1.85)	1.0000	-		
ε_F (repayment)	-0.0179	(-1.42)	0.0112	(-0.31)	-0.0152	-	1.0141	(8.18)

Variance of Expectation Errors

$Std(\Delta_L) = 1.1336\ (26.95), \quad Std(\Delta_R) = 1.2760\ (35.48)$

Notes: This table reports the marginal effects estimated using the demand model. Reported estimates in the first three columns show the marginal effects of a 1 unit change in each of the explanatory variables on the probabilities of completing page 1, page 2 and accepting the loan offers. Estimates in the fourth column show the effects of a 1 unit change in each explanatory variable on the fraction of payments made. For dummy variables, this is computed by taking the difference between the probability of sale when the variable is equal to 1 and the probability when the variable is equal to 0 (holding other variables fixed). For continuous variables, this is computed by taking a numerical derivative of the probability of sale with respect to the continuous variable. t-statistics are calculated based on 70 bootstrap samples and are presented in parentheses.

Table 9: Borrower Welfare

(a) Discount Factor and Loan Value

	D - page 1		D - page 2		T	
loan (000s)	0.2028	(6.48)	0.0826	(9.01)	0.0643	(9.70)
repayment (000s)	-0.1387	(-14.53)	-0.0540	(-7.63)	-0.0427	(-7.32)
a. Implied Discount Factor						
monthly	0.9400		0.9325		0.9350	
annual	0.4759		0.4323		0.4464	
b. Value of Loans						
value per USD/RMB	0.76		0.77		0.77	
average loan size ($)	203		220		536	
average loan size (RMB)	1,418		1,536		3,770	
value per loan ($)	155		171		419	
value per loan (RMB)	1,087		1,197		2,938	

(b) Borrower Welfare by Groups

	Avg. Utility (borrowers)		Avg. Utility (applicants)	
	level ($)	diff%	level ($)	diff%
delete all	505	7.4%	35.0	13.4%
no QQ	492	4.7%	34.0	7.1%
no marriage	468	0.0%	32.4	2.4%
no landline	502	6.8%	35.8	12.7%
control	470	-	31.7	-

Notes: This table presents the implied discount factor, value of loans, average borrower utility and average applicant utility, all of which are calculated using the estimates in Table 8. In Subtable (a), columns 1-3 corresponds to all applicants, applicants who have completed page 1, and successful borrowers, respectively. In subtable (b), columns 1 and 3 show the respective average utility in monetary terms for successful applicants and all applicants, and columns 2 and 4 present the percentage difference in the average utility relative to the control group.

s

Table 10: Lending Cost and Profit

(a) Cost and Revenue Estimates

	Cost%	t	Cost per loan	Revenue per loan		Expected revenue per applicant	
				platform	combined	platform	combined
All applicants	11.5%	(11.74)	59	69	74	4.27	4.69
by disclosure group							
delete all	11.7%	(12.41)	58	72	78	4.48	4.88
no QQ	11.5%	(11.74)	59	71	78	4.31	4.71
no marriage	10.9%	(11.90)	59	66	73	4.09	4.52
no landline	11.3%	(11.23)	59	69	76	4.37	4.81
control	10.9%	(10.85)	59	67	74	4.08	4.51

(b) Counterfactual Revenues per Applicant

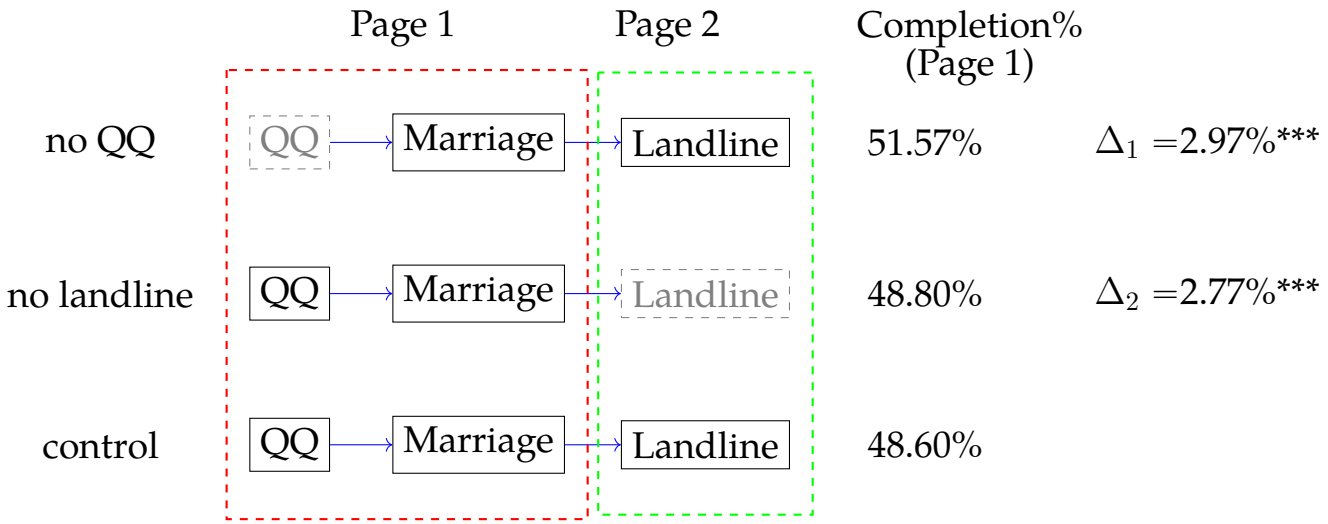
	Actual		Full info. collection		Difference	
	platform	combined	platform	combined	platform	combined
delete all	4.48	4.88	4.00	4.40	-11.8%	-9.9%
no QQ	4.31	4.71	4.08	4.49	-5.5%	-4.8%
no marriage	4.09	4.52	4.04	4.47	-1.2%	-1.2%
no landline	4.37	4.81	4.10	4.53	-6.5%	-5.7%
control	4.08	4.51	4.08	4.51	-	-

Notes: Panel a presents the estimated lending costs and associated per-applicant revenues (in USD). Panel b presents the actual and counterfactual expected revenue per applicant for the five treatment groups in the disclosure RCT separately. The first two columns of panel b present the expected per-applicant revenues for the platform and for the combined lending entity under the current disclosure requirements, respectively, while columns 3-4 show the expected revenue per applicant in the counterfactual case where the full set of information, including the "social network ID", "marital status" and "employer contact" questions, is required. Columns 5-6 present the percentage change in the expected revenue per applicant. t-statistics are calculated based on 70 bootstrap samples and are presented in parentheses.

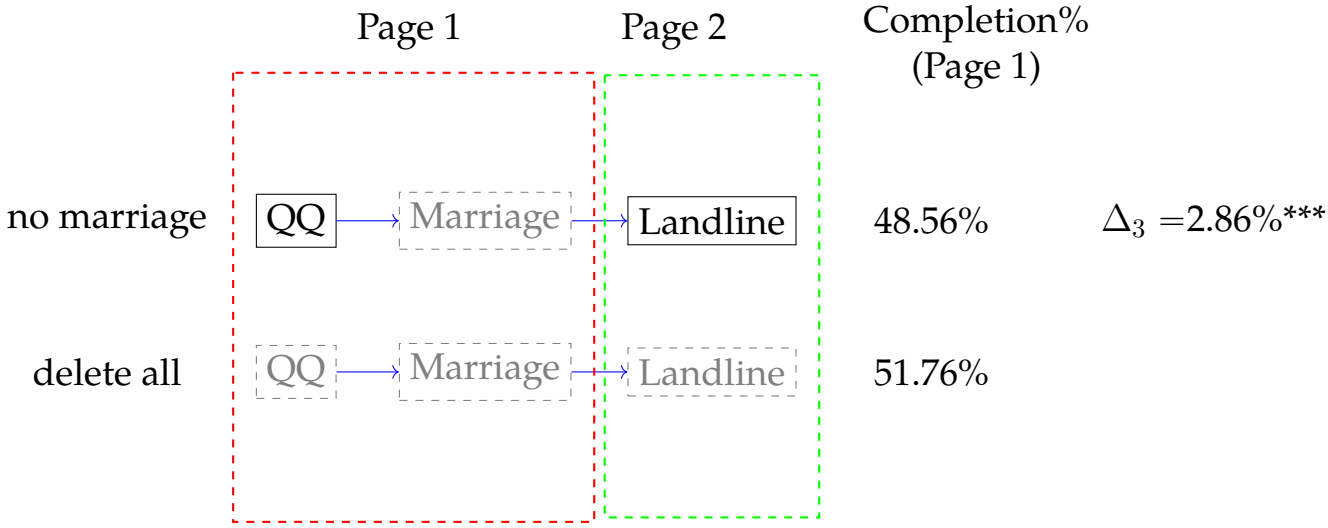
Appendix A Additional Figures

Figure A1: The Treatment Effect of the Social Network ID Question

(a) comparing "no QQ" to control or "no landline"



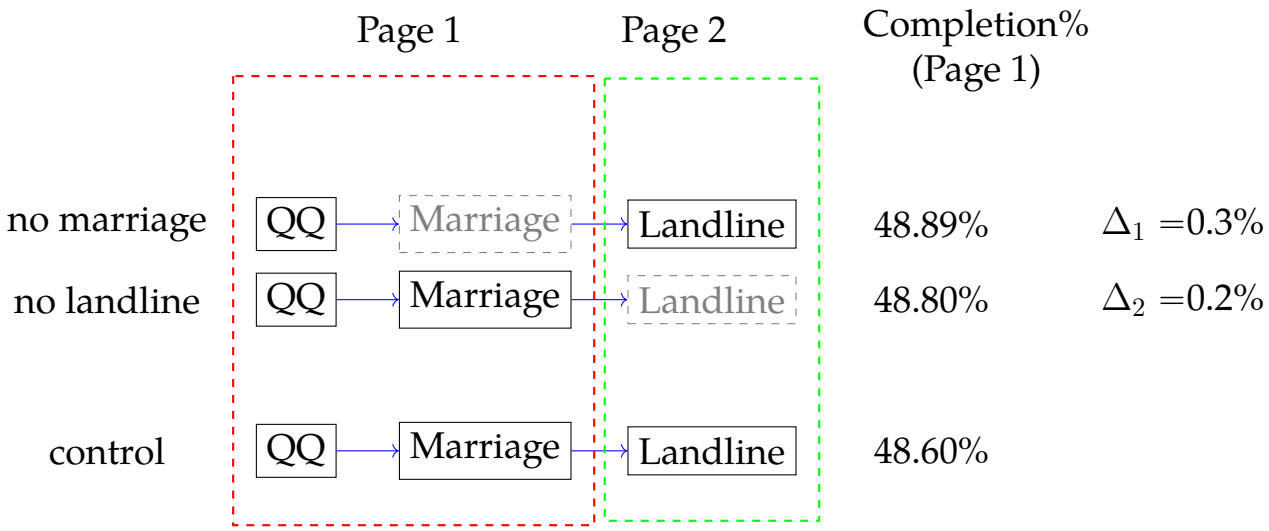
(b) comparing "no marriage" to "delete all"



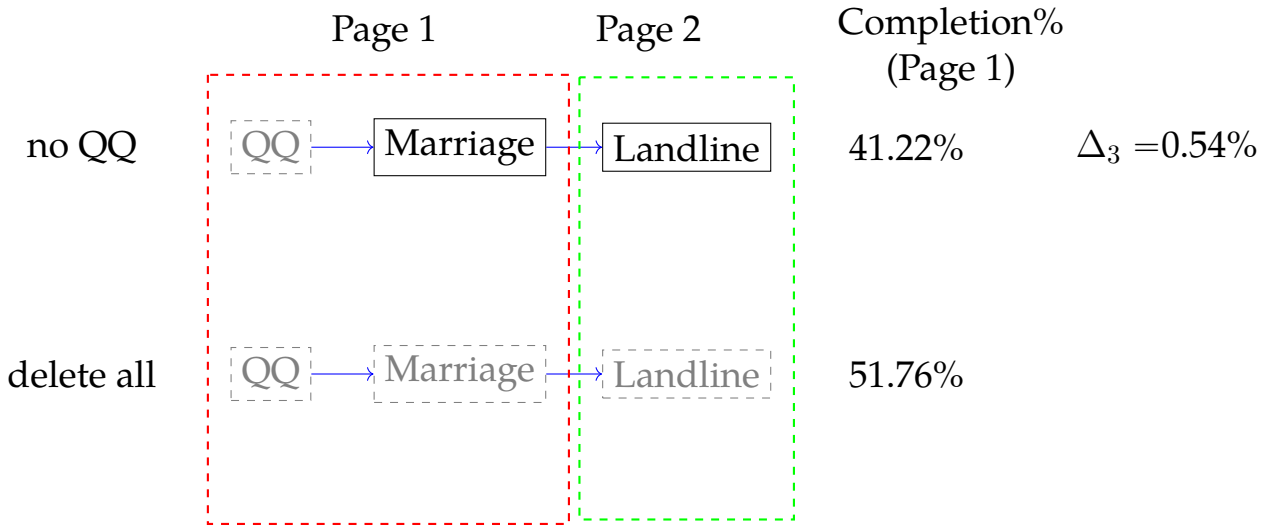
Notes: This figure presents the treatment effect of removing the social network ID question on page-1 completion rates, which can be obtained using three pairs of comparisons: "no QQ" versus "control", "no QQ" versus "landline" and "no marriage" versus "delete all".

Figure A2: The Treatment Effect of the Marriage Question

(a) comparing "no marriage" to "control" or "no landline"



(b) comparing "no QQ" to "delete all"



Notes: This figure presents the treatment effect of the marriage question on page-1 completion rates, which can be obtained using three pairs of comparisons, i.e., "no marriage" versus "control", "no marriage" versus "landline" and "no QQ" versus "delete all".

Appendix B Additional Tables

Table B1: Loan Characteristics: Platform Versus Industry

	Platform	P2P Industry	Credit Card
loan size (in RMB)	4,000	18,000	25,000
maturity	9	11.5	-
return	10%	9.62%	-
default	2%	1-5%	1.5%
cost of borrowing	>24 %	>24%	13-18%

Notes: This table compares characteristics of P2P loans originated on the platform, to loans in the entire industry and to credit card loans. The summary statistics include average loan size, maturity, annualized return and default rate, and are calculated for the first quarter of 2018. The statistics on P2P loans on the platform is from the data used in this study, while statistics on P2P industry and credit card loans are collected manually by the author from public sources including a report published by Lufax.com (http://blog.lendit.com/wp-content/uploads/2015/04/Lufax-white-paper-Chinese-P2P-Market.pdf). All statistics for the platform are reduced to the nearest round number.

Table B2: Application Speed by Age, Gender, Mobile Device Brand and Education

	Speed page 1				Speed page 2			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
old	0.99***				0.34***			
	(30.80)				(23.56)			
female		-0.86***				-0.10***		
		(-22.52)				(-5.99)		
iOS			-0.18***				-0.43***	
			(-3.73)				(-18.48)	
bachelor				-0.01				-0.38***
				(-0.36)				(-25.87)
Constant	6.87***	7.56***	9.76***	7.35***	1.75***	1.93***	2.03***	2.06***
	(301.38)	(415.82)	(387.73)	(368.53)	(173.70)	(232.98)	(156.06)	(221.72)
Observations	131,213	131,213	58,038	129,757	81,205	81,205	42,819	81,414
R^2	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.01

Notes: This table presents application speed for different demographic groups. The dependent variables in the first and last three columns are the logarithm of the number of minutes spent on page 1 and 2, respectively. The omitted reference groups for age, gender and mobile device type are male, age 50+ and other brands, respectively.