Bank Competition and Household Privacy in a Digital Payment Monopoly*

Itai Agur
IMF, iagur@imf.org

Anil Ari
IMF, aari@imf.org

Giovanni Dell'Ariccia
IMF, gdellariccia@imf.org

August 6, 2024

Abstract

Lenders can exploit households' payment data to infer their creditworthiness. When households value privacy, they then face a tradeoff between protecting such privacy and attaining better credit conditions. We study how introducing an informationally more intrusive digital payment vehicle affects households' cash use, credit access, and welfare. A tech monopolist controls the intrusiveness of the new payment method and manipulates information asymmetries among households and oligopolistic banks to extract data contracts that are more lucrative than lending on its own. The laissez-faire equilibrium entails a digital payment vehicle that is more intrusive than socially optimal, providing a rationale for regulation.

Keywords: Privacy; Financial intermediation; BigTech; Data regulation.

JEL codes: D82, E41, G21, G28

*We would like to thank Philip Bergmann, Yuteng Cheng, Rodney Garratt, Florian Heider, Peter Hoffmann, Jing Huang, Charles Kahn, Artashes Karapetyan, Michael Junho Lee, Joseph Stiglitz, Sergio Vicente, and Brandon Tan for discussions and audiences at the 2024 AEA / IBEFA Annual Meetings, 2024 Bank of Portugal - CEPR Conference on Financial Intermediation, 2024 Bundesbank and Bank of Canada Conference on Payments and Securities Settlement, 2024 Federal Reserve Board Conference on the Interconnectedness of Financial Systems, 2023 Swiss Winter Conference on Financial Intermediation, 2023 Luiss-CEPR Finance Workshop, 2023 Oxford Financial Intermediation Theory Conference, 2023 CEPR-Bocconi Conference on The Future of Payments and Digital Assets, 2023 IMF Annual Macro-Financial Conference, the BIS, Collegio Carlo Alberto Turin, and an internal IMF seminar for helpful comments. The views expressed are those of the authors only and do not represent the views of the IMF, its Executive Board or IMF management. Postal address (all authors): IMF, 700 19th St NW, 20431, Washington DC.

1 Introduction

New payment technologies are making it increasingly easy to collect information about consumer behavior, while network externalities limit the number of viable payment providers and endow them with significant market power in the collection of personal data.[1] This data can be parsed to infer a variety of personal characteristics that corporations can profitably exploit, including by deriving credit ratings based on an individual's spending behavior. In credit markets, such use of payment data can reduce informational asymmetries and unlock market segments where adverse selection may otherwise hinder the provision of credit. There is growing evidence on the expansion of credit access associated with new means of payment, including those issued by major ecommerce or social media platforms known as BigTechs (Beck et al., 2022; Doerr et al., 2024; Frost et al., 2020).[2] But this may come at the expense of personal privacy. This is an acute concern for many households, as documented in empirical literature discussed in the next section. Further, market power in data collection may give payment providers an edge that alters the competitive landscape in banking markets.

How will markets navigate tradeoffs between credit provision and privacy? Who will be the winners and losers of the introduction of new forms of data collection? Does the rise of these new technologies call for government intervention? In this paper, we shed light on these questions by building a model in which a digital payment issuer competes against an anonymous form of payment, cash. This presents households with a choice between a payment method that precludes information collection and another for which the issuer of the digital payment vehicle determines the extent of informational intrusiveness, i.e., what consumer data to collect and retain. For instance, for each transaction the only data collected and retained could be the vendor and the amount spent, similar to what happens with credit

[1] Payment data is increasingly migrating to the platforms of leading information technology companies, some of which also encourage payment data migration beyond their platforms. E.g., Amazon rewards consumers that share receipts of purchases outside Amazon (https://panel.amazon.com/); Alibaba and WeChat provide widely used means of payment in China; and Meta has explored digital currency issuance.

[2] The case of payment and credit provision by Alibaba in China has been most extensively studied. See, for example, Hau et al. (2019), Huang et al. (2020), and Ouyang (2022).

cards, or instead reach the level of the specific items bought. Further, the data collector could choose what segments of payment data to connect to other sources of information about a specific individual: e.g., payment data could be linked to users' activity on social media accounts or ecommerce platforms, as has been the case for, respectively, WeChat Pay and Alipay in China. Alternatively, the payment provider could introduce a digital currency with customizable privacy features. In the model, intrusiveness is a choice variable for the payment provider that determines the probability with which the digital payment vehicle reveals the creditworthiness of its users.

Households are consumer-entrepreneurs that differ in their creditworthiness and their care for privacy, both of which are private information. They are endowed with a project but need external financing in order to realize it. The payment provider sells the information it collects about consumer behavior to banks, which can exploit this data to reduce adverse selection when competing for loans to borrowers in an imperfectly competitive market. Spatial differentiation of households and banks provides individual banks with market power, as monitoring costs are lower for nearby borrowers.

When optimizing households choose between using cash and the digital payment vehicle, they take into account how being observed by the digital payment issuer affects their personal cost in terms of loss of privacy and the potential revelation of their creditworthiness (and thereby their borrowing conditions). Similarly, the payment provider takes this optimizing behavior into account when determining the intrusiveness of the payment vehicle. How much data the provider manages to collect depends on the intrusiveness of the vehicle and on the portion of consumers that chooses to use it.

The model's equilibrium delivers four main results. First, the payment provider makes the payment vehicle as intrusive as possible (leading to certain revelation of creditworthiness type for all users). The payment provider's optimal strategy stems from adverse selection, which it exploits to instigate a partially unravelling Lemons' Market. This plays on the way that households sort into means of payment. Households with low privacy costs and high

creditworthiness want to use the digital payment vehicle, while households with high privacy costs and low creditworthiness opt for cash. In between, there are households that would choose cash if everyone chose cash, but are swept along in a disclosure cascade, because high quality households' departure from cash worsens the average creditworthiness of the pool of borrowers using cash and hence raises their equilibrium loan rate. As a result of this disclosure externality, the market share of the digital payment vehicle increases with intrusiveness, despite the higher privacy cost experienced by individual users.

A second result is that the payment provider maximizes data sale revenues by offering free data access to banks outside their home market. Banks compete à la Bertrand in the credit market and (under symmetric information) a price-limit equilibrium emerges wherein each bank serves and extracts surplus from borrowers in its home market and unsuccessfully contests other markets. Hence, the home bank stands to lose the most (and thus is willing to pay the most) from being cut out of private information about borrowers' creditworthiness, and its willingness to pay for data is maximized by the threat of an asymmetric equilibrium when out-of-home-market information is free to all banks.

The third result is that the payment provider prefers to sell information to the banks rather than becoming a lender and walling off the information for its own use. This resembles, for example, the case of Paytm, a major payment provider India, which does not provide credit but instead uses its data to act as a (fee-earning) bridge between borrowers and lenders. In our model, this outcome stems from the fact that the presence of cash makes the payment provider's monopoly power partly contestable. Should the provider become a lender itself, it would use the payment information to gain monopoly power over loan markets and thus extract all the surplus from borrowers/consumers. Foreseeing this, consumers would all use cash, denying the digital provider any informational advantage versus competing lenders. Competition among banks here acts as a commitment device for the payment provider to share some of the surplus with borrowers. The ability to credibly commit to lower loan rates would assuage this outcome, but in the absence of perfect commitment devices, this

provides a counterbalance to the integration of data collection and credit provision in a data monopoly from, e.g., economies of scope.

These results come about in an economy where enough borrowers have positive NPV projects that in equilibrium every project is financed. Borrowers with negative NPV projects that choose to transact with cash get financed because, even after disclosure cascades, there remain enough borrowers with high creditworthiness and high care for privacy that also choose cash. This in turn implies that, while there will be winners (households with a low care for privacy and high creditworthiness) and losers (other households), the aggregate effects of introducing the digital payment vehicle are negative: relative to a cash-only world, aggregate output is unchanged and there are positive privacy costs. It follows that under these circumstances a social planner would always prefer a non-informative form of payment.

To speak to the potential benefits of an information producing payment vehicle, we consider an economy where under certain circumstances information disclosure leads to greater aggregate credit access. Put simply, we consider states of the world under which the absence of information about individual borrowers leads to a market shutdown. When the loan market for undifferentiated households freezes, type differentiation generates social value. More intrusive data collection now gives households with good credit quality a way to stand out and obtain credit. This raises total credit provision and aggregate welfare.

We use this framework to draw insights about the optimal regulation of a data monopoly, which constitutes our fourth key result. Regulation is envisaged as a slow-moving policy, determined before the economic state materializes, and implemented as a constraint on the intrusiveness of the monopolist's data collection. Optimal regulation here trades off the costs of intrusiveness when credit provision is assured against its benefits during a credit crunch. We also analyze how optimal regulation responds to society's average privacy preferences and market power in lending.

Furthermore, we find that there can be a role for a second policy instrument, implemented as a tax or subsidy on households that opt for the monopolist's payment method. The reason

that optimal policy with two tools can raise welfare is that the amount of revealed household data consists of two margins: one extensive (the share of households opting for disclosure) and one intensive (the extent of revelation about such households).

We highlight the extent to which the payment data monopoly amplifies externalities in an extension wherein the digital payment system is owned by a consortium of banks rather than a monopolist. The welfare outcome is in between the monopolist and the social planner: the bank consortium neither actively exploits nor leans against the disclosure externalities.

Based on extensions, we also find that our results do not hinge on payments being the only source of credit quality data. The key assumption is that the new means of payment reveals more credit quality information, not that it is the unique source of such information. Lastly, we find that our results hold if the ability of the data monopolist to infer credit quality depends positively on the number of users of its payment system.

The remainder of the paper is organized as follows. The next section reviews the related literature. Section 3 presents the setup of the model. Section 4 derives the equilibria wherein the data monopolist sells the data to lenders and Section 5 shows why it chooses to do so instead of engaging in credit provision itself. Section 6 analyzes welfare and socially optimal policy design. Section 7 summarizes the model extensions. Proofs and additional modeling detail behind some of the extensions are provided in the Online Appendix.

2 Related Literature

Our paper builds on classical work on optimal disclosure (Grossman and Hart, 1980; Jovanovic, 1982) and relates to a growing literature on emerging financial technologies in payments and credit provision. Like us, Parlour et al. (2022) and He et al. (2023) consider borrowers that are heterogeneous in both credit quality and a second dimension (respectively, bank affinity and privacy), and data externalities among such borrowers take center stage. In Parlour et al. (2022), payment providers compete against a bank and externalities

stem from coupling data garnered from payment services with credit provision. He et al. (2023) also consider lending market competition between banks and financial-technology driven entrants, and show that household data ownership is suboptimal as individual data sharing decisions lead to aggregate credit quality inferences by lenders.[3]

Different from these studies, we consider a data monopolist that stands "upstream" and controls the flow of information, actively exploiting information asymmetries and data externalities. In essence, we model a BigTech, whereas Parlour et al. (2022) and He et al. (2023) model competing FinTechs (companies that automate financial services, but generally do not enter the market with the dominant networks that BigTechs derive from ecommerce or social media platforms).[4] Unlike the FinTechs, our BigTech determines the extent of intrusiveness of its data collection, while it competes with cash, which allows households to opt out of its data web. This lays bare the incentives of a data monopolist to play off households and banks against each other.

Several papers consider the role of a Central Bank Digital Currency (CBDC) in the preservation of privacy. In Garratt and Van Oordt (2021), the social value of CBDC derives from the fact that privacy is a public good: taking actions to protect digital privacy is costly to a consumer, while firms use data collected through payments to price discriminate future consumers. In Garratt and Lee (2021), the CBDC raises welfare compared to household ownership of their data: either policy can fend off the endogenous formation of data monopolies, but household data ownership leads to data underprovision, which hampers firms' ability to match products to consumer preferences. In Ahnert et al. (2022) and Brunnermeier and Payne (2022), a CBDC can be more than an electronic equivalent of cash, by including data

[3] In Huang (2021) banks compete with lenders that rely on data from linked ecommerce platforms while banks rely on physical collateral, leading to different borrower type specializations.

[4] Information flow on ecommerce platforms provides a different angle on BigTechs. Bouvard et al. (2022) focus on monitoring efficiency differences between an ecommerce platform and banks in the competition for loans towards merchants that are active on the platform. In Gambacorta et al. (2022) and Li and Pegoraro (2023) platforms have an enforcement advantage over banks, because defaulting firms can face exclusion from the platform. The broader (not credit related) literature on platform data collection and data externalities, is too extensive to review here. We refer to, e.g., Acemoglu et al. (2022), Bergemann and Bonatti (2024), Bergemann et al. (2022), and Cong and Mayer (2023) and discussions of the literature therein.

sharing features that counteract payment data monopolies. Different from us, these papers do not focus on hidden quality problems in credit markets.

Empirically, although the value of digital privacy in general is debated (Goldfarb and Que, 2023), for payments and lending specifically most studies find an important role for privacy. Doerr et al. (2023) and Tang (2023) document borrowers' willingness to pay to limit the intrusiveness of data disclosure in loan applications; and various studies find that privacy is an important determinant of payment choice.[5] Moreover, several studies report extensive heterogeneity in privacy preferences among households, in line with our assumption.[6]

3 Model

We consider an economy populated by three sets of agents. Consumer-entrepreneurs, to which we refer to as "households", form the first set. Second, there is a monopolist digital currency issuer.[7] Lenders comprise the third set of agents.[8] All agents are risk neutral. The economy is distributed across $N > 3$ connected islands. Each island hosts one lender and a continuum of households with mass 1. A lender on the same island as a borrower is called the borrower's "home lender", and borrowers on the same island as a lender are referred to as the lender's "on island" borrowers**.** All islands share the same physical currency (cash) and digital currency issuer.

[5] See , Borgonovo et al. (2021), Brits and Jonker (2023), Choi et al. (2023), Cunliffe (2023), ECB (2021), Engels et al. (2022), and Li (2023).

[6] See, e.g., Armantier et al. (2024), Bian et al. (2023), Budak et al. (2023), Collis et al. (2024), Goldfarb and Tucker (2012), Jiang et al. (2023), Lin (2022), Lin and Strulov-Shlain (2023) and Prince and Wallsten (2022).

[7] The model's setup applies to a digital payment monopolist in general but, to fix ideas, we call this company a "digital currency issuer". Its business model, based on the collection and sale of data, is akin to Meta Platforms, while the notion of a digital currency resembles the Libra and Diem initiatives. We do not model what gives rise to a monopoly here, such as network effects in means of payment (Agur et al., 2022).

[8] We here use the term "lenders" instead of banks, because the way they are modeled (Section 3.3) is general enough to encompass nonbank lenders, including FinTechs.

3.1 Households

Each household is characterized by three variables: its creditworthiness and preference for privacy, which are private information, and the island where it lives, which is publicly observable. The last is self-explanatory. Let us turn to the first two. Each household is born with an investment project that can only be brought to fruition with financing from a lender. All projects yield the same payoff, y, when successful and 0 otherwise. However, the probability that a project succeeds, q, differs across households. We refer to q as a household's credit quality or creditworthiness. Households also differ in their preference for privacy or, put differently, their perceived cost of having their consumption patterns scrutinized. Formally, each household attaches disutility φ to its loss of privacy. On each island, the mass 1 of households is uniformly distributed on a two-dimensional plane with $q \in \left[\frac{1}{2}, 1\right]$ and $\varphi \in [0, 2]$.[9] A household's creditworthiness and privacy preferences are independent of each other.

3.1.1 Household consumption

Households are born with an endowment that they use towards consumption. To avoid complicating the model excessively, we assume that this consumption is needed to meet subsistence needs (i.e., none of the endowment can be saved or used as collateral when requesting credit). However, an individual household's consumption pattern, if observed, may carry information about its creditworthiness.

Each household chooses between cash and the digital currency (DC) to pay for consumption. Cash fully protects the household's privacy. If instead the household uses the DC then with some probability, θ, the DC issuer is able to infer the household's creditworthiness.

The household experiences privacy disutility φ from being observed by the DC issuer.

[9] We choose $q \in \left[\frac{1}{2}, 1\right]$ instead of $q \in [0, 1]$, as the latter creates additional complexities (individual breakeven loan rates go to infinity as $q \to 0$, which complicates the identification of simple parameter conditions to characterize equilibria; see, e.g., Table 1). Using $q \in \left[\frac{1}{2}, 1\right]$ is without loss of generality, because all possibilities on the functioning of the loan market are covered by our framework (Section 4). Given $q \in \left[\frac{1}{2}, 1\right]$, we let $\varphi \in [0, 2]$ instead of $\varphi \in [0, 1]$ to obtain a unit mass of households on each island.

Households understand the *extent* to which their behavior is observed by the DC issuer. That extent, captured by θ, determines both the household's privacy disutility and the probability of type revelation. We assume that households are unable to predict how their consumption choices affect their credit score (i.e., households have no visibility into the algorithm used to extract creditworthiness information from consumption patterns). For a practical example, consider the case of a social media company that can choose whether to link households' payment and social media data (e.g., WeChat has derived credit ratings from linked payment and social media data). The decision to link these data is observed by households, and disliked (to varying degrees) from a privacy angle by those who choose to use DC. But the way in which social media chats interact with payment data is likely poorly understood by households, and they therefore neither consume nor chat strategically.

3.1.2 Household borrowing and payoff

Households are protected by limited liability. They may borrow from any lender. The expected payoff for a household of type (q, φ), considering both the potential gains from borrowing and the costs of privacy, can then be written as

$$u(q, \varphi) = q \max\{y - R, 0\} - \alpha\varphi\Theta \tag{1}$$

where the term $q \max\{y - R, 0\}$ represents the expected benefit from the household's project under limited liability, and $-\alpha\varphi\Theta$ is privacy costs. In $q \max\{y - R, 0\}$, R is the gross loan interest rate charged; and the max operator reflects the assumed limited liability protection.[10] In $-\alpha\varphi\Theta$, $\alpha > 0$ is the relative weight on privacy preferences and Θ represents the probability that the household's credit quality will become revealed: $\Theta = \theta$ if the household chooses DC and $\Theta = 0$ if it chooses cash.

For simplicity, the baseline model does not incorporate an additional term to represent

[10] The expression is of the form $q \max\{., 0\}$ because even a successful project could, in principle, violate the limited liability constraint when R is large enough. This will not occur in equilibrium, however, and the max operator therefore becomes moot by optimal play in Section 4.

an ease-of-transactions benefit of DC relative to cash. We note, however, that cash is merely one example of an alternative payment instrument. For example, our modeled alternative to the DC could equally well be a privacy-preserving crypto asset, which may not have a transaction disadvantage relative to the DC.

3.2 Digital currency issuer

The DC issuer is in control of the design of the digital currency. Formally, it chooses the probability, θ, with which it will be able to learn the creditworthiness of individual households using the DC. The DC issuer is a monopolist in the provision of digital payments and hence also sets the price at which its collected digital payment data are made available to lenders. Note that in this section we focus on the case of a DC issuer that puts these data up for sale to lenders and does itself not engage in lending.[11] Section 5 considers the possibility that the DC issuer would instead provide credit, either directly or through a subsidiary lender, and shows that this will never emerge as an equilibrium.

The DC issuer may charge lenders differentiated fees for information on borrowers on and off their islands. Here, Ω_{ij} denotes the data access fee that the digital currency provider charges lender i for access to data on households on island j.[12] We define $D_{ij}(\theta, \Omega_{ij})$ as lender i's demand function for data on households on island j. This demand function will take value 0 (do not purchase) or 1 (purchase) depending on the pricing of the data access, Ω_{ij}, and the value that a lender derives from the data. How much a lender is willing to spend on data gathered from DC users, will depend on how much borrower data the DC issuer

[11] The DC issuer can market credit quality data but not a list of which households open a DC account. The latter would lead to two separate pools of unrevealed households: cash users and DC users that remain unrevealed with probability $1-\theta$, opening the model to mimicking strategies (cash preferring households may gamble on DC use, hoping that they end up among the unrevealed DC users), considerably complicating the analysis. However, allowing households to open *unused* DC accounts would restore our baseline setup. Opening such an account would directly sort a household into the pool of unrevealed DC users, which have a higher higher average credit quality than cash users (see Section 4) . All cash users would do so, as it would come without privacy costs. This implies the same single pool of unrevealed households as in our baseline model and all results would be identical.

[12] The numbering of lenders by i and household island locations by j is symmetric. E.g., Ω_{11} is the data access fee for lender 1 with respect to households that have lender 1 as their home lender; Ω_{1j} for $j > 1$ are the fees for lender 1 to access the credit quality data of households that do not have it as their home lender.

obtained, which in turn relates to the intrusiveness of the DC. Therefore, D_{ij} is written as a function of (θ, Ω_{ij}). The DC issuer's objective is to maximize data fee revenue by optimally choosing θ and Ω_{ij}. This can be written as

$$\max_{\theta \in [0,1], \Omega_{ij}} \sum_{i \in [1,N]} \sum_{j \in [1,N]} \Omega_{ij} D_{ij} \left(\theta, \Omega_{ij}\right) \tag{2}$$

3.3 Lenders

Lenders in our model face an infinitely elastic supply of funding at cost $c^f \geq 1$ and engage in Bertrand competition for loans to households.[13] They have two sources of differentiation. The first is their access to information about households' credit quality. If one lender purchases detailed data on a set of households, while other lenders do not, the purchasing lender can charge differentiated loan rates to attract the subset of households with high credit quality, leaving other lenders a pool of lower credit quality borrowers. In our baseline model, the DC's issuer's privately owned payment data is the only source of credit quality information, but Section 7.2 considers lender access to additional public sources of partial information on household credit quality, such as from a credit registry.

The second source of differentiation stems from the cost of monitoring borrowers. Unmonitored borrowers are assumed to always run off with the loan principal. Monitoring prevents loan theft with certainty but is costly. Monitoring a borrower on the same island as the lender costs c_h^m. This is the monitoring cost for a home lender. Monitoring at distance is more costly: away (off-island) lenders incur a cost of $c_a^m = c_h^m + \tau$, where $\tau > 0$ parameterizes the monitoring cost advantage of home lenders. The relation between physical distance and monitoring costs is well-established in the literature (Agarwal and Hauswald, 2010; Heitz et al., 2023). To simplify notation, we let $c = c^f + c_h^m$ represent the combined cost of loan provision and monitoring to a home lender and therefore $c + \tau$ is that cost for away lenders. Since borrowers always abscond when unmonitored, lenders will always choose to pay the

[13] To guarantee that limit pricing can be sustained as a Nash equilibrium in Section 4, we additionally assume a household breakeven preference in favor of the home lender.

monitoring cost. We therefore do not model this as a separate decision stage in the game, but rather incorporate the cost directly as part of the cost of providing a loan. I.e., providing a loan costs c for a home lender and $c+\tau$ for an away lender.

We can express lender i's expected profits in general form (with closed forms following as part of the derivations in Section 4):

$$\begin{aligned} &\left(q_{ik}R_{ik}\left(q_{ik}\right)-c\right)m_{ih}+\left(q_{ik}R_{ik}\left(q_{ik}\right)-\left(c+\tau\right)\right)m_{ia} \\ &\left(E\left[q\middle|u\right]R_{iu}-c\right)\upsilon_{ih}+\left(E\left[q\middle|u\right]R_{iu}-\left(c+\tau\right)\right)\upsilon_{ia}-\sum_{j\in[1,N]}\Omega_{ij}D_{ij}\left(\theta,\Omega_{ij}\right) \end{aligned} \tag{3}$$

Lender i's profit consists of three parts. The first is the expected profit on revealed borrowers whose credit type the lender has obtained. Here, m_{ih} and m_{ia} represent, respectively, the masses of on-island and off-island revealed households who choose to borrow from lender i. A household, k, whose credit quality data the lender has obtained, and who has chosen to borrow from lender i, is denoted by ik. A lender's expected profit on a loan with customized loan rate $R_{ik}\left(q_{ik}\right)$, is $q_{ik}R_{ik}\left(q_{ik}\right)-c$ if the borrower is on the same island as the lender and $q_{ik}R_{ik}\left(q_{ik}\right)-\left(c+\tau\right)$ if the borrower is on another island. The second term is the expected profit that the lender makes from unrevealed borrowers. The masses of such borrowers on-island and off-island, respectively, are denoted by υ_{ih} and υ_{ia}. For every loan that the lender provides to an unrevealed borrower, the expected profit is $E\left[q\middle|u\right]R_{iu}-c$ and $E\left[q\middle|u\right]R_{iu}-\left(c+\tau\right)$, respectively, on-island and off-island. Here, R_{iu} is the pooled loan rate that lender i charges unrevealed borrowers, and $E\left[q\middle|u\right]$ represents the expected credit quality of households conditional on being unrevealed: iu denotes the unrevealed households that choose to borrow from lender i. The third term is the lender's cost if it chooses to purchase data on revealed borrowers from the DC issuer, as per (2).

3.4 Parameter conditions

We include three parameter conditions. First, we let

$$c < y \leq 3 \tag{4}$$

where $c < y$ ensures that the highest quality borrower ($q = 1$) always has a positive NPV project. This is a necessary condition for credit provision to take place. Moreover, setting an upper bound on y proves useful in some of our proofs where we need to calculate upper and lower bounds within allowed parameter ranges.[14] Second, we let

$$\tau < \frac{1}{6}c \tag{5}$$

which means that the added cost of monitoring at distance (off-island) is at most one-sixth of the combined other costs that lenders face in providing credit (namely, the cost of their own funding and the baseline cost of monitoring that is independent of distance). The reason for this condition is that key mechanisms of our model rely on the competition between lenders and when home lender market power becomes too strong, such mechanisms break down. Third, we assume that

$$\frac{c+\tau}{9} < \alpha < \frac{c+\tau}{6} \tag{6}$$

where $\alpha < \frac{1}{6}(c+\tau)$ is necessary and sufficient to ensure that households with the highest credit quality ($q = 1$) will always want to disclose their type. This helps preserve tractability. Lastly, a lower bound on α facilitates a focus on the setting where the Lemons' Market dynamics are bounded by the sufficiently high care for privacy that some high quality households have. This will allow the use of different means of payment by different households to be sustained in equilibrium.

[14] We note that $y \leq 3$ is, quantitatively, a very wide bound, as $y = 3$ implies a 300 percent project return.

3.5 Timing

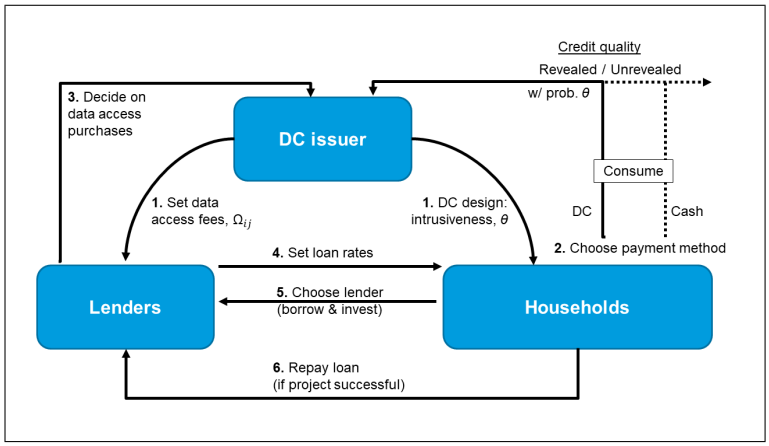


Figure 1: Timing and actions.

The timing of agents' actions and the realization of events is shown in Figure 1. At Stage 1, households are born with an endowment and a project opportunity, and the DC provider chooses the DC design, θ, and data access fees, Ω_{ij}. At Stage 2, households observe θ and decide whether to transact with DC or cash and consume, after which DC users' creditworthiness becomes revealed with probability θ. At Stage 3, lenders decide on data access purchases and, subsequently, each lender announces its loan rates at Stage 4 (namely, individualized loan rates for revealed households on which data has been purchased and a single loan rate for other households). Based on these loan rates, in turn, households decide from which lender to borrow to finance their projects at Stage 5. Lastly, at Stage 6, project returns materialize and households with successful projects repay their lenders.

4 Equilibria

Below we derive the subgame-perfect equilibrium of the game laid out in Section 3.5 in separate propositions that relate to different credit market outcomes. These propositions are defined for cases shown in Table 1. When y is well above c, households will have positive NPV projects ($qy > c$) unless their probability of success q is very low. Referring to borrowers

with positive (negative) NPV projects as good (bad) borrowers, the share of bad borrowers is small when y is high. The equilibrium that ensues from Table 1's "High y" case is discussed in Section 4.1. Conversely, when the payoff on a successful project is only barely above c, then households will have negative NPV projects ($qy < c$) unless their credit quality is high, leading to a large share of bad borrowers on the loan market. Section 4.2 analyzes the outcomes of Table 1's "Low y" case. The "Intermediate y" case is a direct extension of the other two cases: its equilibrium is derived and discussed in the Online Appendix.

Table 1: Parameterizing the share of bad borrowers

Case	Share of bad (NPV < 0) borrowers	Value of y
High y	Small	$y > 2(c+\tau)$
Medium y	Intermediate	$y \in \left(\frac{4}{3}c, 2(c+\tau)\right)$
Low y	Large	$y \in \left(c, \frac{4}{3}c\right)$

4.1 Equilibrium with few bad borrowers

Proposition 1 *When $y > 2(c+\tau)$, all households borrow from their respective home lenders at Stage 5. At Stage 4, home lenders offer differentiated loan rates*

$$R_k(q) = \frac{c+\tau}{q} \tag{7}$$

to revealed borrowers and a single loan rate

$$R_u = \frac{c+\tau}{E[q|u]} \tag{8}$$

to unrevealed borrowers.[15] *All lenders obtain data access at Stage 3. At Stage 2, households sort into DC use if*

$$\varphi \leq \frac{c+\tau}{\alpha}\left(\frac{q}{E[q|u]} - 1\right) \tag{9}$$

[15] We drop the subscripts i in (7) and (8), given home lender symmetry (each lender is a home lender to a given set of households).

where

$$E\left[q\right|u] = \frac{2\theta - 1 + \sqrt{1 - \theta + \frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2 (4\theta - 3)\,\theta}}{2\theta\left(1 - \frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2\right)} \tag{10}$$

and choose cash otherwise. The mass of revealed borrowers (per island) is given by

$$m = \frac{2\theta\left((c+\tau)^2 - \frac{4}{3}\alpha^2\right) - (c+\tau)(\alpha + c + \tau)\left((2\theta - 1) + \sqrt{1 - \theta + \frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2 (4\theta - 3)\,\theta}\right)}{(c+\tau)^2 - \frac{4}{3}\alpha^2} \tag{11}$$

and, at Stage 1, the DC issuer charges each lender data access fees $\Omega_{ii} = m\tau$ for information about its on-island households, and offers each lender free information about off-island households ($\Omega_{ij} = 0$ when $j \neq i$). The DC issuer optimally sets $\theta = 1$, which maximizes m and thereby data fee revenues.

Proof. See Appendix A.1. ■

While Appendix A.1 contains the full proof, we here describe the main mechanisms at play. We follow the decision stages (Stages 1-5) of the game in backward order, discussing at each stage why deviation from the equilibrium is suboptimal for the players.

Stage 5 requires little description: households borrow from the lender that offers them the lowest loan rate.

Stage 4. When lenders are symmetrically informed (or uninformed), meaning they all bought (or refused to buy) information from the DC issuer at Stage 3, Stage 4 is a standard Bertrand game which admits a unique limit-pricing equilibrium. When all lenders are informed, on each island, competition among symmetric off-island lenders pushes their expected profits from on-island borrowers to zero, leading them to offer revealed borrowers a rate conditional on their type defined by (7) and unrevealed borrowers the single rate in (8). The on-island lender matches these rates and (due to the assumed breakeven rule) serves all borrowers on the island, making positive profits off them due to its positional rent.

For the case in which no lender is informed, the same logic applies, with off-island lenders offering all borrowers the rate shown in (8). Again, all borrowers are served by the on-island

lender that obtains positive profits.

Now consider the asymmetric case in which a lender does not buy data about its on-island borrowers, while other lenders do obtain the data on these borrowers. The off-island lenders will Bertrand compete and, as before, this will push rates down to their breakeven levels in (7) and (8). The uninformed on-island lender will be unable to discriminate across credit types and will have to charge a pooling rate, which we call R_{dev}, to all on-island borrowers (see Appendix A.1 for a formal treatment). It is immediate that because of the adverse selection on revealed borrowers, its profits (excluding data access fees) will decrease relative to the case in which the home lender is informed. We denote by $\Delta\Pi_k$ the difference between the home lender profits (excluding data access fees) when informed and when not.

Stage 3. At this stage, lenders decide whether or not to buy the data provided by the DC issuer. It is immediate from the result at Stage 4 that it will be optimal for lenders to buy data about their on-island borrowers if and only if the price of data access is (weakly) below the difference in profits between being informed and uninformed. That is, lenders will buy the data when $\Omega_{ii} \leq \Delta\Pi_k$. The results at Stage 4 imply that, in equilibrium, lenders cannot make any profits from data about their respective off-island borrowers.[16] Hence, lenders will never accept to pay a positive price for data about off-island borrowers.

Stage 2. At the start of Stage 2, households observe DC design, θ, and the data access fees set by the DC issuer at Stage 1. They are then able to foresee the subgame equilibrium prevailing at Stage 4 and hence the expected loan rates and payoffs from using cash versus DC. Based on its q and φ, every household compares the payoffs in (1) from using cash versus the DC. Solving this problem leads to the sorting condition in (9). Note that since expected loan rates for unrevealed borrowers depend on the credit quality of the pool of households using cash and DC, a household's payment choice is affected by other households' payment choices. A comparison to a cash-only (or, equivalently, $\theta = 0$) setting is useful to highlight these interactions. Without DC, all households use cash so that $E[q|u] = E[q] = \frac{3}{4}$ and

[16] Unless they obtain a monopoly contract from the DC issuer, a case that cannot be supported in equilibrium as discussed in Section 5.

$R_u = \frac{4}{3}(c+\tau)$. At this loan rate, some households will find it attractive to use a $\theta > 0$ DC, when given that option. This set of households is represented by the households to right of and below the dashed line in Figure 2. Since these households are located in the corner with the highest credit quality, their choice to use DC lowers $E[q|u]$ and raises R_u, as some of them become revealed. With a lower $E[q|u]$, more households find that the condition in (9) holds, and choose DC over cash. These households are near (to the left of) the threshold given by the dashed line, meaning that their credit quality is high among cash users. Their choice to use the DC in turn further lowers $E[q|u]$, and so on.

The equilibrium in Proposition 1 documents the ultimate settle point of such a process, as represented by the unbroken line in Figure 2. But behind this equilibrium lies a cascade of disclosure: households that only choose to use the DC because other, higher credit quality households choose to do so. Households in between the dashed and unbroken lines Figure 2 can be worse off than if the DC did not exist (i.e., in the cash-only economy), which highlights the public good nature of privacy here. The implications for individual and aggregate welfare are further explored in Section 6.

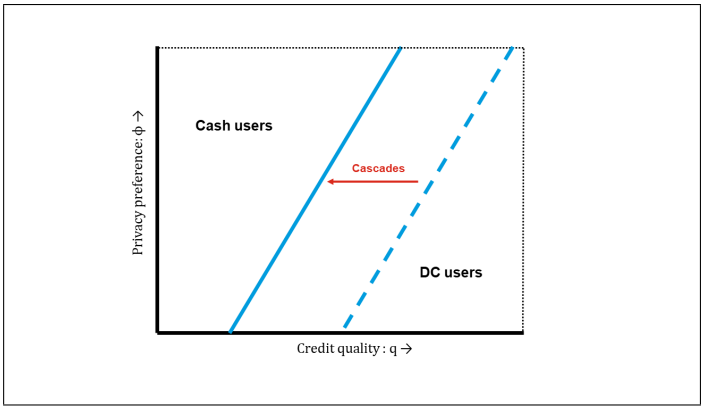


Figure 2: Payment choice and disclosure cascades.

Households at the left end of Figure 2 are bad borrowers. Bad borrowers self-select into cash use, because they know that credit quality revelation will preclude them from receiving loans. However, when $y > 2(c+\tau)$, there are always enough good borrowers on the market,

including good borrowers with high privacy preferences that choose to use cash, so that the loan market for unrevealed borrowers remains open.

Stage 1. The DC issuer sets θ and data access fees to maximize its profits. Based on the analysis for Stages 3 and 4, this boils down to maximizing the difference in equilibrium profits between informed and uninformed lenders. To implement this strategy, the DC issuer offers free data access to each lender on its off-island households (as discussed, lenders are unwilling to pay a positive price for these data). This maximizes adverse selection among on-island borrowers should a lender choose not to buy the associated data. The DC issuer charges each lender the highest fee that it is willing to pay for data on its on-island borrowers, which is $m\tau$. To earn the most off this fee, the DC issuer sets $\theta = 1$, which makes the set of revealed households, m, as large as possible (augmenting adverse selection for uninformed lenders). A higher θ not only increases the odds of revelation per DC user but, as Appendix A.1 shows, also increases the mass of DC users. The pool of the unrevealed consists of two types of households - cash users and unrevealed DC users - and the worst quality unrevealed DC user is always of better credit quality than the average cash user. When θ increases, the mass of unrevealed DC users shrinks, $E[q|\, u]$ declines and this induces more households to opt for DC use, despite the higher privacy cost associated with DC use.

4.2 Equilibrium with many bad borrowers

Proposition 2 *When $y \in \left(c, \frac{4}{3}c\right)$, then if $\tau > y - c$, no credit provision takes place and all households use cash. Instead, when $\tau < y - c$, households that choose to borrow do so from their respective home lenders at Stage 5. At Stage 4, only revealed households are offered loans, with loan rates given by (7). At Stage 3, lenders obtain data access from the DC issuer. At Stage 2, households with*

$$q \geq \frac{\alpha\varphi + c + \tau}{y} \tag{12}$$

choose to use the DC and others use cash. At Stage 1, the DC issuer optimally sets $\theta = 1$*, while offering each lender free data on its off-island borrowers and charging home lenders data access fee* $\Omega_{ii} = m\tau$ *where*

$$m = \begin{cases} \frac{\theta}{2\alpha y}\left(y - c - \tau\right)^2 & \text{if } y < 2\alpha + c + \tau \\ 2\theta\left(1 - \frac{\alpha + c + \tau}{y}\right) & \text{if } y > 2\alpha + c + \tau \end{cases} \tag{13}$$

Proof. See Appendix A.2. ■

Stage 5. As in as in Section 4.1, a household borrows from the lender that offers it the lowest loan rate.

Stage 4. Lenders never make loan offers to unrevealed households. When $y \in \left(c, \frac{4}{3}c\right)$, no lender could break even on lending to such households, because there are too many bad borrowers among them. Among revealed households, there are two groups. For households with $q > \frac{c+\tau}{y}$, competition among lenders establishes limit pricing in line with equation (7). Instead, households with $\frac{c}{y} < q < \frac{c+\tau}{y}$ can only be made loan offers above a lender's breakeven rate if they borrow on-island. Hence, the home lender is a monopolist towards them, charging a loan rate that takes up the full return on a successful project, y.

Stage 3. At this stage, if the home lender refuses to buy the data on revealed households and offers a single loan rate that attracts some revealed households, then all the unrevealed households will borrow from the home lender too. The home lender cannot distinguish between the household types and, given $y \in \left(c, \frac{4}{3}c\right)$, there are too many bad borrowers among the unrevealed, which leads to a loss making loan portfolio. The home lender is therefore willing to purchase data access at any fee below what it can earn on revealed households. However, lenders are not willing to pay a positive fee for data on households located on other islands, because they foresee zero profit on such households at Stage 4.

Stage 2. Households with $q < \frac{c+\tau}{y}$ realize that at Stage 4 they cannot obtain a loan rate that leaves them with project value and therefore choose cash from the angle of privacy preservation. When $\tau > y - c$, this applies to all households and therefore everyone uses

cash and no credit provision takes place. When $\tau < y - c$, there are households with $q > \frac{c+\tau}{y}$ and, among such households, those that satisfy the condition in (12) choose DC.

Stage 1. The DC issuer optimally provides free data access on off-island households. It needs the loan market to be competitive for some households to use its payment instrument (provided $\tau < y - c$). The DC issuer charges home lenders a fee (marginally below) the full profits made on revealed households, which it knows home lenders will accept at Stage 3. The DC issuer optimizes these data sale revenues by maximizing the mass of revealed borrowers and this is achieved by setting $\theta = 1$ (as apparent from (13)).

For the remainder of the paper, we let $\tau < y - c$, since we are interested in cases where credit provision can take place.[17]

5 Data usage strategies

So far, we have focused on a DC issuer that markets its data to lenders. This section considers a DC issuer that uses its data advantage to engage in lending on its own, without selling data to other lenders. The setup has an identical structure of lenders as in the baseline, with one lender based on each island. We assume that, in addition, the DC issuer has a branch on every island. Thus, if the DC issuer provides credit, its cost of monitoring borrowers is the same as an on-island lender. The timing of the game is as described in Section 3.5, except that at Stage 1, the DC issuer only sets θ but does not set data access fees; Stage 3 on lender data purchases is inactive (since no data are put up for sale); and at Stage 4, both the DC issuer and all lenders set loan rates towards households.

This setup leads to a stark outcome, formalized in Proposition 3: all households choose cash and the DC issuer makes no profit. The reason is that, regardless of its q, a revealed borrower never obtains a lower borrowing cost than unrevealed borrowers. Since DC comes with a privacy cost while cash does not, cash is then preferred.

[17] A sufficient condition is to tighten $y > c$ from in (4) to $y > \frac{7}{6}c$, because $y > \frac{7}{6}c$ together with $\tau < \frac{1}{6}c$ from (5) imply $\tau < y - c$. This sufficient condition effectively tightens the Low y case in Table 1 to $y \in \left(\frac{7}{6}c, \frac{4}{3}c\right)$.

Proposition 3 *The DC issuer makes zero profit if it provides loans rather than selling data access to lenders.*

Proof. See Appendix A.3. ■

Only the DC issuer could offer a differentiated loan rate that benefits the highest quality borrowers, since other lenders are uninformed, but the DC issuer lacks the incentive to do so. At Stage 4, revealed households have nowhere else to obtain differentiated loan rates. In the "High y" case, the DC issuer offers $\varepsilon \to 0$ below the undifferentiated loan rate; in the "Low y" case, it offers $R_k(q) = y$ since no other lender can provide loans. In either case, households foresee at Stage 2 that there is no benefit to choosing DC.

When everyone uses cash, the DC issuer has no advantage on the loan market. In the "High y" case, the DC issuer competes on loan rates against the on-island lender on every island, so that the equilibrium entails breakeven loan rates. In the "Low y" case no credit provision takes place at all. In either case, the DC issuer makes no profit.

The fundamental problem for the DC issuer here is the ability of households to opt out before being charged monopolistic loan rates. Unmodeled factors, like economies of scope, may plead for the integration of data collection and credit provision. Nevertheless, the notion that there is a benefit for the data monopolist to be a "puppet master", who manipulates information asymmetries from above but does not directly get involved in credit markets, is powerful. So is the absence of a fully credible way to precommit not to exploit the market power inherent in a data monopoly, after households have given up their data (Liu et al., 2023). Market power in lending is a double-edged sword in a setting like this: it could facilitate higher loan revenues, but only if enough households stay put and the market power does not compel too many of them to opt out of disclosure altogether.[18]

[18] The DC issuer faces a similar problem under a different data usage strategy, where it sells exclusive data access to the highest bidding lender. Any lender that becomes the sole proprietor of the payment data has no incentive to offer revealed borrowers customized loan rates below the undifferentiated loan rate. Foreseeing this at Stage 2, all households choose cash. At Stage 1, the DC issuer then knows that exclusive data contracts will lead to zero data revenues. Hence, it will choose the non-exclusive sales strategy outlined in Section 4, rather than exclusive sales to one lender.

6 Policy analysis

We now draw insights from our model for the analysis and design of socially optimal policy. In particular, we consider the role of regulation to constrain the DC's intrusiveness.

To enable the welfare analysis, we extend our model by adding two more stages to the game described in Section 3.5. First, we assume that the policy maker determines regulation at the beginning of the game, which represents the notion that regulation is a long-term policy that is infrequently revised. We denote socially optimal DC design by $\overline{\theta}$ and implement regulation as a constraint on the intrusiveness of data collection.[19] That is, the DC issuer now faces a constrained design choice such that $\theta \in \left[0, \overline{\theta}\right]$. Second, we let y represent an uncertain aggregate state, exogenously determined in a third stage: both DC policy and DC design are set long-term, before y is known (and the remaining stages, namely 2-6 in Section 3.5, now become Stages 4-8). With probability $\gamma \in [0, 1]$, y takes the value y_h and the good economic state of Proposition 1 materializes. With probability $(1 - \gamma)$, a bad state, y_l, occurs, in which credit risks are profound and there are relatively many bad borrowers, as in Proposition 2. That is, we let y_h represent "High y" in Table 1, with $y_h > 2(c + \tau)$, while y_l represents "Low y" in Table 1 with $y_l \in \left(c, \frac{4}{3}c\right)$. We further let $\tau < y - c$, to ensure that credit provision does take place in the "Low y" case, as discussed in Section 4.2.

6.1 Welfare expression

We define aggregate social welfare as the sum of the expected payoffs of all households and the total profits of lenders and the DC issuer. Welfare is then determined by two factors only. The first is the value added of the projects that receive funding. From Propositions 1 and 2, in equilibrium households borrow from home lenders, so that a project's expected NPV is $qy - c$. Some of this value added accrues to households, through positive net returns on projects, and some of the value added is captured as corporate profits. The second factor

[19] An example of constraining intrusiveness could be not allowing a BigTech to pair (all of) its social media or ecommerce data with the collected payment data. For example, China has imposed regulatory bounds on the use of payment and platform data in BigTech credit provision.

is the privacy costs borne by DC users. In general form, welfare can be expressed as:

$$W = \int_{\text{borrowers}} (qy - c)\, f\left(q \middle| \text{"borrowing"}\right) dq - \int_{\text{DC users}} (\alpha\varphi\theta)\, g\left(\varphi \middle| \text{"DC use"}\right) d\varphi \tag{14}$$

At the policy maker's decision stage, the state of the economy is uncertain, and therefore the policy maker's objective is to maximize expected welfare, $E[W]$, with respect to θ:

Proposition 4 *The policy maker's objective is given by*

$$\max_{\theta} E[W] = \max_{\theta} \{\gamma W_h + (1-\gamma) W_l\} \tag{15}$$

$$W_h = \frac{3}{4} y_h - c - 2\alpha\theta \left[1 - \left(1 + \frac{4\alpha}{3(c+\tau)}\right) E[q|u]\right] \tag{16}$$

$$W_l = \begin{cases} \theta \frac{(y_l - (c+\tau))^2 (y_l - c + 2\tau)}{6\alpha y_l} & \text{if } y_l < 2\alpha + c + \tau \\ \theta \frac{\left((y_l - c)(y_l - c - 2\alpha) + \frac{4}{3}\alpha^2 - \tau^2\right)}{y_l} & \text{if } y_l > 2\alpha + c + \tau \end{cases} \tag{17}$$

where W_h and W_l represent welfare in the good and bad states, respectively, and the closed-form for $E[q|u]$ is in (10).

Proof. See Appendix A.4. ■

6.2 Socially optimal regulation

In the bad state, welfare increases linearly as θ increases, as seen from (17). A higher θ helps households who want to reveal their types to do so, while means of payment choice is purely individual and there are no disclosure externalities when the loan market for unrevealed households is inoperative. Instead, in the good state, all households receive loans and therefore the total value added of projects is independent of θ. The DC's impact on aggregate welfare then runs through privacy costs. A higher θ has a direct effect on privacy costs in (1) and in addition has an indirect effect from disclosure cascades. These effects are seen from, respectively, the terms θ and $E[q|u]$ in (16). Put together, direct and indirect

effects imply that W_h decreases more than linearly as θ increases. This nonlinearity creates the potential for an interior solution for socially optimal policy, derived in Proposition 5.

Proposition 5 *When the good state is certain ($\gamma = 1$), privacy costs dominate and the policy maker bans the DC, $\overline{\theta} = 0$. When the bad state is certain ($\gamma = 0$), credit inclusion from type differentiation dominates and laissez-faire, $\overline{\theta} = 1$, is socially optimal. When $\gamma \in (0,1)$, the policy maker for some parameterizations chooses to regulate the DC with $\overline{\theta} \in (0,1)$ given by*

$$\overline{\theta} = \frac{3}{32\alpha^2}\left[4\alpha^2 + (c+\tau)^2 - \sqrt{\frac{\left((c+\tau)^2-12\alpha^2\right)\left(8\alpha^2\gamma(\alpha+c)+3(1-\gamma)\lambda(c+\tau)^2+4\alpha^2(2\gamma\tau-(1-\gamma)\lambda)\right)^2}{16\alpha^4\gamma^2+48\alpha^3(1-\gamma)\gamma\lambda+9(1-\gamma)^2\lambda^2(c+\tau)^2+12\alpha^2(1-\gamma)\lambda(4\gamma(c+\tau)-(1-\gamma)\lambda)}}\right] \tag{18}$$

where

$$\lambda \equiv \begin{cases} \frac{(y_l-(c+\tau))^2(y_l-c+2\tau)}{6\alpha y_l} & \text{if } y_l < 2\alpha + c + \tau \\ \frac{\left((y_l-c)(y_l-c-2\alpha)+\frac{4}{3}\alpha^2-\tau^2\right)}{y_l} & \text{if } y_l > 2\alpha + c + \tau \end{cases} \tag{19}$$

Proof. See Appendix A.5. ■

6.3 Distributional effects

The distributional effects from DC design differ between the two states of y. In the bad state, DC users' disclosure is a Pareto improvement, creating value to them, no externalities to others, and profits to the corporate sector. In the good state, some (high q, low φ) households gain from lower customized loan rates. But their gains are exactly offset by the losses of unrevealed households, who face a higher loan rate.[20] Moreover, total corporate profits do not depend on θ in the good state: a higher θ transfers profits within the corporate sector, as the DC issuer gains data access fee revenues at lenders' expense. The impact of θ on W_h therefore derives from the one effect that has no offset: the privacy loss of some DC users that were swept along in disclosure cascades.

[20] In practice, such distributional effects could materialize along socially undesirable dimensions, such as race or gender (Chen et al., 2023; Fuster et al., 2022).

6.4 Comparative statics

The solution for $\overline{\theta}$ in (18) permits a numerical analysis of the comparative statics of optimal regulation to underlying parameters (as (18) does not readily lend itself to analytically determining the signs of derivatives). Of particular interest are the comparative statics with respect to τ and α, for which numerical examples are shown in Appendix B.4.

These numerical examples consistently indicate that $\frac{\partial \overline{\theta}}{\partial \tau} < 0$. A higher τ translates into increased market power for home lenders. The larger this market power is, the more regulation optimally leans against household disclosure. The intuition for this result is most easily understood in the context of the bad state. In the bad state, a higher τ increases the loan rate of DC users, but does not affect cash users, who do not obtain credit. A higher τ then pushes more households into cash use. With fewer DC users in the bad state, the social benefits of a larger θ in the bad state decline, because these stem from DC users who can become revealed and thereby obtain credit. Thus, $\overline{\theta}$ decreases.

In (1), α represents the weight that society places on privacy preferences. The relation between α and optimal regulation is potentially intricate. On the one hand, the "size" of cascades weakens as α rises. Indeed, if households did not care about privacy at all ($\alpha = 0$, which we preclude in (6)), then everyone would use the DC, even the households with the lowest q: cascades would be all-encompassing. However, cascades would not matter to welfare in that case. Thus, although a higher α weakens the extent of cascades, such cascades become more painful to society. In the numerical examples in Appendix B.4, the latter effect tends to dominate the welfare tradeoff. However, we find that for some parameterizations the relation between $\overline{\theta}$ and α can be u-shaped, and the fact that households "take on more of the regulator's job" for higher α (i.e., weakening disclosure cascades as households care more about privacy) can come to dominate for sufficiently large α.

7 Model extensions

This section considers four extensions to the baseline model: optimal policy with two instruments; a DC that is issued by a lender consortium rather than a separate monopoly; alternative sources of credit quality information for lenders; and a DC issuer that can make better inferences on household credit quality when more households use DC. The mathematical derivations for two of these extensions are contained in the Online Appendix, which also contains several additional extensions.[21]

7.1 A second policy instrument

Household data disclosure consists of two margins in our model: who sorts into DC use, which we can refer to as the extensive margin of disclosure; and how much information is disclosed about DC-using households, which we term the intensive margin. By constraining the intensive margin, regulation also affects sorting into means of payment and thereby the extensive margin. But the fact that there are two margins and one instrument implies that there could be room for a second instrument to improve welfare outcomes. We explore such a second instrument in the form of a government tax or subsidy on DC use (e.g., on DC transactions or on opening a user wallet), denoted by T, where $T > 0$ represents a tax and $T < 0$ is a subsidy.[22] We rederive the model with the inclusion of T in Appendix B.1 and find that for some parameterizations the availability of a second instrument does not raise welfare (e.g., when banning DC is optimal), while for other parameterizations it does. The most surprising outcome is that subsidizing DC use can sometimes be socially optimal.

[21] The contents of the Online Appendix are as follows. Appendix A contains the proofs of the propositions in the paper. Appendices B.1 and B.2 contain the derivations behind the extensions in, respectively, Sections 7.1 and 7.2. Appendix B.3 covers the "Medium y" case of Table 1. Appendix B.4 shows numerical examples behind the comparative statics discussed in Section 6.4. Appendices B.5 and B.6 analyze two additional extensions: Appendix B.5 considers an extension where the two dimensions of household heterogeneity, q and φ, are correlated. Appendix B.6 extends the model to empower households with a choice to only partially disclose by mixing cash and digital payments for consumption.

[22] An implicit assumption is that reduced (increased) DC-based consumption from taxation (subsidization) does not to affect the DC issuer's ability to infer household credit quality data. We also note that $T < 0$ can be alternatively interpreted as an additional term that represents ease-of-transactions benefit of DC relative to cash.

As discussed in Appendix B.1, a subsidy can optimally lean against underdisclosure, which occurs in the bad state: as lender market power establishes loan rates above c, some positive NPV households choose cash and remain credit excluded.

7.2 Alternative sources of credit quality

In the baseline model, the DC is the unique source of credit quality information. While adding lender-specific household data lies outside the scope of the current analysis, we can incorporate alternative public sources of information. Appendix B.2 considers two possibilities. The first is pre-existing public information on borrowers, such as from credit registries or FICO scores. The second is public payment system data, such as from a data-sharing CBDC, where we assume that the DC is more informationally intrusive than the public alternative. Appendix B.2 finds that both possibilities lead back to our baseline results, because sorting conditions between means of payment are unchanged. The key factor for our results is that the DC is more informationally intrusive than other payment alternatives, not that it is the unique source of credit quality data.

7.3 Lender consortium

This extension lets a consortium of lenders own the DC and decide on θ. We assume that the collected DC data is now freely available to all lenders. While lenders cooperate in DC design, their mode of competition on the credit market is unchanged compared to the baseline. In the good state, a lender's total equilibrium profit is τ.[23] Therefore, lenders' profits do not depend on θ and the lender consortium is indifferent about θ in this case: privately optimal θ is now indeterminate. Unlike the DC issuer, the lender consortium does not actively exploit disclosure externalities; unlike a social planner, the consortium does not lean against such externalities either. Instead, in the bad state, only revealed households

[23] Namely, a mass 1 of on-island borrowers, which have expected profit $q\left(\frac{c+\tau}{q}\right) - c = \tau$ per revealed borrower and $E[q|u]\left(\frac{c+\tau}{E[q|u]}\right) - c = \tau$ per unrevealed borrower.

receive credit and the mass of such households rises as θ increases, as seen from (13), so that the consortium sets $\theta = 1$. In the bad state, $\theta = 1$ creates value for households and lenders, and is optimal for DC issuer, lender consortium and social planner alike.

7.4 User base and inference quality

The quality of inferences that a data monopolist makes could depend on the size of its user base, as screening algorithms can become more effective when trained on a larger sample (Ghosh et al., 2024). We here write the probability of type revelation as a function $p(\theta, \mu(\theta))$ where $\mu(\theta)$ is the mass of DC users and p rises both as θ increases and as $\mu(\theta)$ increases. This setup further strengthens the incentives of the DC issuer to set $\theta = 1$ because, due to disclosure cascades, $\mu'(\theta) > 0$ (Section 4.1). That is, $\frac{dp(\theta,\mu(\theta))}{d\theta} = \frac{\partial p(\theta,\mu(\theta))}{\partial \theta} + \frac{\partial p(\theta,\mu(\theta))}{\partial \mu(\theta)}\mu'(\theta) > 0$ as all the partial derivatives are positive. If we normalize $p(1, \mu(1)) = 1$ (the largest attainable mass of DC users, which occurs at $\theta = 1$, is large enough that the inference technology works perfectly), then equilibrium expressions are identical to our baseline. With $\theta = 1$, the DC issuer now optimally plays on both households' disclosure externality and the network effect that improves its inference technology.

8 Conclusion

New technologies are fundamentally changing the way in which payments and credit provision take place, and create an unprecedented scope for data monopolization. Such developments can bring both unique opportunities and risks to consumers and borrowers, and policy institutions are actively grappling with the extent to which they should step in and regulate. This paper builds a framework to analyze how private and social optima diverge when a monopolist offers privacy-valuing households a digital payment system that collects data on their habits but can help them signal their creditworthiness to lenders. We find that the monopolist uses its pole position to play off both households and lenders against

each other. To maximize the value of household data in its hands, the monopolist makes its data collection as intrusive as possible, which squeezes households into a partially unraveling Lemons' Market, whereby the willingness of some high quality households to disclose prods more tranches of households to differentiate themselves to the detriment of their privacy.

Socially optimal data regulation aims at shielding households from being played off against each other. The challenge in doing so, is that regulation is a long-term policy, while the benefits from counteracting intrusive data collection are state dependent. In a bad state, where projects become less profitable, credit provision to unrevealed borrowers can freeze. The monopoly's data gathering can then be a net positive, not just to a subset of households, but to aggregate welfare, because the data unfreeze part of the credit market by differentiating better quality households. Socially optimal regulation trades off the partial preservation of credit provision during downturns against the privacy losses of households that are induced into disclosure cascades in upturns.

This paper focuses on a single payment provider and one outside payment option. Future work can build on the framework provided here to consider competition between multiple payment providers, which could include a public provider, such as a CBDC. Notably, our framework can already be interpreted from the angle of a benchmark case wherein CBDC is introduced with data-sharing features as the sole alternative to cash. What we discussed as optimal regulation ($\overline{\theta}$) of a private digital payment system, can equivalently be considered the optimal degree of data collection with a welfare-maximizing CBDC. Furthermore, a setting with multiple digital payment providers could provide a microfoundation for survey findings that consumers trust BigTechs less with their data than banks or FinTechs (Armantier et al., 2021), because consumers internalize the exploitative potential of BigTech data control.

Future work can also extend this framework to encompass elements of household data ownership. Policies that allow households to port their payment data between lenders have been implemented in various countries, and the implications of such "open banking policies" for credit market outcomes have been studied empirically (Babina et al., 2024; Nam, 2022).

Moreover, at least two major jurisdictions (the EU and California) have recently implemented regulation for private digital data that endows households with a degree of data ownership. Such regulation includes the ability for households to opt out of sharing their data and covers a wide variety of online footprints, such as browsing and social media data. For instance, Doerr et al. (2023) study how California's Consumer Privacy Act has affected credit market outcomes, particularly by FinTechs that rely on the purchase of such data.

In our framework, the digital payment monopolist's ability to profitably sell data about households' creditworthiness stems in part from the adverse selection suffered by any lender that would choose not to purchase this information. In a setting that would incorporate costs to collecting and processing payment data, policies that allow households to credibly transmit their payment data to lenders would reduce the incentives to collect their data in the first place. From this standpoint, data ownership policies could potentially complement or replace intrusive regulation directed at the design of the payment technology itself. We leave a formal analysis of this issue to future research.

References

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. (2022). Too Much Data: Prices and Inefficiencies in Data Markets. *American Economic Journal: Microeconomics*, 14(4):218–56.

Agarwal, S. and Hauswald, R. (2010). Distance and Private Information in Lending. *The Review of Financial Studies*, 23(7):2757–2788.

Agur, I., Ari, A., and Dell'Ariccia, G. (2022). Designing Central Bank Digital Currencies. *Journal of Monetary Economics*, 125:62–79.

Ahnert, T., Hoffmann, P., and Monnet, C. (2022). Payments and Privacy in the Digital Economy. ECB Working Papers 2662, European Central Bank.

Armantier, O., Doerr, S., Frost, J., Fuster, A., and Shue, K. (2021). Whom Do Consumers Trust With Their Data? US Survey Evidence. Bulletin 42, BIS.

Armantier, O., Doerr, S., Frost, J., Fuster, A., and Shue, K. (2024). Nothing to hide? Gender and Age Differences in Willingness to Share Data. Working Papers 1187, BIS.

Babina, T., Bahaj, S., Buchak, G., De Marco, F., Foulis, A., Gornall, W., Mazzola, F., and Yu, T. (2024). Customer Data Access and Fintech Entry: Early Evidence from Open Banking. *Journal of Financial Economics forthcoming*.

Beck, T., Gambacorta, L., Huang, Y., Li, Z., and Qiu, H. (2022). Big Techs, QR Code Payments and Financial Inclusion. Working Paper 1011, Bank for International Settlements.

Bergemann, D. and Bonatti, A. (2024). Data, Competition, and Digital Platforms. *American Economic Review*, 114(8):2553–2595.

Bergemann, D., Bonatti, A., and Gan, T. (2022). The Economics of Social Data. *The RAND Journal of Economics*, 53(2):263–296.

Bian, B., Ma, X., and Tang, H. (2023). The Supply and Demand for Data Privacy: Evidence from Mobile Apps. SSRN Electronic Journal.

Borgonovo, E., Caselli, S., Cillo, A., Masciandaro, D., and Rabitti, G. (2021). Money, Privacy, Anonymity: What Do Experiments Tell Us? *Journal of Financial Stability*, 56:100934.

Bouvard, M., Casamatta, C., and Xiong, R. (2022). Lending and Monitoring: Big Tech vs Banks. TSE Working Papers 1386, Toulouse School of Economics.

Brits, H. and Jonker, N. (2023). The Use of Financial Apps: Privacy Paradox or Privacy Calculus? WP 794, De Nederlandsche Bank.

Brunnermeier, M. K. and Payne, J. (2022). Platforms, Tokens, and Interoperability. WP 2022-8, Princeton University.

Budak, J., Rajh, E., and Škrinjarić, B. (2023). Resilience to Online Privacy Violation: Developing a Typology of Consumers. *Sc. Annals of Econ. and Business*, 70(3):379–98.

Chen, S., Doerr, S., Frost, J., Gambacorta, L., and Shin, H. (2023). The Fintech Gender Gap. *Journal of Financial Intermediation*, 54:101026.

Choi, S., Kim, B., Kim, Y.-S., and Kwon, O. (2023). Central Bank Digital Currency and Privacy: A Randomized Survey Experiment. WP 1147, Bank for International Settlements.

Collis, A., Moehring, A., Sen, A., and Acquisti, A. (2024). Information Frictions and Heterogeneity in Valuations of Personal Data. SSRN Electronic Journal.

Cong, L. W. and Mayer, S. (2023). Antitrust, Regulation, and User Union in the Era of Digital Platforms and Big Data. WP 30881, National Bureau of Economic Research.

Cunliffe, J. (2023). Money and Payments: A Black Ships Moment? Speech delivered at the Economics of Payments XII Conference at the Federal Reserve Board on Oct. 26, 2023.

Doerr, S., Frost, J., Gambacorta, L., and Shreeti, V. (2024). Big Techs in Finance. In *Oxford Handbook of Banking (forthcoming)*. Oxford University Press.

Doerr, S., Gambacorta, L., Guiso, L., and del Villar, M. S. (2023). Privacy Regulation and Fintech Lending. BIS Working Papers 1088, Bank for International Settlements.

ECB (2021). Eurosystem Report on the Public Consultation on a Digital Euro. Technical report, European Central Bank.

Engels, C., Francis, B. B., and Philip, D. (2022). The Cost of Privacy Failures: Evidence from Bank Depositors' Reactions to Breaches. SSRN Electronic Journal.

Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., and Zbinden, P. (2020). BigTech and the Changing Structure of Financial Intermediation. *Economic Policy*, 34(100):761–799.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance*, 77(1):5–47.

Gambacorta, L., Khalil, F., and Parigi, B. M. (2022). Big Techs vs Banks. WP 1037, BIS.

Garratt, R. and Lee, M. J. (2021). Monetizing Privacy. Staff Reports 958, NY Fed.

Garratt, R. J. and Van Oordt, M. R. (2021). Privacy as a Public Good: A Case for Electronic Cash. *Journal of Political Economy*, 129(7):2157–2180.

Ghosh, P., Vallee, B., and Zeng, Y. (2024). FinTech Lending and Cashless Payments. *Journal of Finance forthcoming*.

Goldfarb, A. and Que, V. F. (2023). The Economics of Digital Privacy. *Annual Review of Economics*, 15(1):267–286.

Goldfarb, A. and Tucker, C. (2012). Shifts in Privacy Concerns. *American Economic Review*, 102(3):349–53.

Grossman, S. J. and Hart, O. D. (1980). Disclosure Laws and Takeover Bids. *The Journal of Finance*, 35(2):323–334.

Hau, H., Huang, Y., Shan, H., and Sheng, Z. (2019). How FinTech Enters China's Credit Market. *AEA Papers and Proceedings*, 109:60–64.

He, Z., Huang, J., and Zhou, J. (2023). Open Banking: Credit Market Competition When Borrowers Own the Data. *Journal of Financial Economics*, 147(2):449–474.

Heitz, A. R., Martin, C., and Ufier, A. (2023). Bank Loan Monitoring, Distance, and Delegation. *AEA Papers and Proceedings*, 113:177–81.

Huang, J. (2021). Fintech Expansion. SSRN Electronic Journal.

Huang, Y., Zhang, L., Li, Z., Qiu, H., Sun, T., and Wang, X. (2020). Fintech Credit Risk Assessment for SMEs: Evidence from China. WP 20/193, International Monetary Fund.

Jiang, E. X., Yu, G. Y., and Zhang, J. (2023). Bank Competition amid Digital Disruption: Implications for Financial Inclusion. SSRN Electronic Journal.

Jovanovic, B. (1982). Truthful Disclosure of Information. *The Bell Journal of Economics*, 13(1):36–44.

Li, J. (2023). Predicting the Demand for Central Bank Digital Currency: A Structural Analysis with Survey Data. *Journal of Monetary Economics*, 134:73–85.

Li, J. and Pegoraro, S. (2023). Borrowing from a Bigtech Platform. SSRN Electronic Journal.

Lin, T. (2022). Valuing Intrinsic and Instrumental Preferences for Privacy. *Marketing Science*, 41(4):60–64.

Lin, T. and Strulov-Shlain, A. (2023). Choice Architecture, Privacy Valuations, and Selection Bias in Consumer Data. In *Proceedings of the 24th ACM Conference*. ACM.

Liu, E., Ma, S., and Veldkamp, L. (2023). Data Sales and Data Dilution. mimeo.

Nam, R. J. (2022). Open Banking and Customer Data Sharing: Implications for Fintech Borrowers. SAFE Working Papers 364, Leibniz Institute for Financial Research.

Ouyang, S. (2022). Cashless Payment and Financial Inclusion. SSRN Electronic Journal.

Parlour, C. A., Rajan, U., and Zhu, H. (2022). When FinTech Competes for Payment Flows. *Review of Financial Studies*, 35(11):4985–5024.

Prince, J. T. and Wallsten, S. (2022). How Much is Privacy Worth Around the World and Across Platforms? *Journal of Economics and Management Strategy*, 31(4):841–861.

Tang, H. (2023). The Value of Privacy: Evidence from Online Borrowers. mimeo.

Online Appendix

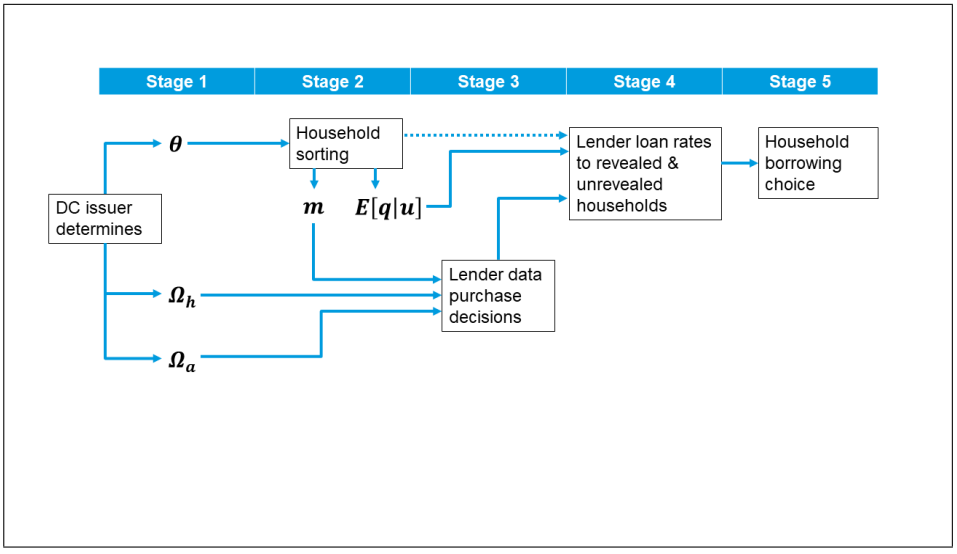
Equations in Appendix A are numbered with "A." and in Appendix B with "B." References to equations without letters refer to equations in the main text of the paper.

A Proofs

This appendix contains the proofs of the propositions in the main text. All calculations are performed in a *Mathematica* file that is included as a supplemental file.

A.1 Proof of Proposition 1

This proof applies backward induction to solve for the Subgame Perfect Nash Equilibrium in pure strategies. We save notation by focusing on a single subset of households (as results extend symmetrically to all subsets) with a given home lender, h, and away (non-home) lenders denoted by a. Thus, $R_{kh}(q)$ and $R_{ka}(q)$ are the differentiated loan rates that, respectively, the home lender and other lenders charge revealed households; and R_{uh} and R_{ua} are such loan rates charged to unrevealed households. We will similarly denote by Ω_h and Ω_a the data fees that the DC issuer charges, respectively, the home lender and the away lenders. We further recall from (3) that m represents the mass of revealed households.



Flow diagram of decisions and variables.

The figure above provides an overview of the decision variables in the model at each stage and how these are affected by the decisions taken at preceding stages. That is, the proof goes backward through the stages, but in doing so, it considers at each stage which decision variables from past decision stages are given. In particular, the DC issuer's choice of θ affects household sorting into payment instruments at Stage 2, which determines which household types become revealed and thereby the mass of revealed households, m, as well

as the creditworthiness of the pool of unrevealed households, $E[q|u]$. In turn, m, together with the data fees, Ω_h and Ω_a, that the DC issuer set at Stage 1, feed into the lenders' data purchase decisions at Stage 3. These data purchase decisions and the household sorting captured by $E[q|u]$ from Stage 2, then determine loan rates at Stage 4. Based on these loan rates, households subsequently (Stage 5) decide from which lender to borrow.

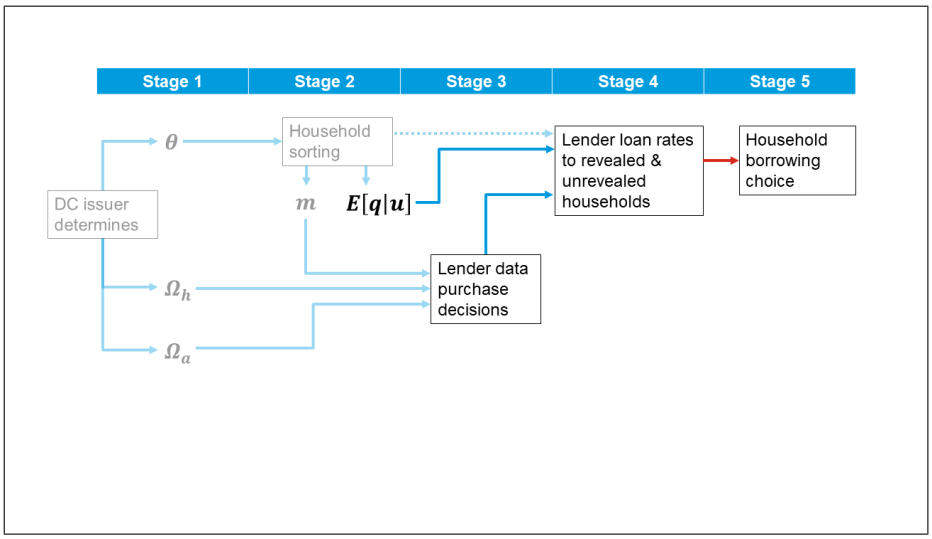
These relations are represented by the solid arrows in the figure above. The figure also includes a dotted arrow between household sorting at Stage 2 and lender loan rates at Stage 4. The difference between this dotted arrow and the solid arrows emerging from household sorting represents the difference between the perspective of an individual household and the aggregate set of households. An individual household is atomistic and knows that its individual choice has no impact on the aggregate household variables that affect subsequent stages. Since aggregate household variables ultimately have to be consistent with the aggregation over individual household optimization problems, the Stage 2 solution technique is based on finding a fixed point where these aggregate variables are indeed consistent with the outcomes of household optimization based on the aggregate variables. This is clarified at Stage 2 of our backward induction analysis.

We also note that Stage 6 is the project realization stage and does not involve player decisions. Therefore, the backward induction in this proof starts from Stage 5.

Stage 5. At this stage, households will straightforwardly choose to borrow from the lender that offers them the lowest loan rate. Formally, at Stage 5, some households' credit qualities are unrevealed (namely, those that chose cash at Stage 2 and those that chose DC at Stage 2 but remained unrevealed (with probability $1-\theta$) at the revelation draw at the end of Stage 2), while others are revealed (namely, those that chose DC at Stage 2 and, with probability θ, became revealed). If a household borrows from the home lender, then the profit on a project, if successful at Stage 6, is $y - R_{uh}$ and $y - R_{kh}(q)$ if the household is, respectively, unrevealed and revealed, and 0 if the project is unsuccessful. If borrowing from another lender, these expressions are $y - R_{ua}$ and $y - R_{ka}(q)$. The Stage 5 optimal choice is therefore for the unrevealed to borrow from their home lender if $R_{uh} \leq R_{ua}$ and from another lender otherwise, and for the revealed this condition is $R_{kh}(q) \leq R_{ka}(q)$. Here, the "smaller or equal than" signs follow from the assumed breakeven preference for the home lender (see footnote 13 in Section 3.3).

Stage 4. This is the stage at which lenders determine their loan rates. At Stage 4, lenders foresee that at Stage 5 households will borrow from the lender that offers them the lowest loan rate, while lenders take as given: θ and the data fees, Ω_h and Ω_a, set by the DC issuer at Stage 1, household sorting at Stage 2, and lenders' Stage 3 decisions on obtaining data access. As shown in the figure below (and detailed in Table S4 below), the decision variables that directly affect the Stage 4 lender decisions are $E[q|u]$ from household sorting at Stage 2 and lender data purchase decisions from Stage 3 (which, in turn, are affected by Stage 1 decisions of the DC issuer on θ and the data fees). In the figure below, the backward induced (Stage 5) outcome is depicted by a red arrow and the key input variables from preceding stages are in blue arrows (while the variables that in turn affect those variables at

preceding stages are shown in faded colors).



Stage 4 decision: input variables and backward induced outcomes.

For Stage 4, we are going to derive the equilibria within each of the four possible subgames. Table S4 summarizes these subgame equilibria, which are derived per case below it. Within these subgame derivations, we show the relevant profit function that is maximized. At the end of the Stage 4 discussion (equation (A.16)), we also show an overall profit function of the home lender that incorporates all subgames. Away lenders' play instead directly follows from Bertrand competition: at Stage 4, away lenders can be modeled as a single player that always sets loan rates equal to the marginal cost of providing the loan.[1]

We further note that the subgames shown in Table S4 incorporate the household breakeven preference in favor of the home lender. For example, in the top-left cell of S4, the home lender's equilibrium loan rates are the limit cases of, respectively, $R_k(q) = \frac{c+\tau}{q} - \varepsilon$ and $R_u = \frac{c+\tau}{E[q|u]} - \varepsilon$ with $\varepsilon \to 0$.

Table S4: Stage 4 loan rate equilibria within subgames

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: $R_{uh} = \frac{c+\tau}{E[q\|u]}$, $R_{kh}(q) = \frac{c+\tau}{q}$ Away lenders: $R_{ua} = \frac{c+\tau}{E[q\|u]}$, $R_{ka}(q) = \frac{c+\tau}{q}$	*"Top-right" subgame:* Home lender: $R_h = \frac{4}{3}(c+\tau)$ Away lenders: $R_a = \frac{4}{3}(c+\tau)$
Home lender has no data access	*"Bottom-left" subgame:* Home lender: $R_{uh} = \frac{c+\tau}{E[q\|u]}$ Away lenders: $R_{ua} = \frac{c+\tau}{E[q\|u]}$, $R_{ka}(q) = \frac{c+\tau}{q}$	*"Bottom-right" subgame:* Home lender: $R_h = \frac{4}{3}(c+\tau)$ Away lenders: $R_a = \frac{4}{3}(c+\tau)$

[1] With $N > 3$ and symmetric pure strategies there are always at least two equally informed away lenders engaging in Bertrand competition.

Subgame 1: all lenders have data access (top-left subgame of Table S4). Due to Bertrand competition, it follows directly that loan rates among away lenders are given by the breakeven rates. For away lenders, the breakeven loan rate on revealed households is $R_{ka}(q) = \frac{c+\tau}{q}$ and on unrevealed households it is $R_{ua} = \frac{c+\tau}{E[q|u]}$, because the marginal cost of a loan by an away lender is $c+\tau$, while the probability of repayment is q for a revealed household and $E[q|u]$ for unrevealed households. Strictly, these are *expected* breakeven rates, given the uncertain realizations of project success. However, with a continuum of borrowers, this distinction is immaterial.

We can now write the home lender's profit function as

$$\Pi_h^{S4,TL} = \int_{\overline{q}}^{1} Z_k \left(qR_{kh}^{TL}(q) - c\right) f(q)dq + Z_u \left(E[q|u] R_{uh}^{TL} - c\right)(1-m) \tag{A.1}$$

where $\Pi_h^{S4,TL}$ stands for the profit of the home lender at Stage 4 ($S4$), in the top-left (TL) subgame. Here, Z_k is an indicator variable that takes value 1 when $R_{kh}(q) \leq \frac{c+\tau}{q}$ and 0 otherwise; and Z_u is an indicator variable that takes value 1 when $R_{uh} \leq \frac{c+\tau}{E[q|u]}$ and 0 otherwise. Moreover, we let $\overline{q}$ stand for the q of the lowest quality revealed household, while $f(q)$ is the distribution of revealed households on $q \in [\overline{q}, 1]$. We do not need to (and cannot) solve for $\overline{q}$ and $f(q)$ at this stage of the game (see backward induction at Stage 2 for solutions). Instead, we can identify the home lender's optimal rates directly from (A.1) as the highest $R_{kh}(q)$ and R_{uh} for which, respectively, $Z_k = 1$ and $Z_u = 1$: when the home lender charges rates above the Bertrand Nash game rates of the away lenders, it faces 0 demand for loans at Stage 5. It then follows directly that the subgame optimal loan rates for the home lender are $R_{kh}(q) = \frac{c+\tau}{q}$ for revealed households and $R_{uh} = \frac{c+\tau}{E[q|u]}$ for the unrevealed. These are the highest rates at which households at Stage 5 are willing to borrow from their home lender.

We further note that $y > 2(c+\tau)$ (i.e., Proposition 1 covers the "High y" case in Table 1) and $q \in \left[\frac{1}{2}, 1\right]$, here imply that $y - \frac{c+\tau}{q} > 0$ and $y - \frac{c+\tau}{E[q|u]} > 0$ and therefore at the subgame equilibrium loan rates, lenders are ensured positive returns on households that borrow from them (and, to avoid repetition, we note that the same holds true in the other Stage 4 subgames).

Subgame 2: the home lender has data access, while the away lenders do not (top-right subgame of Table S4). The uninformed away lenders Bertrand compete on a single loan rate to all households, where the breakeven rate is $\frac{c+\tau}{E[q]}$ and since $E[q] = \frac{3}{4}$ (as households are uniformly distributed on $q \in \left[\frac{1}{2}, 1\right]$), this yields $R_a = \frac{4}{3}(c+\tau)$ as the subgame equilibrium loan rate of the away lenders. In the same manner as for Subgame 1, we can now write the home lender profit function as

$$\Pi_h^{S4,TR} = \int_{\overline{q}}^{1} Z_{all} \left(qR_{kh}^{TR}(q) - c\right) f(q)dq + Z_{all} \left(E[q|u] R_{uh}^{TR} - c\right)(1-m) \tag{A.2}$$

where Z_{all} is an indicator variable that takes value 1 when the associated home lender interest rate (either $R_{kh}^{TR}(q)$ or R_{uh}^{TR}) is set at or below $\frac{4}{3}(c+\tau)$ and 0 otherwise. It follows that the

subgame optimal loan rate for the home lender is $R_h^{TR} = R_{kh}^{TR}(q) = R_{uh}^{TR} = \frac{4}{3}(c+\tau)$. All households, including the highest quality types, are "captive" at that rate (cannot do better at other lenders, which cannot type differentiate because they are unrevealed to other lenders). There is therefore no incentive for the home lender to offer higher quality borrowers customized (lower) loan rates.

Put differently, $\Pi_h^{S4,TR}$ can be more succinctly written as

$$\Pi_h^{S4,TR} = Z_{all}\left(E[q] R_h^{TR} - c\right) \tag{A.3}$$

where $Z_{all} = 1 \Leftrightarrow R_h \leq \frac{4}{3}(c+\tau)$ and therefore $R_h = \frac{4}{3}(c+\tau)$ is optimal.

Subgame 3: the home lender has no data access, while the away lenders do (bottom-left subgame of Table S4). Informed away lenders' Bertrand compete against each other in the same way that they do in Subgame 1. Therefore, $R_{ka}(q) = \frac{c+\tau}{q}$ and $R_{ua} = \frac{c+\tau}{E[q|u]}$. As it cannot differentiate between borrowers, the home lender can offer only one loan rate, R_h, to all households. Our derivations next center on finding the R_h that the home lender optimally offers in this subgame. We can first note that $R_h > \frac{c+\tau}{E[q|u]}$ cannot be optimal: the home lender would fail to attract any borrowers, because all households can do better at the rates offered by the away lenders. Therefore, we center attention on $R_h \leq \frac{c+\tau}{E[q|u]}$.

Our proof focuses on showing that for any $R_h \leq \frac{c+\tau}{E[q|u]}$ it holds that $\frac{\partial \Pi_h}{\partial R_h} > 0$, where Π_h is the home lender's profit. When this is the case, it follows that $R_h = \frac{c+\tau}{E[q|u]}$ is optimal for the home lender. The approach for this proof is based on sufficient conditions. In particular, we consider a setting that makes it maximally attractive for the home lender to set $R_h < \frac{c+\tau}{E[q|u]}$. If even for this setting, $R_h = \frac{c+\tau}{E[q|u]}$ remains optimal (i.e., $\frac{\partial \Pi_h}{\partial R_h} > 0$ holds unambiguously), then it will be true in general.

The first step is to note that setting $R_h < \frac{c+\tau}{E[q|u]}$ is more attractive when more households are revealed. The cost of cutting R_h is losing out on profits on unrevealed households. The home lender earns $(1-m)(E[q|u] R_h - c)$ on unrevealed households (who all borrow from it rather than away lenders, because $R_h \leq \frac{c+\tau}{E[q|u]} = R_{ua}$). Thus, a lower R_h leads to lower profits on the unrevealed. The larger is m (the mass of revealed households), the smaller is $(1-m)$ and therefore the smaller the profits associated with lending to the unrevealed, and the less the home lender has to lose from cutting R_h. The upside of cutting R_h is that the home lender manages to attract revealed households. Revealed households with q such that $R_h < R_{ka}(q) = \frac{c+\tau}{q}$ will choose to borrow from the home lender. Thus, the larger is m the more revealed borrowers there are to be gained by setting $R_h < \frac{c+\tau}{E[q|u]}$ and the smaller are the total losses on unrevealed households as compared to setting $R_h = \frac{c+\tau}{E[q|u]}$. What we are looking for is therefore $\sup m$: the largest mass of revealed households that the model can generate, which maximizes the incentives to set $R_h < \frac{c+\tau}{E[q|u]}$ and thereby provides us with a sufficient condition for the optimality of $R_h = \frac{c+\tau}{E[q|u]}$.

We next find $\sup m$ from considering what is the minimum set of households that would always choose to use cash and thereby certainly remain unrevealed (i.e., this will give $\inf(1-m)$ and thus also $\sup m$). Cash is least attractive if remaining unrevealed leads to the worst possible inference about a household's credit quality. For the purpose of this

sufficient condition, assume that all households that choose cash expect to be seen by lenders as though they are the worst quality household with $q = \frac{1}{2}$. I.e., households in this sufficient condition act from the most cash-unattractive prior. Note that, since this entire proof is conducted by backward induction, we cannot yet solve for household Stage 2 behavior foreseeing the actual interest rate determination process at Stage 4. Rather, at Stage 4, we solve for equilibria within loan rate setting subgames, and within this particular subgame we work with a sufficient condition that shows that $R_h = \frac{c+\tau}{E[q|u]}$ must be optimal for the home lender, regardless of what household sorting actually looks like at Stage 2: even at the most adverse sorting from the perspective of setting $R_h = \frac{c+\tau}{E[q|u]}$, that loan rate nevertheless remains optimal for the home lender.

A different way to put it is that households in this sufficient case act with an assumed prior $E[q|\,u] = \frac{1}{2}$ at Stage 2. Calculating the equilibrium value of $E[q|\,u]$ involves solving for Stage 4 subgames and rolling back the subgames to Stage 2. But we must first solve for the Stage 4 subgames, and the manner in which we solve this particular subgame is through a "worst case prior" from the perspective of $R_h = \frac{c+\tau}{E[q|u]}$ optimality. Even though that prior on household sorting will ultimately be incorrect in equilibrium, it does correctly identify an absolute lower bound on the share of unrevealed households, and therefore the most unattractive case for the home lender to maintain $R_h = \frac{c+\tau}{E[q|u]}$, which it nevertheless is shown to optimally maintain in this case.

Any household with a privacy preference strong enough that it would choose to use cash even when this leads to an inference that it is $q = \frac{1}{2}$, would certainly choose cash under any circumstance. The advantage to choosing DC over cash is further maximized when households know that they will obtain their type differentiated rates, $\frac{c+\tau}{q}$, with certainty when they choose DC (i.e., it is sufficient to let $\theta = 1$), because these rates are always lower than the rate $\frac{c+\tau}{1/2}$ that cash users expect.

In this setting, a household prefers cash over DC (see (1)) if and only if

$$q\left(y - \frac{c+\tau}{q}\right) - \alpha\varphi < q\left(y - \frac{c+\tau}{1/2}\right) \tag{A.4}$$

which becomes

$$\varphi > (2q-1)\frac{c+\tau}{\alpha} \tag{A.5}$$

and this condition will be tightest (i.e., having the smallest mass of households opting for cash and remaining unrevealed) when α is at its infimum which, from (6), is $\alpha = \frac{1}{9}(c+\tau)$. Replacing this gives

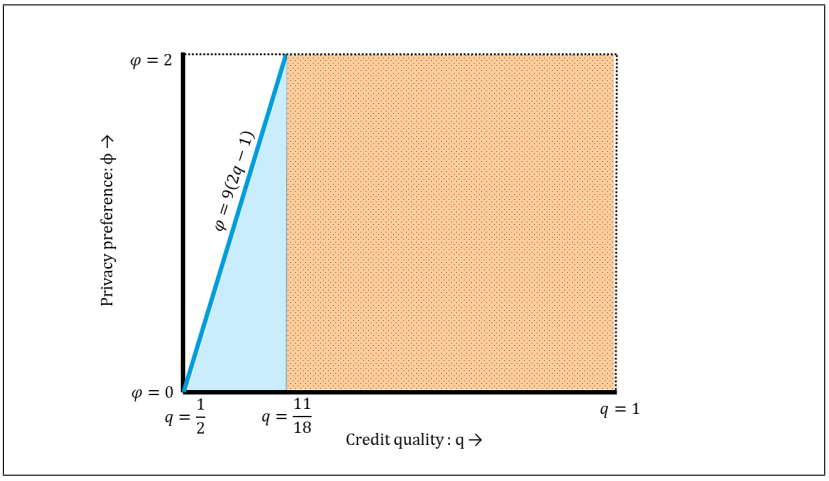
$$\varphi > 9(2q-1) \tag{A.6}$$

To get from this expression to $\sup m$ requires finding the integral over all revealed households implied by the fact that only $\varphi > 9(2q-1)$ households choose cash. We here get

$$\sup m = \int_{\frac{1}{2}}^{\frac{11}{18}} 9(2q-1)\,dq + \int_{\frac{11}{18}}^{1} 2dq = \frac{8}{9} \tag{A.7}$$

because: 1) any $\varphi = 0$ household always chooses DC (i.e., the condition in (A.6) is violated when $\varphi = 0$). Therefore, the integral over DC users starts from $q = \frac{1}{2}$, which the lowest

q; 2) between $q = \frac{1}{2}$ and $q = \frac{11}{18}$ the probability density function of DC users is given by $9(2q-1)$ from (A.6). This is the blue shaded area underneath the indifference frontier in the figure below and is captured by the term $\int_{\frac{1}{2}}^{\frac{11}{18}} 9(2q-1)\,dq$; 3) At $q = \frac{11}{18}$, we have that $9(2q-1) = 2$ and therefore for any $q \geq \frac{11}{18}$ the condition in (A.6) is always violated, regardless of a household's φ. Hence, for $q \geq \frac{11}{18}$ it follows that $f(q) = 2$ because the mass of revealed households is maximized: $f(q) = 2$ occurs where the indifference frontier as displayed in the figure below has "hit" the upper bound of $\varphi = 2$ and all households choose the same payment instrument (DC). This is shown as the beige dotted area and is formally captured by the term $\int_{\frac{11}{18}}^{1} 2dq$. Taken together, the shaded areas in the figure below encompass the mass of revealed households in the calculation of $\sup m$.



Calculation of $\sup m$: a visual representation.

The next step is to take this setting to $\frac{\partial \Pi_h}{\partial R_h}$, for which we need an expression for Π_h. Here, Π_h can be expressed as:

$$\Pi_h = \begin{cases} \int_{\frac{1}{2}}^{\frac{c+\tau}{R_h}} (qR_h - c)\,9(2q-1)\,dq + \frac{1}{9}\left(E[q|u]\,R_h - c\right) & \text{if } R_h > \frac{c+\tau}{11/18} \\ \int_{\frac{1}{2}}^{\frac{11}{18}} (qR_h - c)\,9(2q-1)\,dq + \int_{\frac{11}{18}}^{\frac{c+\tau}{R_h}} (qR_h - c)\,2dq + \frac{1}{9}\left(E[q|u]\,R_h - c\right) & \text{if } R_h < \frac{c+\tau}{11/18} \end{cases} \tag{A.8}$$

where the level of R_h determines from which part of the figure above the home lender is attracting revealed households.

When $R_h > \frac{c+\tau}{11/18}$, some of the revealed household in the blue shaded area choose to borrow from the home lender. In particular, households with $q < \frac{c+\tau}{R_h}$ find that R_h is lower than the alternative offer of $R_{ka}(q) = \frac{c+\tau}{q}$ from away lenders. The home lender expects to earns $(qR_h - c)$ per revealed household and since all revealed households with $q \in \left[\frac{1}{2}, \frac{c+\tau}{R_h}\right]$ choose to borrow from it and $R_h > \frac{c+\tau}{11/18}$, which implies that all these revealed households are underneath the indifference frontier in the blue shaded area of the figure above, then the home lender's profit on revealed households is $\int_{\frac{1}{2}}^{\frac{c+\tau}{R_h}} (qR_h - c)\,9(2q-1)\,dq$. Instead, when

$R_h < \frac{c+\tau}{11/18}$, all revealed households in the blue shaded area (i.e., all revealed households with $q \in \left[\frac{1}{2}, \frac{11}{18}\right]$) borrow from the home lender and instead the cutoff households that prefer to borrow from away lenders with $R_{ka}(q) = \frac{c+\tau}{q}$ (namely the households marginally above $q = \frac{c+\tau}{R_h}$) are within the beige, dotted area of the figure. Therefore, the home lender's profit on revealed households is $\int_{\frac{1}{2}}^{\frac{11}{18}} (qR_h - c)\, 9\,(2q-1)\, dq + \int_{\frac{11}{18}}^{\frac{c+\tau}{R_h}} (qR_h - c)\, 2dq$. The home lender's profit on unrevealed households, all of whom borrow from the home lender, is always $(1-m)\,(E[q|u]\,R_h - c)$ and from (A.7) this becomes $\frac{1}{9}\,(E[q|u]\,R_h - c)$.

The next step involves finding the $E[q|u]$ that is commensurate with this sufficient condition derivation. That is, given sorting implied by households' worst-case prior, what is the average quality of the unrevealed household? It is this average quality that feeds into Π_h in (A.8). With this solved for, we will be able to take $\frac{\partial \Pi_h}{\partial R_h}$ and show that $\frac{\partial \Pi_h}{\partial R_h} > 0$.

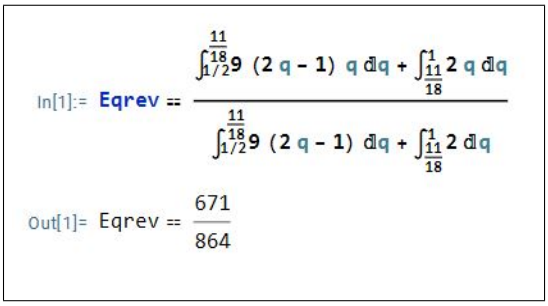
We first note that $(1-m)\,E[q|u] + mE[q|rev] = E[q] = \frac{3}{4}$, where $E[q|rev]$ is the expected quality of revealed households. Replacing from (A.7), this means that here

$$\frac{1}{9}E[q|u] + \frac{8}{9}E[q|rev] = \frac{3}{4} \tag{A.9}$$

We can find $E[q|rev]$ as a weighted average of expected values with respect to the integral areas of revealed households:

$$E[q|rev] = \frac{\int_{\frac{1}{2}}^{\frac{11}{18}} 9\,(2q-1)\,qdq + \int_{\frac{11}{18}}^{1} 2qdq}{\int_{\frac{1}{2}}^{\frac{11}{18}} 9\,(2q-1)\,dq + \int_{\frac{11}{18}}^{1} 2dq}$$

the solution to which is $E[q|rev] = \frac{671}{864}$:



Mathematica calculation of $E[q|rev]$.

and replacing $E[q|rev] = \frac{671}{864}$ into (A.9) gives

$$E[q|u] = \frac{29}{54} = \inf E[q|u] \tag{A.10}$$

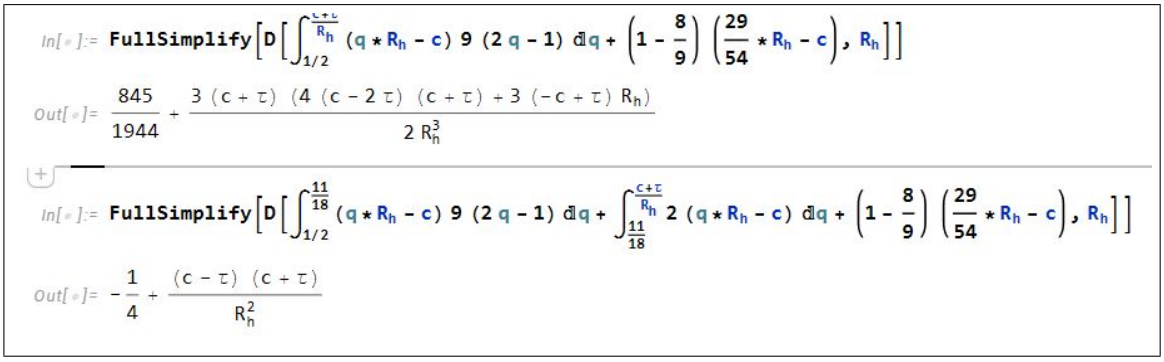
where we note that this is the infimum of $E[q|u]$ in the same way that $\frac{8}{9}$ is the supremum of m. That is, $E[q|u]$ could never be lower than $\frac{29}{54}$ under any household sorting, because we are in the setting that makes cash the least attractive it could possibly be, and the gap

to DC use as large as it could possibly be (because all DC users become revealed and get customized loan rates), so that the set of cash users is at the lowest range of q (and therefore the lowest $E[q|u]$) that it could ever be.

Replacing $E[q|u] = \frac{29}{54}$ back into Π_h, we obtain

$$\Pi_h = \begin{cases} \int_{\frac{1}{2}}^{\frac{c+\tau}{R_h}} (qR_h - c)\, 9\,(2q-1)\, dq + \frac{1}{9}\left(\frac{29}{54}R_h - c\right) \text{ if } R_h > \frac{c+\tau}{11/18} \\ \int_{\frac{1}{2}}^{\frac{11}{18}} (qR_h - c)\, 9\,(2q-1)\, dq + \int_{\frac{11}{18}}^{\frac{c+\tau}{R_h}} (qR_h - c)\, 2dq + \frac{1}{9}\left(\frac{29}{54}R_h - c\right) \text{ if } R_h < \frac{c+\tau}{11/18} \end{cases} \tag{A.11}$$

so that we can now take the derivative $\frac{\partial \Pi_h}{\partial R_h}$, as shown in below:

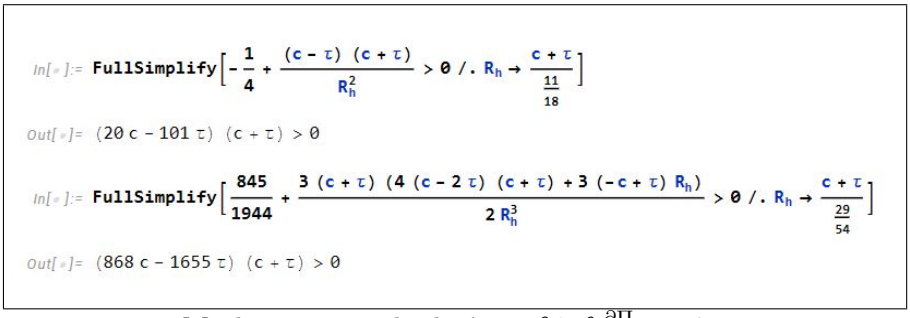


Derivatives calculated in *Mathematica*.

which we can write as:

$$\frac{\partial \Pi_h}{\partial R_h} = \begin{cases} \frac{3(c+\tau)(4(c-2\tau)(c+\tau)-3(c-\tau)R_h)}{2R_h^3} + \frac{845}{1944} \text{ if } R_h > \frac{c+\tau}{11/18} \\ \frac{(c-\tau)(c+\tau)}{R_h^2} - \frac{1}{4} \text{ if } R_h < \frac{c+\tau}{11/18} \end{cases} \tag{A.12}$$

Our aim is to find $\inf \frac{\partial \Pi_h}{\partial R_h}$ and show that $\inf \frac{\partial \Pi_h}{\partial R_h} > 0$ so that $\frac{\partial \Pi_h}{\partial R_h} > 0$ always holds. Since both terms in $\frac{\partial \Pi_h}{\partial R_h}$ decrease for a larger R_h, we can find $\inf \frac{\partial \Pi_h}{\partial R_h} > 0$ for $\sup R_h$. Within the case "$\frac{(c-\tau)(c+\tau)}{R_h^2} - \frac{1}{4}$ if $R_h < \frac{c+\tau}{11/18}$", we have that $\sup R_h = \frac{c+\tau}{11/18}$. And within the case "$\frac{3(c+\tau)(4(c-2\tau)(c+\tau)-3(c-\tau)R_h)}{2R_h^3} + \frac{845}{1944}$ if $R_h > \frac{c+\tau}{11/18}$" we have that $\sup R_h = \frac{c+\tau}{E[q|u]} = \frac{c+\tau}{29/54}$ (since $R_{ua} = \frac{c+\tau}{E[q|u]} = \frac{c+\tau}{E[q|u]} = \frac{c+\tau}{29/54}$). Applying these suprema to (A.12) yields $\inf \frac{\partial \Pi_h}{\partial R_h} > 0$ as



Mathematica calculation of $\inf \frac{\partial \Pi_h}{\partial R_h} > 0$.

and therefore

$$\inf \frac{\partial \Pi_h}{\partial R_h} > 0 \Leftrightarrow 20c > 101\tau \wedge 868c > 1655\tau \tag{A.13}$$

where $20c > 101\tau$ is a tighter constraint than $868c > 1655\tau$ and we can write $20c > 101\tau$ as

$$\inf \frac{\partial \Pi_h}{\partial R_h} > 0 \Leftrightarrow \tau < \frac{20}{101} c \tag{A.14}$$

which is true since $\tau < \frac{1}{6}c$ from (5) and $\frac{20}{101} > \frac{1}{6}$. And $\inf \frac{\partial \Pi_h}{\partial R_h} > 0$ for any $R_h \leq \frac{c+\tau}{E[q|u]}$ implies that $R_h = \frac{c+\tau}{E[q|u]}$ is optimal for the home lender.

Subgame 4: all lenders are uninformed (bottom-right subgame of Table S4). The uninformed away lenders Bertrand compete on a single loan rate to all households, where the breakeven rate is $\frac{c+\tau}{E[q]}$ and since $E[q] = \frac{3}{4}$ (as households are uniformly distributed on $q \in \left[\frac{1}{2}, 1\right]$), this yields $R_a = \frac{4}{3}(c+\tau)$ as the subgame equilibrium loan rate of the away lenders. The profit function of the uninformed home lender is then given by:

$$\Pi_h^{S4,BR} = Z_{all}\left(E[q] R_h^{BR} - c\right) \tag{A.15}$$

and therefore, as in Subgame 2, $R_h = \frac{4}{3}(c+\tau)$ is optimal for the home lender.

The home lender's overall profit function for Stage 4. Subgames 1-4 identified the home lender profit function within each subgame. We can write the home lender's Stage 4 profit function over all subgames as the following function:

$$\begin{aligned} \Pi_h^{S4} \;=\; & \mathbf{TL}\left(\int_{\overline{q}}^{1} Z_k\left(qR_{kh}^{TL}(q) - c\right) f(q)dq + Z_u\left(E[q|u] R_{uh}^{TL} - c\right)(1-m)\right) \\ & + \mathbf{TR}\left(Z_{all}\left(E[q] R_h^{TR} - c\right)\right) + \mathbf{BR}\left(Z_{all}\left(E[q] R_h^{BR} - c\right)\right) \\ & + \mathbf{BL}\left(\int_{\overline{q}}^{\frac{c+\tau}{R_h^{BL}}} (qR_h^{BL} - c) f(q)dq + (1-m)\left(E[q|u] R_h^{BL} - c\right)\right) \end{aligned} \tag{A.16}$$

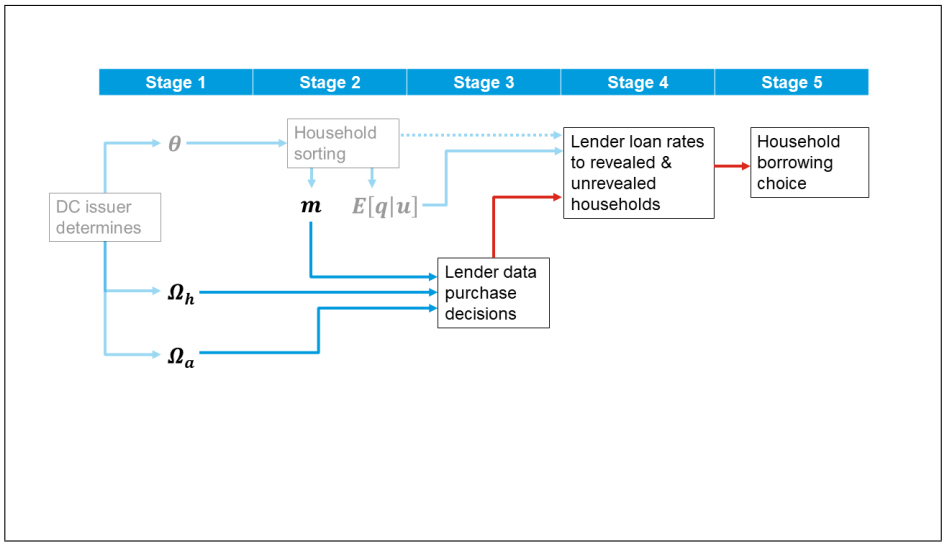
where **TL, TR, BL,** and **TR** are indicator variables that represent the different subgames in Table S4 (i.e., **TL** stands for "top-left", **TR** stands for "top-right", **BL** stands for "bottom-left", and **BR** stands for "bottom-right"). For example, **TL** $= 1$ if all lenders obtained data access at Stage 3 and is 0 otherwise. Similarly, the loan rates in the respective subgames are given the subgame title as a superscript to note that the home lender can choose different rates in each subgame.

Overall, as depicted in the figure at the beginning of this stage's derivation, the Stage 4 loan rate optimization takes as given $E[q|u]$ from Stage 2 household sorting, as well as the data purchase decisions from Stage 3 that are given by **TL, TR, BL,** and **TR**.

From the subgame derivations we have that the optimal Stage 4 play of the home lender is given by

$$R_{kh}^{TL}(q) = \frac{c+\tau}{q},\ R_{uh}^{TL} = R_h^{BL} = \frac{c+\tau}{E[q|u]},\ R_h^{TR} = R_h^{BR} = \frac{4}{3}(c+\tau) \tag{A.17}$$

Stage 3. At Stage 3, lenders foresee the Stage 4 loan rate subgame equilibria and the associated Stage 5 household borrowing choices. They internalize these, shown with red arrows in the figure below, and take as given the DC issuer's data fees, Ω_h and Ω_a, and the mass of revealed households, m, that emerges from household sorting at Stage 2 (which in turn is determined by the DC issuer's choice of θ at Stage 1).



Stage 3 decision: input variables and backward induced outcomes.

Table S3a summarizes the profits in each subgame and Table S3b the subgame equilibria depending on data access pricing determined by the DC issuer at Stage 1. Below, we derive the entries in each cell of these two tables.

Table S3a: Stage 3 lender profit within subgames

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: $\Pi_h = \tau - \Omega_h$ Away lenders: $\Pi_a = -\Omega_a$	*"Top-right" subgame:* Home lender: $\Pi_h = \tau - \Omega_h$ Away lenders: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* Home lender: $\Pi_h = (1-m)\,\tau$ Away lenders: $\Pi_a = -\Omega_a$	*"Bottom-right" subgame:* Home lender: $\Pi_h = \tau$ Away lenders: $\Pi_a = 0$

Table S3b: Data fees determine which cell of Table S3a is the subgame equilibrium

Data fee on away lenders $\rightarrow$ Data fee on home lenders $\downarrow$	$\Omega_a = 0$	$\Omega_a > 0$
$\Omega_h = 0$	Top-left	Top-right
$\Omega_h \in (0, m\tau)$	Top-left	Bottom-right
$\Omega_h > m\tau$	Bottom-left	Bottom-right

We first derive the profits by subgame shown in Table S3a and subsequently the equilibrium play per combination of data fees shown in Table S3b.

Lender profits in subgame 1: all lenders have data access (top-left subgame of Table S3a). When all lenders have data access, the Stage 4 subgame (Table S4, top-left) has revealed households borrowing from the home lender at $R_{kh}(q) = \frac{c+\tau}{q}$ and unrevealed households borrowing from the home lender at $R_{uh} = \frac{c+\tau}{E[q|u]}$. As no household borrows from away lenders in this subgame equilibrium, away lender profit as seen at Stage 3 is simply $\Pi_a = -\Omega_a$: i.e., the away lenders pay Ω_a and get nothing in return. For a given revealed household, the profit that the home lender makes in this subgame is $q\left(\frac{c+\tau}{q}\right) - c = \tau$. This means that the profit (excluding data access fees) that the home lender makes on all revealed households in equilibrium is $m\tau$. Similarly, the home lender profit per unrevealed household is $E[q|u]\left(\frac{c+\tau}{E[q|u]}\right) - c = \tau$ and therefore the total profit on unrevealed households is $(1-m)\,\tau$. Putting these terms together, the total profit on lending to all households by the home lender in this subgame is $m\tau + (1-m)\,\tau = \tau$. Against these profits on lending stands the cost that the home lender pays for data access. Therefore, its profit is $\Pi_h = \tau - \Omega_h$, as seen in the top-left cell of Table S3a.

Lender profits in subgame 2: the home lender has data access, while the away lenders do not (top-right subgame of Table S3a). When only the home lender has data access, the Stage 4 subgame (Table S4, top-right) sees all households borrowing from the home lender at $R_h = \frac{4}{3}(c+\tau)$. As no household borrows from away lenders in this subgame equilibrium, away lender profit as seen at Stage 3 is simply $\Pi_a = 0$: i.e., the away lenders do not pay a data access fee and do not lend in the subgame equilibrium. For a given household, the profit

that the home lender makes in this subgame is $E[q]\left(\frac{c+\tau}{E[q]}\right) - c = \tau$. This means that the profit (excluding data access fees) that the home lender makes on the mass 1 of households is also τ. Against these profits on lending stands the cost that the home lender pays for data access. Therefore, its profit is $\Pi_h = \tau - \Omega_h$, as seen in the top-right cell of Table S3a.

Lender profits in subgame 3: the home lender does not have data access, while the away lenders do (bottom-left subgame of Table S3a). When only the away lenders have data access, then in the Stage 4 subgame (Table S4, bottom-left) all unrevealed households borrow from the home lender at $R_{uh} = \frac{c+\tau}{E[q|u]}$ while all revealed households borrow from away lenders at $R_{ka}(q) = \frac{c+\tau}{q}$. Since $R_{ka}(q) = \frac{c+\tau}{q}$ is the away lenders' breakeven rate, away lenders make 0 profit from lending. Hence, away lenders' total profits are given by the cost of data access: $\Pi_a = -\Omega_a$. The home lender profit per unrevealed household is $E[q|u]\left(\frac{c+\tau}{E[q|u]}\right) - c = \tau$ and therefore the total profit on lending to unrevealed households is $(1-m)\tau$. This is the only entry in the home lender's profit function in this subgame, because it does not lend to revealed households and has not purchased data access. Thus, $\Pi_h = (1-m)\tau$.

Lender profits in subgame 4: all lenders are uninformed (bottom-right subgame of Table S3a). In the Stage 4 subgame between uninformed lenders, all households end up borrowing from the home lender at the same rate: $R_h = \frac{4}{3}(c+\tau)$. For a given household, the profit that the home lender makes in this subgame is $E[q]\left(\frac{c+\tau}{E[q]}\right) - c = \tau$. This means that the profit (excluding data access fees) that the home lender makes on the mass 1 of households is also τ: i.e., $\Pi_h = \tau$. No one borrows from the away lenders and the away lenders did not pay any cost for data access, so that $\Pi_a = 0$.

Table 3Sa summarizes the profits of lenders in the four subgames. These profits are expressed as functions of data fees, Ω_h and Ω_a. We next consider which subgame emerges as the Stage 3 equilibrium depending on the values of Ω_h and Ω_a. These subgame outcomes in relation to data fees are summarized in Table 3Sb.

Subgame equilibrium when $\Omega_h = \Omega_a = 0$. Free data directly places lenders in the top-left subgame of Table S3a because there is no data purchase decision to be made and lenders all have data access.

Subgame equilibrium when $\Omega_h = 0$ and $\Omega_a > 0$. Table S3a now centers on the top row, given that the home lender has been given the data for free. In this case, the profit structure looks as represented below. Therefore, the away lenders choose not to purchase data access and the subgame equilibrium in Table S3a is top-right.

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: $\Pi_h = \tau$ Away lenders: $\Pi_a < 0$	*"Top-right" subgame:* Home lender: $\Pi_h = \tau$ Away lenders: $\Pi_a = 0$

Subgame equilibrium when $\Omega_h \in (0, m\tau)$ and $\Omega_a = 0$. The game dynamics in Table S3a now center on the left column, because away lenders have been granted free data access. Given $\Omega_h \in (0, m\tau)$, it follows that $\tau - \Omega_h > (1-m)\,\tau$ and therefore the home lender earns more profit when it buys the data, as also seen in the table included below. Therefore, the subgame equilibrium in Table S3a is top-left.

	Away lenders have data access
Home lender has data access	*"Top-left" subgame:* Home lender: $\Pi_h = \tau - \Omega_h > (1-m)\,\tau$ Away lenders: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* Home lender: $\Pi_h = (1-m)\,\tau$ Away lenders: $\Pi_a = 0$

Subgame equilibrium when $\Omega_h \in (0, m\tau)$ and $\Omega_a > 0$. Table S3a now looks like this:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: $\Pi_h = \tau - \Omega_h > (1-m)\,\tau$ Away lenders: $\Pi_a < 0$	*"Top-right" subgame:* Home lender: $\Pi_h = \tau - \Omega_h$ Away lenders: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* Home lender: $\Pi_h = (1-m)\,\tau$ Away lenders: $\Pi_a < 0$	*"Bottom-right" subgame:* Home lender: $\Pi_h = \tau > \tau - \Omega_h$ Away lenders: $\Pi_a = 0$

and this means that the evolution of game play can be represented as follows:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: **Stay** Away lenders: *Right*	*"Top-right" subgame:* Home lender: *Down* Away lenders: **Stay**
Home lender has no data access	*"Bottom-left" subgame:* Home lender: *Up* Away lenders: *Right*	*"Bottom-right" subgame:* Home lender: **Stay** Away lenders: **Stay**

where an equilibrium occurs at a settle point where all players "stay" (i.e., have no incentives to adjust their play given the choice of other players): therefore, the bottom-right is the subgame equilibrium.

Subgame equilibrium when $\Omega_h > m\tau$ and $\Omega_a = 0$. The game dynamics in Table S3a now center on the left column, because away lenders have been granted free data access. Given $\Omega_h > m\tau$, it follows that $\tau - \Omega_h < (1-m)\,\tau$ and therefore the home lender earns more profit when it does not buy the data, as also seen in the table included below. Therefore, the subgame equilibrium in Table S3a is bottom-left.

	Away lenders have data access
Home lender has data access	*"Top-left" subgame:* Home lender: $\Pi_h = \tau - \Omega_h$ Away lenders: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* Home lender: $\Pi_h = (1-m)\,\tau > \tau - \Omega_h$ Away lenders: $\Pi_a = 0$

Subgame equilibrium when $\Omega_h > m\tau$ and $\Omega_a > 0$. Table S3a now looks like this:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: $\Pi_h = \tau - \Omega_h < (1-m)\,\tau$ Away lenders: $\Pi_a < 0$	*"Top-right" subgame:* Home lender: $\Pi_h = \tau - \Omega_h$ Away lenders: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* Home lender: $\Pi_h = (1-m)\,\tau$ Away lenders: $\Pi_a < 0$	*"Bottom-right" subgame:* Home lender: $\Pi_h = \tau > \tau - \Omega_h$ Away lenders: $\Pi_a = 0$

and this means that the evolution of game play can be represented as below, and therefore the bottom-right cell is the subgame equilibrium:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: *Down* Away lenders: *Right*	*"Top-right" subgame:* Home lender: *Down* Away lenders: **Stay**
Home lender has no data access	*"Bottom-left" subgame:* Home lender: **Stay** Away lenders: *Right*	*"Bottom-right" subgame:* Home lender: **Stay** Away lenders: **Stay**

This completes the derivations behind Tables S3a and S3b. Based on Tables S3a and S3b, in turn, we can also write down an expression for the optimization problem of the home lender at Stage 3 as we previously did in (A.16) for Stage 4 (while, due to Bertrand competition

among them at Stage 4, the optimal strategy of the away lenders is straightforwardly given by "don't buy data access" whenever $\Omega_a > 0$, as can also be seen from Tables S3a and S3b). We let $\mathbf{AL}$ be an indicator variable that stands for "Away Lender access" (to data) that takes value 1 if away lenders have access (i.e., left column of Table S3a) and 0 otherwise. Moreover, $(1 - \mathbf{AL})$ is then the inverse indicator variable that takes value 1 if and only if away lenders have no access (i.e., right column of Table S3a). Moreover, we let $\mathbf{B}$ be a $0-1$ choice variable for the home lender, where 1 means "Buy data access" and 0 means "don't buy data access" (and the case where $\Omega_h = 0$ automatically implies $\mathbf{B} = 1$ as the data is given for free). Put together, this gives the following Stage 3 optimization problem for the home lender:

$$\max_{\mathbf{B}=\{0,1\}} \Pi_h^{S3} = \mathbf{AL}\left(\left(\tau - \Omega_h\right)\mathbf{B} + \left(\left(1-m\right)\tau\right)\left(1-\mathbf{B}\right)\right) + \left(1-\mathbf{AL}\right)\left(\left(\tau-\Omega_h\right)\mathbf{B} + \tau\left(1-\mathbf{B}\right)\right) \tag{A.18}$$

which can be further simplified to

$$\max_{\mathbf{B}=\{0,1\}} \Pi_h^{S3} = \tau - \Omega_h\mathbf{B} - \mathbf{AL}\left(1-\mathbf{B}\right)m\tau \tag{A.19}$$

This means that the optimal choice of home lender can be found by comparing profits under $\mathbf{B} = 1$, which gives $\Pi_h^{S3} = (\tau - \Omega_h)$, to profits under $\mathbf{B} = 0$, which gives $\Pi_h^{S3} = \tau - \mathbf{AL}m\tau$. This implies the following optimal policy structure:

$$\text{Home lender optimal strategy} \begin{cases} \text{if } \mathbf{AL} = 1, \text{ set } \mathbf{B} = 1 \Leftrightarrow \Omega_h < m\tau \\ \text{if } \mathbf{AL} = 0, \text{ set } \mathbf{B} = 0 \end{cases} \tag{A.20}$$

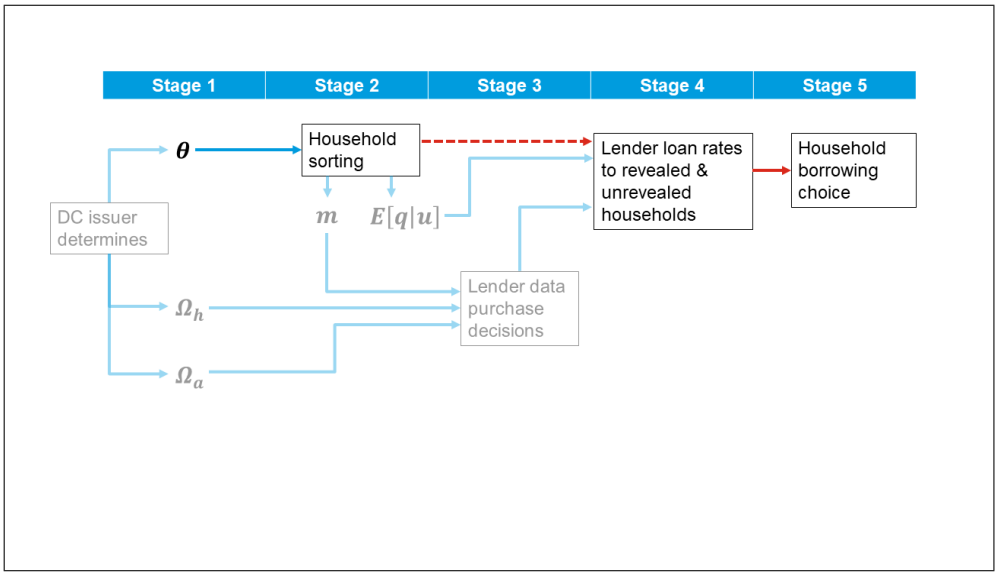
because when $\mathbf{AL} = 1$, the home lender compares $(\tau - \Omega_h)$ under $\mathbf{B} = 1$ to $(1-m)\tau$ under $\mathbf{B} = 0$, where $(\tau - \Omega_h) > (1-m)\tau \Leftrightarrow \Omega_h < m\tau$; and when $\mathbf{AL} = 0$, the home lender compares $(\tau - \Omega_h)$ under $\mathbf{B} = 1$ to τ under $\mathbf{B} = 0$ and $\tau > \tau - \Omega_h$ for any $\Omega_h > 0$.

Stage 2. The Stage 2 decision problem of households differs in an important way from the decision problems of lenders and of the DC issuer: households are atomistic and each individual household has zero impact on aggregate variables. In aggregate, households impact key disclosure and creditworthiness variables in the model, namely the mass of revealed households, m, and the expected creditworthiness of unrevealed households, $E[q|u]$. These affect lender decisions at subsequent stages. But each individual household decides for itself which means of payment to use and does not affect the aggregate. Instead, when choosing between DC and cash, an individual household takes as given DC design θ from Stage 1, and further bases its choice of money on its privacy preference φ and the expected loan rates from Stage 4 that it foresees conditional on its given q and the revelation probability associated with its chosen form of money.

The figure below provides a visual representation of the individual household's problem at Stage 2. Here, the dashed red arrow between Stage 2 and Stage 4 represents the individual's perspective (as opposed to the blue arrows towards m and $E[q|u]$ at Stage 2 that represent the aggregate household impact). The individual household backward induces the outcome of Stage 4, in terms of the loan rates that it will be offered conditional on revelation and, if

revealed, its q.

One aspect that is not captured by the visual representation below, but which will be captured by the mathematical derivations, is the externalities that households impose on each other during the Stage 2 decision. The aggregate decisions of all households determine $E[q|u]$, which affects the expected loan rates at Stage 4, and therefore the individual household's expected benefit of increasing the probability of credit quality disclosure by choosing DC over cash. The Stage 2 equilibrium of household sorting is the result of reaching a fixed point, where given the equilibrium play of all other households, each individual household chooses money in such a way that it brings about the equilibrium play of all households. I.e., one can see this as an "iteration" between households' response given $E[q|u]$ and the resulting $E[q|u]$ that emerges from households' response. This is not solved through a numerical iteration but rather through a fixed point problem the solution to which is found analytically, as discussed below.



Stage 2 decision: input variables and backward induced outcomes.

At Stage 2, a household prefers DC over cash if and only if

$$q(y - R_{dc}) - \alpha\varphi\theta > q(y - R_{cash}) \tag{A.21}$$

where R_{dc} and R_{cash} are, respectively, the loan rates that the household would expect if it chooses DC or cash.

Backward inducing from Stage 4, there are two possibilities regarding R_{dc}. If the Stage 4 subgame equilibrium is in the right column of Table S4 (top-right or bottom-right), then all households are offered the same loan rates. There is no type differentiation in loan rates, which means that $R_{dc} = R_{cash}$. That is, based on the entries in Table S4, the tradeoff in (A.21) becomes that a household prefers DC over cash if and only if

$$q\left(y-\frac{4}{3}(c+\tau)\right)-\alpha\varphi\theta \;>\; q\left(y-\frac{4}{3}(c+\tau)\right)\Leftrightarrow \tag{A.22}$$

$$-\alpha\varphi\theta \;>\; 0 \tag{A.23}$$

which never holds. That is, if the Stage 4 subgame equilibrium yields undifferentiated loan rates, then all households, foreseeing this, choose cash, because DC comes with a privacy cost but no benefit in terms of lower loan rates. Therefore, $m=0$, because cash users never become revealed, and hence $E[q|u]=E[q]=\frac{3}{4}$.

Instead, if households at Stage 2 foresee that at Stage 4 the subgames in the left column of Table S4 will prevail, then they know that type differentiated loan rates will be offered to them. For both the top-left and bottom-left subgames, the equilibrium loan rates are $R_u=\frac{c+\tau}{E[q|u]}$ for households that are unrevealed and $R_k(q)=\frac{c+\tau}{q}$ for revealed households, as seen from Table S4.

That is, from a household's perspective, the prospective loan rates, depending on the prevailing Stage 4 subgame equilibrium, look as follows:

Table S2a: Households' expected loan rates depending on foreseen Stage 4 subgame equilibrium

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Revealed: $R_k(q)=\frac{c+\tau}{q}$ Unrevealed: $R_u=\frac{c+\tau}{E[q\|u]}$	*"Top-right" subgame:* $R=\frac{4}{3}(c+\tau)$
Home lender has no data access	*"Bottom-left" subgame:* Revealed: $R_k(q)=\frac{c+\tau}{q}$ Unrevealed: $R_u=\frac{c+\tau}{E[q\|u]}$	*"Bottom-right" subgame:* $R=\frac{4}{3}(c+\tau)$

and households' equilibrium sorting in response is

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* The differentiated case $m=?,\ E[q\|u]=?$	*"Top-right" subgame:* Everyone chooses cash $m=0,\ E[q\|u]=E[q]=\frac{3}{4}$
Home lender has no data access	*"Bottom-left" subgame:* The differentiated case $m=?,\ E[q\|u]=?$	*"Bottom-right" subgame:* Everyone chooses cash $m=0,\ E[q\|u]=E[q]=\frac{3}{4}$

The remainder of the Stage 2 derivation is devoted to solving the differentiated case (i.e., finding the solutions to the question marks in the table above). In the differentiated case, $R_{dc}\neq R_{cash}$, depending on q. We thus write $R_{dc}(q)$, where, at Stage 2, a household prefers DC over cash if and only if

$$q(y-R_{dc}(q))-\alpha\varphi\theta > q(y-R_{cash}) \tag{A.24}$$

From Table S2a, it follows that the loan rate on cash users is $R_{cash} = R_u = \frac{c+\tau}{E[q|u]}$. Instead, $R_{dc}(q) = \theta R_k(q) + (1-\theta) R_u$, because a DC user's credit quality is revealed with probability θ in which case that household borrows at the equilibrium loan rate $R_k(q) = \frac{c+\tau}{q}$, as shown in Table S2a, while with probability $1-\theta$ the household remains unrevealed and pays $R_u = \frac{c+\tau}{E[q|u]}$ on its loan. We can now simplify (A.24) to:

$$\alpha\varphi < q\left(R_u - R_k(q)\right) \tag{A.25}$$

and replacing from $R_k(q) = \frac{c+\tau}{q}$ and $R_u = \frac{c+\tau}{E[q|u]}$, this becomes:

$$\varphi \leq \frac{c+\tau}{\alpha}\left(\frac{q}{E[q|u]} - 1\right) \tag{A.26}$$

The sorting condition in (A.26) has $E[q|u]$ as an unsolved variable. To come to a full solution for household sorting, a closed-form expression for $E[q|u]$ is needed. As a first step towards this solution, we first note that $(1-m)E[q|u] + mE[q|dc] = E[q] = \frac{3}{4}$, where $E[q|dc]$ is the expected quality of DC users (and since all revealed households are randomly drawn DC users, this equals the expected quality of revealed households). This can be written to

$$E[q|u] = \frac{3/4 - mE[q|dc]}{1-m} \tag{A.27}$$

To make further headway, the next step requires an expression for m. We note that $m = \theta\mu$ where μ is the mass of DC users. I.e., the mass of revealed households equals the mass of DC users times the probability of DC user type revelation. We can find μ by integrating over the indifference frontier displayed as the unbroken line in Figure 2. Defining the area of integration requires finding two values of q: 1) the lowest q household that is a DC user (which is the starting point of the integral); 2) the lowest value of q above which all users (regardless of φ) are DC users. From the first to the second point, we integrate over the indifference frontier and from the second point to $q = 1$ we integrate over a uniformly distributed mass of households.

To find point 1), we note that only DC users can become revealed and, thus, the lowest quality a revealed household can have is that of the lowest quality DC user. The credit quality of this household is obtained by setting $\varphi = 0$ in (A.26) because, as can be seen from Figure 2, the DC user with the lowest privacy costs is also the DC user with the lowest credit quality. This yields $0 = c\left(\frac{q}{E[q|u]} - 1\right) \Leftrightarrow q = E[q|u]$. Hence, the lowest quality revealed household has $q = E[q|u]$ and all other revealed households have $q > E[q|u]$.

Point 2) is derived from entering $\varphi = 2$ in (A.26) and setting (A.26) to equality, because beyond point 2) all households, including those with the highest privacy preference, choose DC (see also Figure 2). This yields $2\alpha = (c+\tau)\left(\frac{q}{E[q|u]} - 1\right)$ and can be written to $q = \frac{2\alpha+c+\tau}{c+\tau}E[q|u]$. Here, we note that $\frac{1}{2} \leq E[q|u] < \frac{2\alpha+c+\tau}{c+\tau}E[q|u] < 1$, where $\frac{2\alpha+c+\tau}{c+\tau}E[q|u] < 1$ follows from the condition in (6).[2] This, in turn, ensures that the two pieces of integration

[2] I.e., $\sup E[q|u] = E[q] = \frac{3}{4}$ which implies $\sup \frac{2\alpha+c+\tau}{c+\tau}E[q|u] = \frac{2\alpha+c+\tau}{c+\tau}\frac{3}{4}$ and this is smaller than 1 given $\alpha < \frac{1}{6}(c+\tau)$ from (6).

shown in the equation below are properly defined, given $q \in \left[\frac{1}{2}, 1\right]$. We can now write

$$\mu = \int_{E[q|u]}^{\frac{2\alpha+c+\tau}{c+\tau}E[q|u]} f(q)\, dq + \int_{\frac{2\alpha+c+\tau}{c+\tau}E[q|u]}^{1} 2dq \tag{A.28}$$

$$f(q) = \frac{c+\tau}{\alpha}\left(\frac{q}{E[q|u]} - 1\right) \tag{A.29}$$

where $f(q)$ is derived from setting (A.26) to equality and writing to φ. Moreover, the 2 in the second integral in (A.28) comes from the uniform distribution over $q \in \left[\frac{1}{2}, 1\right]$. We note that $E[q|u]$ is a function of θ (derived further below). Solving (A.28) gives:

```
In[5]:= FullSimplify[∫_Equ^((2 α+c+τ)/(c+τ) Equ) (c + τ)/α (q/Equ - 1) dq + ∫_((2 α+c+τ)/(c+τ) Equ)^1 2 dq]

Out[5]= 2 - (2 Equ (c + α + τ))/(c + τ)
```

Mathematica calculation of μ.

so that

$$\mu = 2\left(1 - \frac{\alpha+c+\tau}{c+\tau}E[q|u]\right) \tag{A.30}$$

and thus

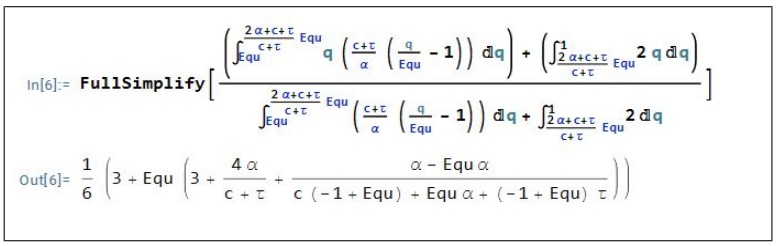
$$m = 2\theta\left(1 - \frac{\alpha+c+\tau}{c+\tau}E[q|u]\right) \tag{A.31}$$

where the term in parentheses is always positive given $E[q|u] \in \left[\frac{1}{2}, \frac{3}{4}\right]$ and $\alpha < \frac{1}{6}(c+\tau)$ from (6).

We now have the elements needed to progress in solving for $E[q|u]$ from (A.27), where $E[q|u]$ is expressed as a function of m and $E[q|dc]$. We have an expression for m from (A.31) and $E[q|dc]$ can be found as a weighted average of expected values with respect to the integral areas identified in (A.28):

$$E[q|dc] = \frac{\int_{E[q|u]}^{\frac{2\alpha+c+\tau}{c+\tau}E[q|u]} q\left(\frac{c+\tau}{\alpha}\left(\frac{q}{E[q|u]} - 1\right)\right) dq + \int_{\frac{2\alpha+c+\tau}{c+\tau}E[q|u]}^{1} 2q dq}{\int_{E[q|u]}^{\frac{2\alpha+c+\tau}{c+\tau}E[q|u]} \frac{c+\tau}{\alpha}\left(\frac{q}{E[q|u]} - 1\right) dq + \int_{\frac{2\alpha+c+\tau}{c+\tau}E[q|u]}^{1} 2dq} \tag{A.32}$$

and solving this gives



Mathematica calculation of $E[q|\,dc]$.

which can be written to:

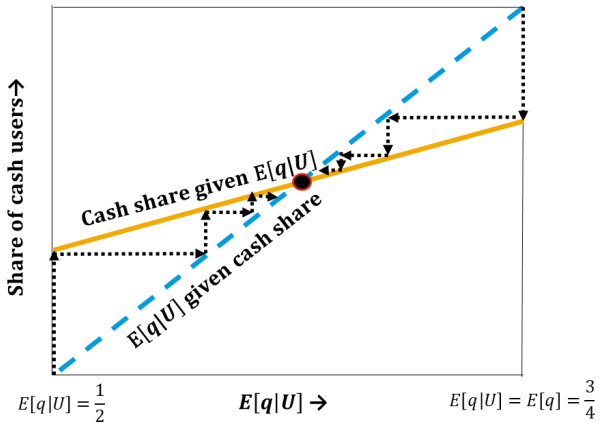
$$E[q|\,dc] = 2\left(1 + E[q|\,u]\right) - \frac{\alpha E[q|\,u]}{6}\left(\frac{1 - E[q|\,u]}{(c+\tau)\left(1 - E[q|\,u]\right) - \alpha E[q|\,u]} - \frac{4}{c+\tau}\right) \quad \text{(A.33)}$$

Replacing from (A.31) and (A.33) into (A.27) and solving leads to:

$$E[q|\,u] = \frac{\begin{array}{c}(c^2+\tau^2)\left(3 + 4\left(-1 + E[q|\,u]^2\right)\theta\right) + \frac{4}{3}E[q|\,u]^2\,\theta\,(4\alpha^2 + 6\alpha\tau) \\ +2c\left((3-4\theta)\,\tau + 4E[q|\,u]^2\,\theta\,(\alpha+\tau)\right)\end{array}}{4\,(c+\tau)\left(c - 2c\left(1 - E[q|\,u]\right)\theta + \tau(1-2\theta) + 2E[q|\,u]\,\theta\,(\alpha+\tau)\right)} \quad \text{(A.34)}$$

Equation (A.34) is the mathematical representation of the fixed-point problem we mentioned before: $E[q|\,u]$ is a function of $E[q|\,u]$, because for any given $E[q|\,u]$ households respond in ways that determine their aggregate sorting, and that sorting has an associated value of $E[q|\,u]$. We can solve this fixed point problem in closed form for $E[q|\,u]$, as shown in Figure A1 (on page 23 of this appendix).

In addition, the figure below provides a visualization of the fixed point problem. For simplicity, this figure is drawn for $\theta = 1$ so that all DC users are revealed and all cash users are unrevealed (instead, the math in Figure A1 keeps θ general). $E[q|\,u]$ is on the horizontal axis and ranges from $\frac{1}{2}$ (if all but the worst quality households were to use DC) to $\frac{3}{4}$ (if all households used cash and therefore $E[q|\,u] = E[q]$). On the vertical axis, the figure plots the share of cash users (i.e., $1-m$). One can "start the convergence towards the fixed point" from any initial prior on $E[q|\,u]$ in $\left[\frac{1}{2}, \frac{3}{4}\right]$. For example, assume initially that all households use cash and therefore $E[q|\,u] = E[q] = \frac{3}{4}$. Given this $E[q|\,u]$, some households (with relatively high q and low φ) will find that the condition in (A.26) is satisfied, and therefore choose to use DC. I.e., the assumption that everyone uses cash is untrue. The share of cash use that emerges given $E[q|\,u] = E[q] = \frac{3}{4}$ is depicted as the right-end of the solid line in this figure. In turn, as relatively high q households have chosen to use DC, the remaining cash users are of lower average creditworthiness and $E[q|\,u]$ will be lower than $\frac{3}{4}$. This is shown by the dashed line. The "iteration" between household optimization and the emerging $E[q|\,u]$ is shown by the dotted arrows between the two lines. The fixed point solution of the process is shown as the black dot where the two lines cross.



Visual intuition for the fixed-point problem.

From the derivation shown in Figure A1, the solution for $E[q|u]$ is:

$$E[q|u] = \frac{2\theta - 1 + \sqrt{1 - \theta + \frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2 (4\theta - 3)\,\theta}}{2\theta\left(1 - \frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2\right)} \tag{A.35}$$

The calculations shown in Figure A1 also verify that for the solution in (A.35), it indeed holds that $E[q|u] \in \left[\frac{1}{2}, \frac{3}{4}\right]$. We also note that the apparent singularity at $\theta = 0$ is not a genuine singularity. Its appearance comes from the fact that (A.35) is the solution of (A.34) when (A.34) is written as a quadratic equation. However, at $\theta = 0$, (A.34) simplifies to $E[q|u] = \frac{3}{4}$ and there is no quadratic equation to solve: when $\theta = 0$, all households are necessarily unrevealed and $E[q|u] = E[q] = \frac{3}{4}$.

Next, we replace from (A.35) into (A.31) and use the calculations shown in Figure A2 (on page 24 of this appendix) to obtain

$$m = \frac{2\theta\left((c+\tau)^2 - \frac{4}{3}\alpha^2\right) - (c+\tau)(\alpha + c + \tau)\left((2\theta - 1) + \sqrt{1 - \theta + \frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2 (4\theta - 3)\,\theta}\right)}{(c+\tau)^2 - \frac{4}{3}\alpha^2} \tag{A.36}$$

where we also conduct a numerical cross-check in Figure A2, showing that $m \in [0, 1]$, as expected.

Putting all the Stage 2 derivations together, now allows us to express the aggregate sorting of households in relation to the prevailing subgame equilibrium among the lenders.

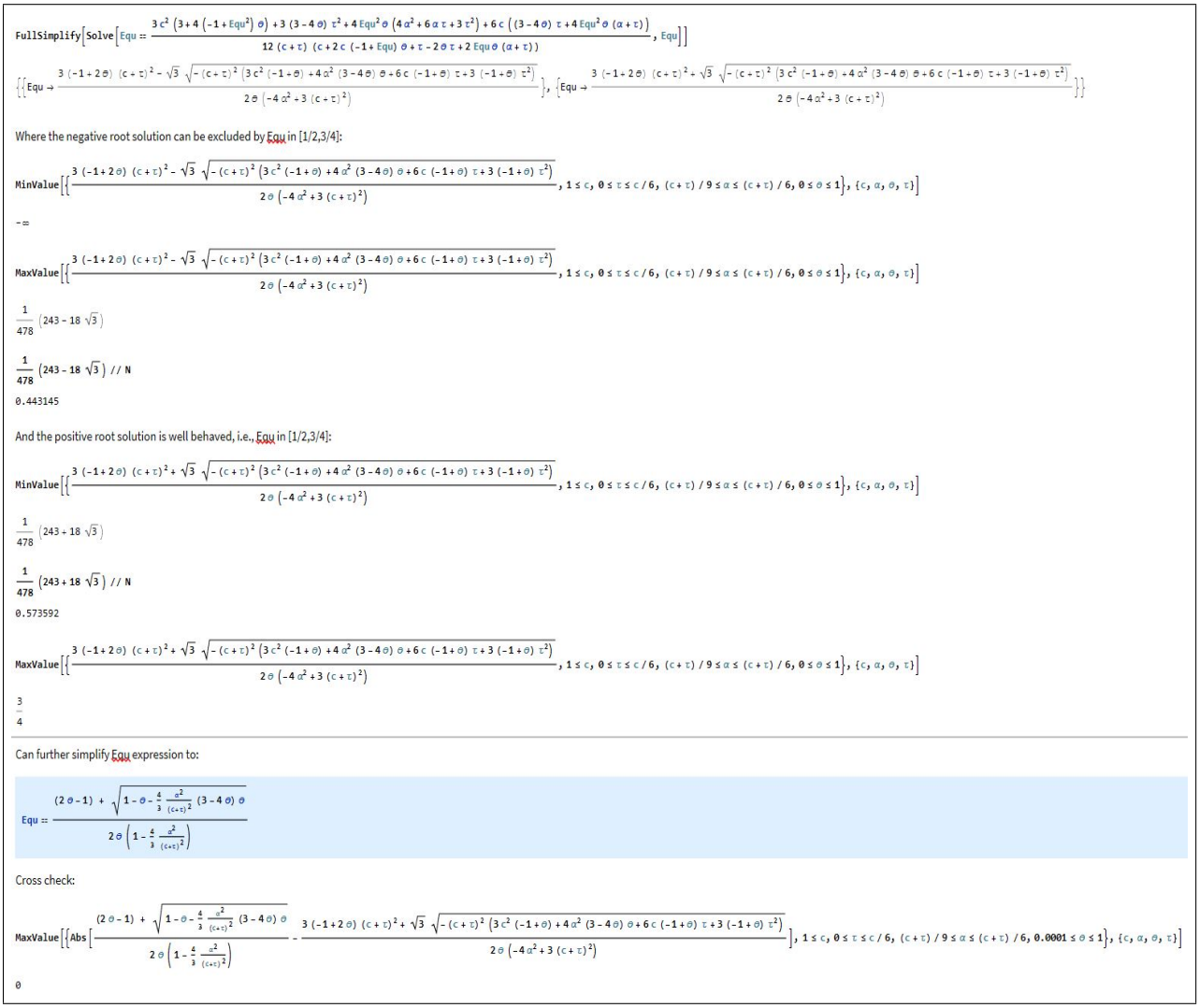


Figure A1: Solving the fixed point problem for $E[q|u]$ in *Mathematica*.

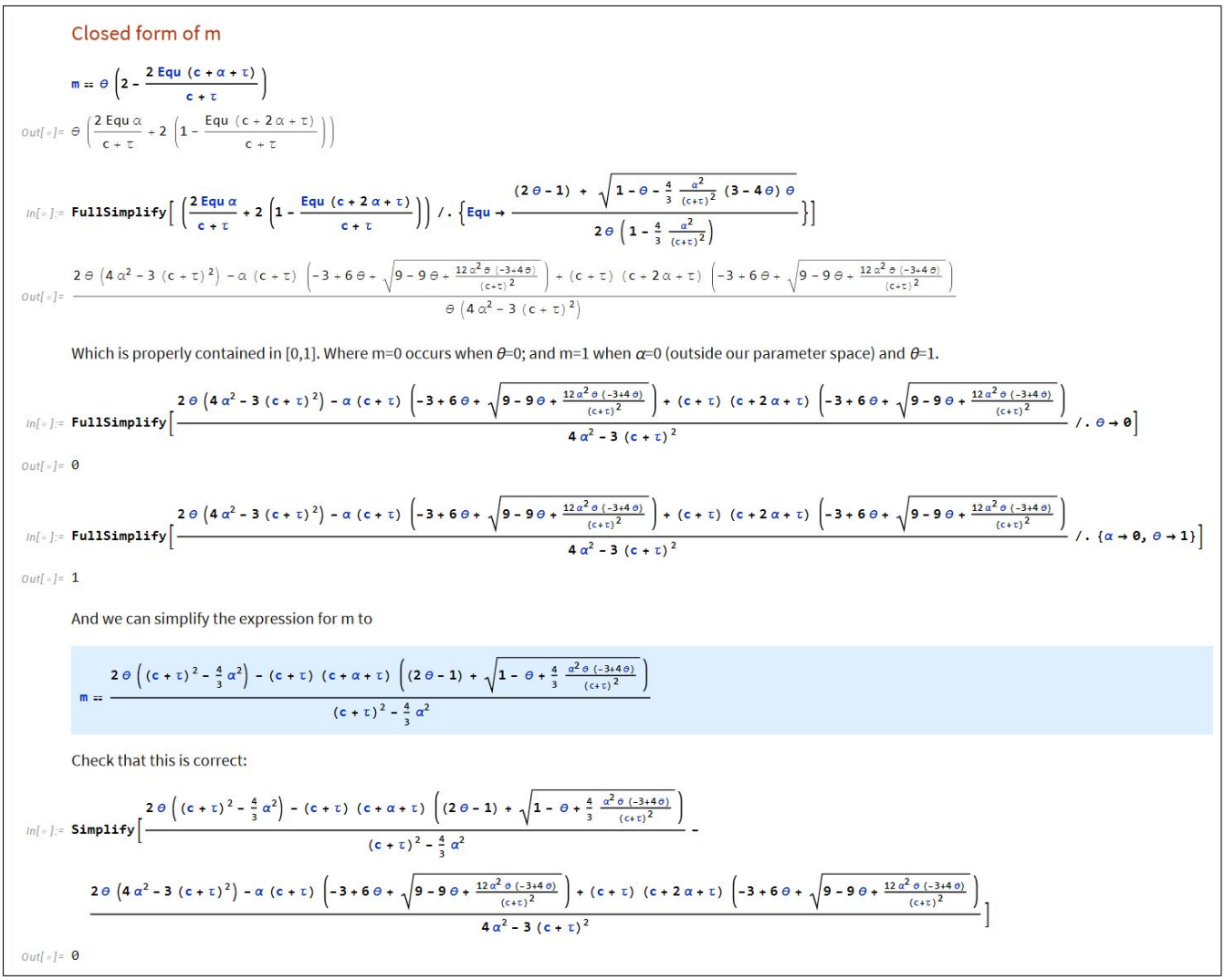
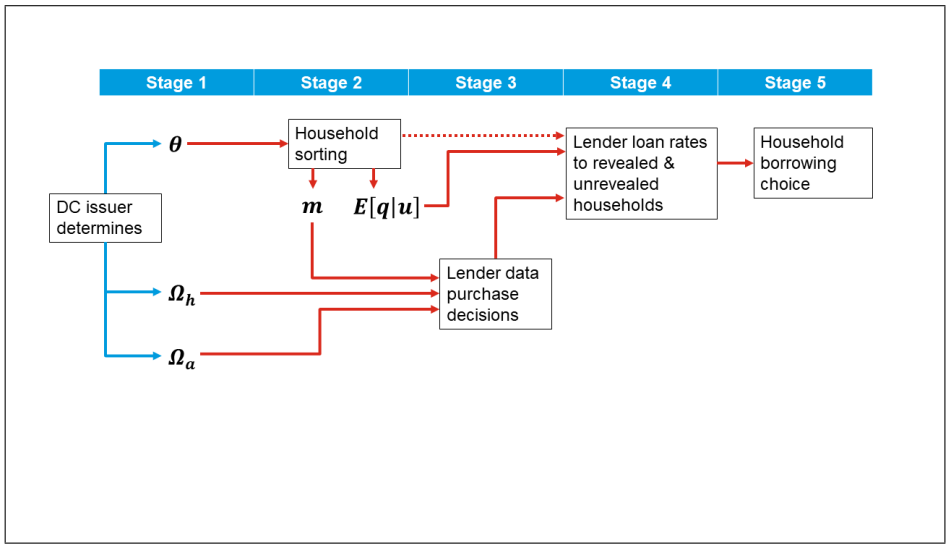


Figure A2: Mathematica derivation of closed form for m.

Table S2b: Aggregate household sorting in relation to foreseen lender subgame equilibria

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* The differentiated case $E[q\|u]$ and m, respectively, given by (A.35) and (A.36)	*"Top-right" subgame:* Everyone chooses cash $m = 0$, $E[q\|u] = E[q] = \frac{3}{4}$
Home lender has no data access	*"Bottom-left" subgame:* The differentiated case $E[q\|u]$ and m, respectively, given by (A.35) and (A.36)	*"Bottom-right" subgame:* Everyone chooses cash $m = 0$, $E[q\|u] = E[q] = \frac{3}{4}$

Stage 1. At Stage 1, the DC issuer internalizes the entire game: the issuer considers how its choice of θ affects household sorting, which determines m and $E[q|u]$; how its data fees together with m shape lender data purchase decisions; how those data purchase decisions jointly with $E[q|u]$ affect the lender loan market equilibrium, and subsequent household borrowing. The way it solves for this is through backward induction. As shown in the figure below, it backward induces the entire game to come to an optimal choice of θ, Ω_h and Ω_a.



Stage 1 decision: backward induction.

Optimal data fees. The objective of the DC issuer is to maximize the profit it makes from selling data. We can refer to the DC issuer profit as π (to differentiate its notation from Π lender profits) and its highest attainable profits (under optimal choice of data fees and θ) as $\widehat{\pi}$. In Table S1, we first express these profits relative to the ranges of data fees (in line with Table S3b) and then discuss how we come to the entries in Table S1, based on the backward induction from the later stages of the game.

Table S1: Profits from data sales

Data fee on away lenders → Data fee on home lenders ↓	$\Omega_a = 0$	$\Omega_a > 0$
$\Omega_h = 0$	$\pi = 0$	$\pi = 0$
$\Omega_h \in (0, m\tau)$	$\pi = \Omega_h \Rightarrow \widehat{\pi} = m\tau$	$\pi = 0$
$\Omega_h > m\tau$	$\pi = 0$	$\pi = 0$

If the DC issuer sets $\Omega_h = \Omega_a = 0$: data revenues are necessarily 0 because lenders get free data access.

If the DC issuer sets $\Omega_h = 0$ *and* $\Omega_a > 0$: the Stage 3 subgame equilibrium is top-right (Table S3b), which means that the away lenders don't buy data access. Since the home lender is given the data for free, this implies $\pi = 0$.

If the DC issuer sets $\Omega_h \in (0, m\tau)$ *and* $\Omega_a = 0$: the Stage 3 subgame equilibrium is top-left (Table S3b), which means that the home lender purchases the data at fee $\Omega_h \in (0, m\tau)$. The optimal data fee is now obviously the highest data fee within the range $\Omega_h \in (0, m\tau)$, which is $m\tau - \varepsilon$ with $\varepsilon \to 0$, which we can express as the limit case where Ω_h is set to $m\tau$ as the optimal data fee on the home lender, from which also follows $\widehat{\pi} = \widehat{\Omega_h} = m\tau$. Here, at Stage 2, households foresee the top-left subgame equilibrium for loan rates and the resulting household sorting implies m as given by the expression in (A.36), as can be seen from Table S2b. Hence, we can express $\widehat{\Omega_h} = m\tau$ as

$$\widehat{\Omega_h} = \frac{2\theta\left((c+\tau)^2 - \frac{4}{3}\alpha^2\right) - (c+\tau)(\alpha+c+\tau)\left((2\theta-1) + \sqrt{1-\theta+\frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2(4\theta-3)\,\theta}\right)}{(c+\tau)^2 - \frac{4}{3}\alpha^2}\tau \tag{A.37}$$

If the DC issuer sets $\Omega_h \in (0, m\tau)$ *and* $\Omega_a > 0$: the Stage 3 subgame equilibrium is bottom-right (Table S3b), in which no lender buys the data. As away lenders are unwilling to buy data at a positive fee, the home lender can make equally much profit off lending to households when it is uninformed as when it is informed (Table S3a) and therefore opts against buying data access. When no lender buys data access, $\pi = 0$.

If the DC issuer sets $\Omega_h > m\tau$ *and* $\Omega_a = 0$: the Stage 3 subgame equilibrium is bottom-left (Table S3b). The home lender finds data access too expensive, because, purchasing it would lower its profit (Table S3a). Since it does not buy data access and the away lenders are given the data for free, the DC issuer has no data revenues and $\pi = 0$.

If the DC issuer sets $\Omega_h > m\tau$ *and* $\Omega_a > 0$: the Stage 3 subgame equilibrium is bottom-right (Table S3b). Neither home nor away lenders purchase the data and therefore $\pi = 0$.

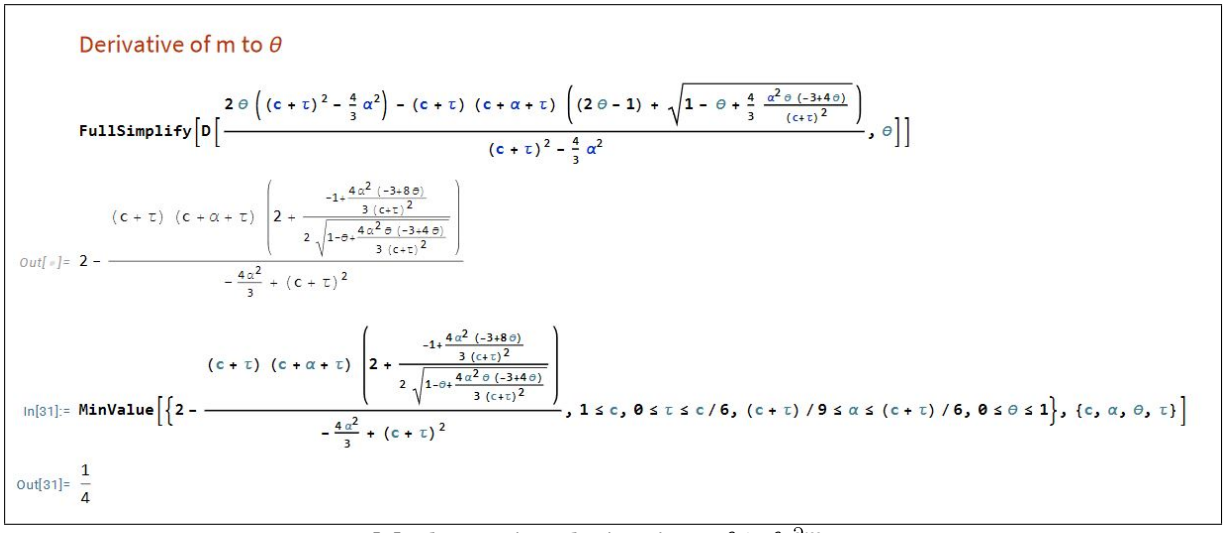
Given these derivations, summarized in Table S1, it is immediately apparent that the optimal data fee structure for the home lender is $\Omega_h = \widehat{\Omega}_h$ and $\Omega_a = 0$. Only the structure with $\Omega_h \in (0, m\tau)$ and $\Omega_a = 0$ leads to positive profits for the DC issuer, and within that structure profits are maximized at $\Omega_h = \widehat{\Omega}_h$.

Optimal θ. Lastly, we turn to finding the optimal θ for the DC issuer. We note that θ is determined jointly with the data fees. However, as shown above, the optimal strategy for data access fees is not affected by θ, in the sense that the expression for $\widehat{\Omega_h}$ is always given by (A.37). But θ does affect the value of $\widehat{\Omega_h}$ in (A.37). That is, our solution technique centers on first finding $\widehat{\Omega_h}$ and then the θ that maximizes $\widehat{\Omega_h}$. This technique, while presented sequentially in the proofs, ultimately yields the joint optimal strategy for θ and the data fees.

In particular, maximizing profit to θ means finding the θ that gives the highest value of $\widehat{\Omega}_h$ in (A.37). We find that

$$\inf \frac{\partial \widehat{\Omega}_h}{\partial \theta} = \inf \frac{\partial m}{\partial \theta} \tau = \frac{1}{4} \tau > 0 \tag{A.38}$$

where we use the *MinValue* function in *Mathematica*, as shown below, to find $\inf \frac{\partial m}{\partial \theta}$ over our allowed parameter range. Therefore, the DC issuer optimally sets $\theta = 1$.



Mathematica derivation of $\inf \frac{\partial m}{\partial \theta}$.

Algebraic expression. The DC issuer optimization problem can be expressed as:

$$\max_{\Omega_h, \Omega_a, \theta} \left\{ \pi = \begin{array}{c} (\Omega_h + (N-1)\,\Omega_a)\,\mathbf{B}\left(\Omega_h, m\left(\theta\right), \mathbf{AL}\left(\Omega_a\right)\right)\mathbf{AL}\left(\Omega_a\right) \\ +\Omega_h \mathbf{B}\left(\Omega_h, m\left(\theta\right), \mathbf{AL}\left(\Omega_a\right)\right)\left(1 - \mathbf{AL}\left(\Omega_a\right)\right) \\ +(N-1)\,\Omega_a\left(1 - \mathbf{B}\left(\Omega_h, m\left(\theta\right), \mathbf{AL}\left(\Omega_a\right)\right)\right)\mathbf{AL}\left(\Omega_a\right) \end{array} \right\} \tag{A.39}$$

where we recall from (A.18) that $\mathbf{B}$ and $\mathbf{AL}$ are indicator variables that take value 1 when, respectively, the home lender and away lenders choose to purchase data access, and take value 0 otherwise. The structure of the profit function is in line with the columns in Tables S4, S3a, S2a and S2b: $\mathbf{B} * \mathbf{AL}$ is the "top-left" case where both the home lender and away lenders have data access, in which case the DC issuer earns data fees $(\Omega_h + (N-1)\,\Omega_a)$; $\mathbf{B} * (1 - \mathbf{AL})$ represents the "top-right" case where only the home lender has data access and where the DC issuer earns Ω_h; $(1 - \mathbf{B})\,\mathbf{AL}$ is the "bottom-left" case where only the away lenders have data access and the DC issuer earns Ω_a on each of the $(N-1)$ away lenders; and, lastly, the "bottom-right" case where no lender has data access is not shown in this expression, because it comes with 0 data sale revenues.

As seen by the DC issuer at Stage 1, $\mathbf{B}$ and $\mathbf{AL}$ are functions of its decision variables, Ω_h, Ω_a and θ. In particular, backward inducing from Stage 3, as seen in Table S3b, $\mathbf{AL}(\Omega_a) = 1 \Leftrightarrow \Omega_a = 0$ and $\mathbf{AL}(\Omega_a) = 0$ otherwise (i.e., for any $\Omega_a > 0$). This directly implies that the third line in (A.39) equals 0 because either $\Omega_a = 0$ or $\mathbf{AL}(\Omega_a) = 0$. We can therefore write the profit maximization problem to

$$\max_{\Omega_h, \Omega_a, \theta} \left\{ \pi = \begin{array}{c} (\Omega_h + (N-1)\,\Omega_a)\,\mathbf{B}(\Omega_h, m(\theta), \mathbf{AL}(\Omega_a))\,\mathbf{AL}(\Omega_a) \\ +\Omega_h \mathbf{B}(\Omega_h, m(\theta), \mathbf{AL}(\Omega_a))\,(1 - \mathbf{AL}(\Omega_a)) \end{array} \right\} \tag{A.40}$$

Next, we note that $\mathbf{B}$ is written as the function $\mathbf{B}(\Omega_h, m(\theta), \mathbf{AL}(\Omega_a))$. Backward inducing (A.20), the DC issuer knows that $\mathbf{B} = 1 \Leftrightarrow \Omega_h < m(\theta)\,\tau \wedge \mathbf{AL}(\Omega_a) = 1$ and $\mathbf{B} = 0$ otherwise. In other words, as can be seen from Table S3b, the Stage 3 purchase decision of the home lender depends on Ω_h, m (which is determined at Stage 2 household sorting and is a function of θ per (A.36)), and away lenders' data purchase decision (i.e., given $\mathbf{AL}(\Omega_a) = 0$, $\mathbf{B} = 0$ is the optimal play for the home lender, as also seen in Table S3a). It follows that $\Omega_h \mathbf{B}(\Omega_h, m(\theta), \mathbf{AL}(\Omega_a))\,(1 - \mathbf{AL}(\Omega_a)) = 0$ in (A.40) because either $(1 - \mathbf{AL}(\Omega_a)) = 0$ (when $\mathbf{AL}(\Omega_a) = 1$) or $\mathbf{B}(\Omega_h, m(\theta), \mathbf{AL}(\Omega_a)) = 0$ (when $\mathbf{AL}(\Omega_a) = 0$). This means that we can further simplify (A.40) to

$$\max_{\Omega_h, \Omega_a, \theta} \{\pi = (\Omega_h + (N-1)\,\Omega_a)\,\mathbf{B}(\Omega_h, m(\theta), \mathbf{AL}(\Omega_a))\,\mathbf{AL}(\Omega_a)\} \tag{A.41}$$

In (A.41), $\mathbf{AL}(\Omega_a) = 0 \Rightarrow \pi = 0$. Therefore, $\Omega_a = 0$ must be optimal for the DC issuer, because only with $\mathbf{AL}(\Omega_a) = 1$ can positive profits be possible. Thus, $\widehat{\Omega}_a = 0$, and we can next write (A.41) to

$$\max_{\Omega_h, \theta} \{\pi = \Omega_h \mathbf{B}(\Omega_h, m(\theta))\} \tag{A.42}$$

where $\mathbf{B} = 1 \Leftrightarrow \Omega_h < m(\theta)\,\tau$ implies that $\widehat{\Omega}_h = m(\theta)\,\tau - \varepsilon$ with $\varepsilon \to 0$, which can be written as the limit case $\widehat{\Omega}_h = m(\theta)\,\tau$. Moreover, $m'(\theta) > 0$ from (A.38) implies that $\widehat{\Omega}_h$ is maximized at $\theta = 1$. Therefore, $\widehat{\Omega}_h = m(\theta)\,\tau$, $\widehat{\Omega}_a = 0$ and $\theta = 1$ are the optimal policy combination for the DC issuer (where the closed form for $m(\theta)$ is given by (A.36)).

Subgame Perfect Nash Equilibrium: Summary. Overall, through backward induction, we find the unique Subgame Perfect Nash Equilibrium, which has the following

equilibrium strategies:

At Stage 1, the DC issuer, foreseeing the entire game, optimally sets $\theta = 1$, $\Omega_a = 0$, and $\Omega_h = \widehat{\Omega}_h$, with the solution for $\widehat{\Omega}_h$ being given by (A.37). Moreover, given $\theta = 1$ the expression for $\widehat{\Omega}_h$ from (A.37) becomes

$$\widehat{\Omega}_h\Big|_{\theta=1} = \frac{2\left((c+\tau)^2 - \frac{4}{3}\alpha^2\right) - (c+\tau)(\alpha+c+\tau)\left(1+\frac{2}{3}\left(\frac{\alpha}{c+\tau}\right)\right)}{(c+\tau)^2 - \frac{4}{3}\alpha^2}\tau \tag{A.43}$$

At Stage 2, households foresee what loan rates they can expect given the DC issuer's Stage 1 choices and the decisions this will induce among the lenders. The households sorting decisions imply that the mass of revealed households is given by (A.36), which is also the expression shown in (11) in the statement of Proposition 1 in Section 4.1. Furthermore, setting $\theta = 1$ in (A.36), we can express $m|_{\theta=1}$ as:

$$m|_{\theta=1} = \frac{2\left((c+\tau)^2 - \frac{4}{3}\alpha^2\right) - (c+\tau)(\alpha+c+\tau)\left(1+\frac{2}{3}\left(\frac{\alpha}{c+\tau}\right)\right)}{(c+\tau)^2 - \frac{4}{3}\alpha^2} \tag{A.44}$$

Similarly, the expected quality of an unrevealed household is given by (A.35), which is also the expression shown in (10) in the statement of Proposition 1. Moreover, setting $\theta = 1$ in (A.35)) we obtain $E[q|u]|_{\theta=1}$ as:

$$E[q|u]|_{\theta=1} = \frac{1+\frac{2}{3}\left(\frac{\alpha}{c+\tau}\right)}{2\left(1-\frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2\right)} \tag{A.45}$$

Households opt for DC use when their q and φ are such that they satisfy (A.26), which is also the expression shown in (9) in the statement of Proposition 1. We can further write the sorting condition given that the DC issuer sets $\theta = 1$ as:

$$\varphi \leq \frac{c+\tau}{\alpha}\left(\frac{1-\frac{4}{3}\left(\frac{\alpha}{c+\tau}\right)^2}{1+\frac{2}{3}\left(\frac{\alpha}{c+\tau}\right)}2q - 1\right) \tag{A.46}$$

At Stage 3, the home lender buys data access at the fee shown in (A.43), while the away lenders have obtained free data access.

At Stage 4, this leads to loan market equilibrium in which all lenders offer revealed households type-dependent loan rates (in line with (7) in the statement of Proposition 1):

$$R_k(q) = \frac{c+\tau}{q} \tag{A.47}$$

and all lenders offer unrevealed households the single loan rate (in line with (8) in the statement of Proposition 1):

$$R_u = \frac{c+\tau}{E[q|u]} \tag{A.48}$$

which, replacing $E[q|u]$ from (A.45) can also be written in the following way given $\theta = 1$:

$$R_u|_{\theta=1} = \frac{2\left(c+\tau-\frac{2}{3}\alpha\right)}{1+\frac{2}{3}\left(\frac{\alpha}{c+\tau}\right)} \tag{A.49}$$

Lastly, at Stage 5 all households borrow from their home lender at the above rates.

A.2 Proof of Proposition 2

This proof is structured by backward induction. We first note that Stage 5 play by a household is the same as for Proposition 1: a household borrows from the lender that offers it the lowest loan rate. The proof therefore focuses on backward induction from Stage 4 to Stage 1. This proof spends less detail on each stage than the Proof of Proposition 1, so as to avoid repetition (on, for example, which players take as given which variables and internalize which future stages, which is the same as in Proposition 1).

Stage 4. The table below summarizes the Stage 4 subgame equilibria, which are discussed below. To limit the column width, the cells within this table refer to the home lender as "H" and the away lenders as "A".

Table Prop.2-S4: Stage 4 loan rate equilibria within subgames

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* H: $R_{kh}(q) = \begin{cases} y \text{ if } q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right] \\ \frac{c+\tau}{q} \text{ if } q > \frac{c+\tau}{y} \end{cases}$ A: $R_{ka}(q) = \frac{c+\tau}{q}$ if $q > \frac{c+\tau}{y}$	*"Top-right" subgame:* H: $R_h = y$ if $q > \frac{c}{y}$ A: N/A
Home lender has no data access	*"Bottom-left" subgame:* H: N/A A: $R_{ka}(q) = \frac{c+\tau}{q}$ if $q > \frac{c+\tau}{y}$	*"Bottom-right" subgame:* H: N/A A: N/A

Note: "N/A" stands for "not available," which represents the fact that uninformed lenders here do not offer any loan rate to households.

Before turning to the specific subgames, we first note that when $y < \frac{4}{3}c$ (i.e., the low y case of Proposition 2) there exists no breakeven R_u and lenders are always unwilling to lend to unrevealed households. The lowest (home lender) breakeven rate on lending to unrevealed households is $\frac{c}{E[q|u]}$ and $\sup E[q|u] = E[q] = \frac{3}{4}$ so that $\inf \frac{c}{E[q|u]} = \frac{4}{3}c > y$. That is, even if everyone, including the highest quality households, were to use cash, no lending takes place to unrevealed households. Equivalently, when a lender is uninformed (no data access obtained), it cannot profitably lend at any loan rate, because it cannot distinguish between borrowers and therefore all unrevealed households would borrow from it. The best pool of borrowers it could conceivably face has $\sup E[q|u] = E[q] = \frac{3}{4}$ and even that is not good enough to prevent a loss-making loan portfolio.

This simplifies the problem as compared to Proposition 1: the Stage 4 subgame derivations center on finding the subgame equilibria for lending by informed lenders to revealed households only.

Subgame 1: all lenders have data access (top-left subgame). Among revealed households, we can distinguish three potential categories by ranges of creditworthiness. First, if $q < \frac{c}{y}$, then a household's project has negative NPV ($qy < c$) and no lender will lend to it. Second, if $q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right]$, then the household's project only has positive NPV when borrowing on-island and therefore only the home lender will make it a loan rate offer. Third, if $q > \frac{c+\tau}{y}$ then all

lenders can make loan offers (i.e., the expected return on the project exceeds the breakeven rate of away lenders, $\frac{c+\tau}{q}$). For households with $q > \frac{c+\tau}{y}$, then, the same arguments as in the Proof of Proposition 1 apply to this stage and therefore the equilibrium loan rates are given by $R_{kh}(q) = R_{ka}(q) = \frac{c+\tau}{q}$. Instead, for households with $q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right]$, the home lender is a monopolist and therefore sets $R_h = y$ to fully appropriate the return on successful projects.

Subgame 2: the home lender has data access, while the away lenders do not (top-right subgame). Away lenders are uninformed and therefore cannot profitably lend to any household. The home lender is a monopolist and, therefore, towards households to which it can profitably lend at all ($q > \frac{c}{y}$), it sets $R_h = y$ to gain the full return on successful projects.

Subgame 3: the home lender has no data access, while the away lenders do (bottom-left subgame). The home lender is uninformed and therefore cannot profitably lend to any household. For households to which they can profitably lend ($q > \frac{c+\tau}{y}$), Bertrand competition among the away lenders drives loan rates down to margin pricing: $R_{ka}(q) = \frac{c+\tau}{q}$.

Subgame 4: all lenders are uninformed (bottom-right subgame). All lenders are uninformed and therefore no loan rate offers are made.

The home lender's overall profit function for Stage 4. Subgames 1-4 identified the home lender profit function within each subgame. Using the same term definitions as in (A.16), we can write the home lender's Stage 4 profit function over all subgames as:

$$\begin{aligned}\Pi_h^{S4} &= \mathbf{TL}\left(\begin{array}{l}\int_{\frac{c+\tau}{y}}^{1} Z_k\left(qR_{kh}^{TL}\left(q|\, q > \frac{c+\tau}{y}\right) - c\right) f(q)\, dq + \\ \int_{\frac{c}{y}}^{\frac{c+\tau}{y}}\left(qR_{kh}^{TL}\left(q|\, q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right]\right) - c\right) f(q)\, dq\end{array}\right) \\ &\quad + \mathbf{TR}\int_{\frac{c}{y}}^{1}\left(qR_{kh}^{TR}\left(q|\, q > \frac{c}{y}\right) - c\right) f(q)\, dq\end{aligned} \tag{A.50}$$

where $R_{kh}^{TL}\left(q|\, q > \frac{c+\tau}{y}\right)$ and $R_{kh}^{TL}\left(q|\, q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right]\right)$ are the loan rate offers in the top-left case to households with, respectively, $q > \frac{c+\tau}{y}$ and $q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right]$; and $R_{kh}^{TR}\left(q|\, q > \frac{c}{y}\right)$ is the loan rate offer in the top-right case to households with $q > \frac{c}{y}$.

From the subgame derivations we have that the optimal Stage 4 play of the home lender is given by

$$\begin{aligned} R_{kh}^{TL}\left(q|\, q > \frac{c+\tau}{y}\right) &= \frac{c+\tau}{q}, \\ R_{kh}^{TL}\left(q|\, q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right]\right) &= R_{kh}^{TR}\left(q|\, q > \frac{c}{y}\right) = y \end{aligned} \tag{A.51}$$

Stage 3. Table Prop2-S3a summarizes the profits in each subgame and Table Prop2-S3b summarizes the subgame equilibria depending on data access pricing determined by the DC issuer at Stage 1. Below, we derive the entries for each cell in these tables.

Table Prop2-S3a: Stage 3 lender profit within subgames

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* H: $\Pi_h = \begin{pmatrix} \tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \\ \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq \\ -\Omega_h \end{pmatrix}$ A: $\Pi_a = -\Omega_a$	*"Top-right" subgame:* H: $\Pi_h = \int_{\frac{c}{y}}^{1} (qy-c) f(q)\,dq - \Omega_h$ A: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* H: $\Pi_h = 0$ A: $\Pi_a = -\Omega_a$	*"Bottom-right" subgame:* H: $\Pi_h = 0$ A: $\Pi_a = 0$

Table Prop2-S3b: Data fees determine which cell of Table Prop2-S3a is the subgame equilibrium

Data fee on away lenders $\rightarrow$ Data fee on home lenders $\downarrow$	$\Omega_a = 0$	$\Omega_a > 0$
$\Omega_h = 0$	Top-left	Top-right
$\Omega_h \in \left(0, \tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq\right)$	Top-left	Top-right
$\Omega_h \in \left(\tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq, \int_{\frac{c}{y}}^{1} (qy-c) f(q)\,dq\right)$	Bottom-left	Top-right
$\Omega_h > \int_{\frac{c}{y}}^{1} (qy-c) f(q)\,dq$	Bottom-left	Bottom-right

Lender profits in subgame 1: all lenders have data access (top-left subgame). Based on the Stage 4 loan market equilibrium in this subgame (Table Prop2-S4): all revealed households borrow from their home lender (and unrevealed households cannot borrow at all). Therefore, away lenders make no profit on lending and $\Pi_a = -\Omega_a$ (as they did obtain data access); the home lender sets $R_{kh}(q) = y$ for households with $q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right]$ and $R_{kh}(q) = \frac{c+\tau}{q}$ for households with $q > \frac{c+\tau}{y}$, which means that its profits, including data fees are:

$$\Pi_h = \int_{\frac{c+\tau}{y}}^{1} \left(q\frac{c+\tau}{q} - c\right) f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq - \Omega_h \tag{A.52}$$

and simplifying $\left(q\frac{c+\tau}{q} - c\right) = \tau$ gives us the expression for Π_h in the top-left cell of Table Prop2-S3a.

Lender profits in subgame 2: the home lender has data access, while the away lenders do not (top-right subgame). Based on the Stage 4 loan market equilibrium in this subgame (Table Prop2-S4): away lenders do not make loan offers (as they are uninformed), and

therefore $\Pi_a = 0$; while the home lender offers $R_h = y$ to all households with $q > \frac{c}{y}$, so that:

$$\Pi_h = \int_{\frac{c}{y}}^{1} (qy - c) f(q) \, dq - \Omega_h \tag{A.53}$$

Lender profits in subgame 3: the home lender has no data access, while the away lenders do (bottom-left subgame). Based on the Stage 4 loan market equilibrium in this subgame (Table Prop2-S4): informed away lenders compete down to margin pricing on the loan market, leaving them with zero profit from lending and, since they obtained data access, $\Pi_a = -\Omega_a$. The home lender does not make loan offers (as it is uninformed), and therefore $\Pi_h = 0$.

Lender profits in subgame 4: all lenders are uninformed (bottom-right subgame). As uninformed lenders make losses on any loan offers in the setting of Proposition 2, no loan offers are made and $\Pi_h = \Pi_a = 0$.

We next consider which subgame emerges as the Stage 3 equilibrium depending on the values of Ω_h and Ω_a. That is, we consider each cell in Table Prop2-S3b.

Subgame equilibrium when $\Omega_h = \Omega_a = 0$. The subgame equilibrium is top-left in Table Prop2-S3a as there is no data purchase decision to be made and lenders all have data access.

Subgame equilibrium when $\Omega_h = 0$ and $\Omega_a > 0$. Since the home lender has been given free data access, the table below represents the profit structure. Given $\Omega_a > 0$, away lenders choose not to buy data access here and therefore the subgame equilibrium is top-right.

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* H: $\Pi_h = \tau \int_{\frac{c+\tau}{y}}^{1} f(q) \, dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy - c) f(q) \, dq$ A: $\Pi_a = -\Omega_a$	*"Top-right" subgame:* H: $\Pi_h = \int_{\frac{c}{y}}^{1} (qy - c) f(q) \, dq$ A: $\Pi_a = 0$

Subgame equilibrium when $\Omega_h \in \left(0, \tau \int_{\frac{c+\tau}{y}}^{1} f(q) \, dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy - c) f(q) \, dq\right)$ and $\Omega_a = 0$. With free data access for away lenders, the remaining data purchase decision is that of the home lender, depicted in the table below. Since $\tau \int_{\frac{c+\tau}{y}}^{1} f(q) \, dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy - c) f(q) \, dq - \Omega_h > 0$, the home lender chooses to buy data access, and the top-left cell in Table Prop2-S3a is the subgame equilibrium.

	Away lenders have data access
Home lender has data access	*"Top-left" subgame:* H: $\Pi_h = \tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq - \Omega_h > 0$ A: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* H: $\Pi_h = 0$ A: $\Pi_a = -\Omega_a$

Subgame equilibrium when $\Omega_h \in \left(0, \tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq\right)$ *and* $\Omega_a > 0$. Here, Table Prop2-S3a becomes:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* H: $\Pi_h = \aleph_1 > 0$ A: $\Pi_a < 0$	*"Top-right" subgame:* H: $\Pi_h = \aleph_2 > \aleph_1 > 0$ A: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* H: $\Pi_h = 0$ A: $\Pi_a < 0$	*"Bottom-right" subgame:* H: $\Pi_h = 0$ A: $\Pi_a = 0$

where $\aleph_1 = \tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq$ and $\aleph_2 = \int_{\frac{c}{y}}^{1} (qy-c) f(q)\,dq$, and

$$\begin{aligned} \aleph_2 - \aleph_1 &= \int_{\frac{c+\tau}{y}}^{1} (qy-c) f(q)\,dq - \tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq \\ &= \int_{\frac{c+\tau}{y}}^{1} (qy-(c+\tau)) f(q)\,dq > 0 \end{aligned} \tag{A.54}$$

This means that the game play can be represented as below, and therefore the top-right cell is the subgame equilibrium:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: **Stay** Away lenders: *Right*	*"Top-right" subgame:* Home lender: **Stay** Away lenders: **Stay**
Home lender has no data access	*"Bottom-left" subgame:* Home lender: *Up* Away lenders: *Right*	*"Bottom-right" subgame:* Home lender: *Up* Away lenders: **Stay**

Subgame equilibrium when $\Omega_h \in \left(\tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq, \int_{\frac{c}{y}}^{1} (qy-c) f(q)\,dq\right)$ *and* $\Omega_a = 0$. Here, Table Prop2-S3a centers on the left column, since away lenders have been

given data access. Moreover, $\Omega_h \in \left(\tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq, \int_{\frac{c}{y}}^{1} (qy-c) f(q)\,dq\right)$ implies that the top-left entry for Π_h is negative. Therefore, the subgame equilibrium is bottom-left.

	Away lenders have data access
Home lender has data access	*"Top-left" subgame:* H: $\Pi_h < 0$ A: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* H: $\Pi_h = 0$ A: $\Pi_a = 0$

Subgame equilibrium when $\Omega_h \in \left(\tau \int_{\frac{c+\tau}{y}}^{1} f(q)\,dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy-c) f(q)\,dq, \int_{\frac{c}{y}}^{1} (qy-c) f(q)\,dq\right)$ *and* $\Omega_a > 0$. Here, Table Prop2-S3a becomes:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* H: $\Pi_h < 0$ A: $\Pi_a < 0$	*"Top-right" subgame:* H: $\Pi_h > 0$ A: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* H: $\Pi_h = 0$ A: $\Pi_a < 0$	*"Bottom-right" subgame:* H: $\Pi_h = 0$ A: $\Pi_a = 0$

This means that the evolution of game play can be represented as below, and therefore the top-right cell is the subgame equilibrium:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: *Down* Away lenders: *Right*	*"Top-right" subgame:* Home lender: **Stay** Away lenders: **Stay**
Home lender has no data access	*"Bottom-left" subgame:* Home lender: **Stay** Away lenders: *Right*	*"Bottom-right" subgame:* Home lender: *Up* Away lenders: **Stay**

Subgame equilibrium when $\Omega_h > \int_{\frac{c}{y}}^{1} (qy-c) f(q)\,dq$ *and* $\Omega_a = 0$. Here, Table Prop2-S3a centers on the left column, since away lenders have been given data access. The subgame equilibrium is bottom-left:

	Away lenders have data access
Home lender has data access	*"Top-left" subgame:* H: $\Pi_h < 0$ A: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* H: $\Pi_h = 0$ A: $\Pi_a = 0$

Subgame equilibrium when $\Omega_h > \int_{\frac{c}{y}}^{1} (qy - c) f(q) dq$ *and* $\Omega_a > 0$. Here, Table Prop2-S3a becomes:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* H: $\Pi_h < 0$ A: $\Pi_a < 0$	*"Top-right" subgame:* H: $\Pi_h < 0$ A: $\Pi_a = 0$
Home lender has no data access	*"Bottom-left" subgame:* H: $\Pi_h = 0$ A: $\Pi_a < 0$	*"Bottom-right" subgame:* H: $\Pi_h = 0$ A: $\Pi_a = 0$

This means that the evolution of game play can be represented as below, and therefore the bottom-right cell is the subgame equilibrium:

	Away lenders have data access	Away lenders have no data access
Home lender has data access	*"Top-left" subgame:* Home lender: *Down* Away lenders: *Right*	*"Top-right" subgame:* Home lender: *Down* Away lenders: **Stay**
Home lender has no data access	*"Bottom-left" subgame:* Home lender: **Stay** Away lenders: *Right*	*"Bottom-right" subgame:* Home lender: **Stay** Away lenders: **Stay**

This completes the derivations behind Tables Prop2-S3a and Prop2-S3b. Based on these tables, in turn, we can also write down an expression for the optimization problem of the home lender at Stage 3 (while, due to Bertrand competition among them, the optimal strategy of the away lenders is straightforwardly given by "don't buy data access" whenever $\Omega_a > 0$, as can also be seen from Tables Prop2-S3a and Prop2-S3b). Following the same notation for indicator variables as previously employed in (A.18), we can write the Stage 3 optimization problem for the home lender as:

$$\max_{\mathbf{B}=\{0,1\}} \Pi_h^{S3} = \mathbf{AL}\left(\left(\tau \int_{\frac{c+\tau}{y}}^{1} f(q) dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy - c) f(q) dq - \Omega_h\right)\mathbf{B}\right) \quad \text{(A.55)}$$

$$+ (1 - \mathbf{AL})\left(\left(\int_{\frac{c}{y}}^{1} (qy - c) f(q) dq - \Omega_h\right)\mathbf{B}\right) \quad \text{(A.56)}$$

from which the home lender optimal strategy on data purchases (for any $\Omega_h > 0$) directly follows as

$$\text{H: optimal strategy} \begin{cases} \text{if } \mathbf{AL} = 1, \text{ set } \mathbf{B} = 1 \Leftrightarrow \Omega_h < \tau \int_{\frac{c+\tau}{y}}^{1} f(q)\, dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy - c) f(q)\, dq \\ \text{if } \mathbf{AL} = 0, \text{ set } \mathbf{B} = 1 \Leftrightarrow \Omega_h < \int_{\frac{c}{y}}^{1} (qy - c) f(q)\, dq \end{cases} \tag{A.57}$$

Together with

$$\text{A: optimal strategy} \begin{cases} \Omega_a = 0 \Rightarrow \mathbf{AL} = 1 \\ \text{if } \Omega_a > 0 \text{ then set } \mathbf{AL} = 0 \end{cases} \tag{A.58}$$

this provides a mathematical formulation that replicates Table S3b.

Stage 2. In the subgames in the right column of Table Prop2-S4, revealed households either receive no loan offer (all revealed households in the bottom-right subgame and revealed households with $q < \frac{c}{y}$ in the top-right subgame) or receive a loan offer that equals the full return on a successful project ($R_h = y$). In either case, there is no benefit to becoming a revealed household, and therefore to choosing DC at Stage 2. All households therefore choose cash at Stage 2, because using cash comes without a privacy cost, while using DC cannot lead to a loan offer that leaves any project return in household hands. The right column of Table Prop2-S2 below therefore has $m = 0$.

Table Prop2-S2: Aggregate household sorting in relation to foreseen lender subgame equilibria

	Away lenders have data access	Away lenders: no data
Home lender has data access	*"Top-left" subgame:* $m = \begin{cases} 0 \text{ if } y < c + \tau \\ \frac{\theta}{2\alpha y} (y - c - \tau)^2 \text{ if } c + \tau < y < 2\alpha + c + \tau \\ 2\theta \left(1 - \frac{\alpha + c + \tau}{y}\right) \text{ if } y > 2\alpha + c + \tau \end{cases}$	*"Top-right" subgame:* $m = 0$
Home lender has no data access	*"Bottom-left" subgame:* $m = \begin{cases} 0 \text{ if } y < c + \tau \\ \frac{\theta}{2\alpha y} (y - c - \tau)^2 \text{ if } c + \tau < y < 2\alpha + c + \tau \\ 2\theta \left(1 - \frac{\alpha + c + \tau}{y}\right) \text{ if } y > 2\alpha + c + \tau \end{cases}$	*"Bottom-right" subgame:* $m = 0$

We next turn to deriving the solution for m in the left-column. We first note that the top-left and bottom-left subgames at Stage 4 in Table Prop2-S4 imply equivalent sorting (over cash and DC) of households at Stage 2. In both the top-left and bottom-left subgames revealed households with $q > \frac{c+\tau}{y}$ receive loan offers of $R_k(q) = \frac{c+\tau}{q}$, while revealed households with $q < \frac{c}{y}$ receive no loan offers. Households with $q \in \left[\frac{c}{y}, \frac{c+\tau}{y}\right]$ receive an offer of $R_h = y$ in the top-left subgame and no loan offer in the bottom-left subgame. However, a loan with $R_h = y$ is equivalent to not getting a loan at all, in terms of household utility: if the project fails, the household earns 0; if the project succeeds, the household is also left with

0 after repaying the loan. Hence, in both the top-left and bottom-left subgames, households with $q < \frac{c+\tau}{y}$ choose cash at Stage 2, because there is no benefit to choosing DC while there is a privacy cost associated with it. The remainder of the derivation of this stage therefore centers on the means of payment choice of $q > \frac{c+\tau}{y}$ households and the derived expression for m applies to both the top-left and bottom-left subgames.

The first case to consider is that $y < c + \tau$. This is possible because $\inf y = c$ from (4). When $y < c+\tau$, there are no households with $q > \frac{c+\tau}{y}$. It then immediately follows from the above paragraph that all households choose cash and therefore $m = 0$. In other words, when $y < c+\tau$, all subgames give $m = 0$ and no credit provision takes place in the model. This is the first sentence in the statement of Proposition 2 (there written equivalently as $\tau > y - c$).

When $y > c + \tau$, there exist households within $q \in \left[\frac{1}{2}, 1\right]$ for which $q > \frac{c+\tau}{y}$. From (1) and Table Prop2-S4, these households prefer DC over cash if and only if

$$\theta q \left(y - \frac{c+\tau}{q} \right) - \alpha\varphi\theta > 0 \tag{A.59}$$

because if they use DC then they become revealed with probability θ and when revealed they obtain a loan rate of $\frac{c+\tau}{q}$, so that their expected project return is $q\left(y - \frac{c+\tau}{q}\right)$. Moreover, DC use comes with privacy cost $\alpha\varphi\theta$. The choice for cash instead leads to a certain 0 payoff, as unrevealed households cannot obtain a loan. Rewriting this condition, a household chooses DC over cash when

$$\alpha\varphi < qy - (c + \tau) \tag{A.60}$$

where we note that (A.60) can alternatively be written as

$$q > \frac{\alpha\varphi + c + \tau}{y} \tag{A.61}$$

To come to an expression for m, an integral over the distribution of revealed households is needed. The bounds of the integrals depend on whether the $q = 1$ and $\varphi = 2$ household chooses to use the DC or not. From (A.60), we have that this household chooses DC if $y > 2\alpha + c + \tau$ and chooses cash if $y < 2\alpha + c + \tau$. We note from (5) and (6) that $\tau \in \left(0, \frac{1}{6}c\right)$ and $\alpha \in \left(\frac{1}{9}(c+\tau), \frac{1}{6}(c+\tau)\right)$, which imply that $2\alpha + c + \tau \in \left(\frac{11}{9}c, \frac{14}{9}c\right)$. Given $y \in \left(c, \frac{4}{3}c\right)$ in the case of Proposition 2, this means that both $y > 2\alpha + c + \tau$ and $y < 2\alpha + c + \tau$ are possible and should be investigated.

If $y < 2\alpha + c + \tau$, then even for $q = 1$ there are still households who prefer cash over DC. In this case, m is given by:

$$m = \theta \int_{\frac{c+\tau}{y}}^{1} \left(\frac{qy - (c+\tau)}{\alpha} \right) dq = \frac{\theta}{2\alpha y} (y - c - \tau)^2 \tag{A.62}$$

where $\left(\frac{qy-(c+\tau)}{\alpha}\right)$ (i.e., $f(q)$) comes setting (A.60) to equality and solving to φ.

When $y > 2\alpha + c + \tau$, we have to consider above which q all households choose DC, even when they have $\varphi = 2$. By (A.60), this occurs when $q > \frac{2\alpha+c+\tau}{y}$. Hence, the mass of revealed

borrowers is now found as

$$m = \theta \left(\int_{\frac{c+\tau}{y}}^{\frac{2\alpha+c+\tau}{y}} \left(\frac{qy - (c+\tau)}{\alpha} \right) dq + \int_{\frac{2\alpha+c+\tau}{y}}^{1} 2dq \right) = 2\theta \left(1 - \frac{\alpha + c + \tau}{y} \right) \tag{A.63}$$

Overall, therefore:

$$m = \begin{cases} 0 \text{ if } y < c + \tau \\ \frac{\theta}{2\alpha y} (y - c - \tau)^2 \text{ if } c + \tau < y < 2\alpha + c + \tau \\ 2\theta \left(1 - \frac{\alpha + c + \tau}{y} \right) \text{ if } y > 2\alpha + c + \tau \end{cases} \tag{A.64}$$

Stage 1. When $y < c + \tau$, there is nothing for the DC issuer to solve, because all households always choose cash at Stage 2. We therefore focus the remainder of the derivation of this stage on the case where $y > c + \tau$ and for which

$$m = \begin{cases} \frac{\theta}{2\alpha y} (y - c - \tau)^2 \text{ if } c + \tau < y < 2\alpha + c + \tau \\ 2\theta \left(1 - \frac{\alpha + c + \tau}{y} \right) \text{ if } y > 2\alpha + c + \tau \end{cases} \tag{A.65}$$

The DC issuer backward induces all subgames and foresees the following relation between data fees and profits:

Table Prop2-S1: Profits from data sales

Data fee on away lenders → Data fee on home lenders ↓	$\Omega_a = 0$	$\Omega_a > 0$
$\Omega_h = 0$	$\pi = 0$	$\pi = 0$
$\Omega_h \in (0, m\tau)$	$\pi = \Omega_h \Rightarrow \widehat{\pi} = m\tau$	$\pi = 0$
$\Omega_h > m\tau$	$\pi = 0$	$\pi = 0$

Below we discuss the entries in each cell of Table Prop2-S1:

If the DC issuer sets $\Omega_h = \Omega_a = 0$: data revenues are necessarily 0 because lenders get free data access.

If the DC issuer sets $\Omega_h = 0$ *and* $\Omega_a > 0$: the Stage 3 subgame equilibrium is top-right (Table Prop2-S3b), which means that the away lenders don't buy data access. Since the home lender is given the data for free, this implies $\pi = 0$.

If the DC issuer sets $\Omega_h \in (0, m\tau)$ *and* $\Omega_a = 0$: the Stage 3 subgame equilibrium is top-left (Table Prop2-S3b). Regarding the entries in Table Prop2-S3b, as seen by the DC issuer at Stage 1, we note that $\int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy - c) f(q) dq = 0$ because, based on the Stage 2 derivation, $\int_{\frac{c}{y}}^{\frac{c+\tau}{y}} f(q) dq = 0$ as all households below $q = \frac{c+\tau}{y}$ choose cash. Hence, $\tau \int_{\frac{c+\tau}{y}}^{1} f(q) dq = m\tau$

and therefore the row entry $\Omega_h \in \left(0, \tau \int_{\frac{c+\tau}{y}}^{1} f(q)\, dq + \int_{\frac{c}{y}}^{\frac{c+\tau}{y}} (qy - c) f(q)\, dq\right)$ equals $\Omega_h \in (0, m\tau)$. The associated subgame equilibrium is top-left, as shown in Table Prop2-S3b. This means that the home lender purchases the data at fee $\Omega_h \in (0, m\tau)$. The optimal data fee is now obviously the highest data fee within the range $\Omega_h \in (0, m\tau)$, which is $m\tau - \varepsilon$ with $\varepsilon \to 0$, which we can express as the limit case where Ω_h is set to $m\tau$ as the optimal data fee on the home lender, from which also follows $\widehat{\pi} = \widehat{\Omega}_h = m\tau$.

If the DC issuer sets $\Omega_h \in (0, m\tau)$ *and* $\Omega_a > 0$: the Stage 3 subgame equilibrium is top-right (Table Prop2-S3b). But the top-right subgame is associated with $m = 0$ at Stage 2 (Table Prop2-S2). Since no one uses the DC, there is no data to sell and $\pi = 0$.

If the DC issuer sets $\Omega_h > m\tau$ *and* $\Omega_a = 0$: the Stage 3 subgame equilibrium is bottom-left (Table Prop2-S3b). The home lender finds data access too expensive, because, purchasing it would lower its profit (Table Prop2-S3a). Since it does not buy data access and the away lenders are given the data for free, the DC issuer has no data revenues and $\pi = 0$.

If the DC issuer sets $\Omega_h > m\tau$ *and* $\Omega_a > 0$: the Stage 3 subgame equilibrium is bottom-right (Table Prop2-S3b). Neither home nor away lenders purchase the data and therefore $\pi = 0$.

Given these derivations, summarized in Table Prop2-S1, it is immediately apparent that the optimal data fee structure is $\Omega_h = \widehat{\Omega}_h$ and $\Omega_a = 0$. Only the structure with $\Omega_h \in (0, m\tau)$ and $\Omega_a = 0$ leads to positive profits for the DC issuer, and within that structure profits are maximized at $\Omega_h = \widehat{\Omega}_h$. The closed-form expression for the optimal data fee is:

$$\widehat{\Omega}_h = m\tau = \begin{cases} \frac{\theta\tau}{2\alpha y}(y - c - \tau)^2 & \text{if } y < 2\alpha + c + \tau \\ 2\theta\tau\left(1 - \frac{\alpha + c + \tau}{y}\right) & \text{if } y > 2\alpha + c + \tau \end{cases} \tag{A.66}$$

In (A.66), $\widehat{\Omega}$ increases linearly as θ rises, regardless of $y \lessgtr 2\alpha + c + \tau$:

$$\frac{\partial \widehat{\Omega}_h}{\partial \theta} = \begin{cases} \frac{\tau}{2\alpha y}(y - c - \tau)^2 > 0 \text{ if } y < 2\alpha + c + \tau \\ 2\tau\left(1 - \frac{\alpha + c + \tau}{y}\right) > 0 \text{ if } y > 2\alpha + c + \tau \end{cases} \tag{A.67}$$

and therefore $\theta = 1$ is optimal for the DC issuer.

The DC issuer's Stage 1 optimization problem can be expressed in algebraic form: this takes the same form as (A.39)-(A.42) and is therefore not repeated here.

Subgame Perfect Nash Equilibrium: Summary

In sum, through backward induction, we find the unique Subgame Perfect Nash Equilibrium, which has the following equilibrium strategies:

At Stage 1, the DC issuer, foreseeing the entire game, optimally sets $\theta = 1$, $\Omega_a = 0$, and $\Omega_h = \widehat{\Omega}_h$, with the solution for $\widehat{\Omega}_h$ being given by (A.66). Moreover, given $\theta = 1$ the expression for $\widehat{\Omega}_h$ from (A.37) becomes

$$\widehat{\Omega}_h\Big|_{\theta=1} = \begin{cases} \frac{\tau}{2\alpha y}\left(y - c - \tau\right)^2 > 0 \text{ if } y < 2\alpha + c + \tau \\ 2\tau\left(1 - \frac{\alpha + c + \tau}{y}\right) > 0 \text{ if } y > 2\alpha + c + \tau \end{cases} \tag{A.68}$$

At Stage 2, households' sorting decisions imply that the mass of revealed households is given by (A.65), which is also the expression shown in (13) in the statement of Proposition 2 in Section 4.2. More specifically, this is m for the case where $y > c + \tau$. For $y < c + \tau$, the outcome is $m = 0$ (equation (A.64)) as all households always choose cash and no credit provision takes place. This is the first sentence in the statement in Proposition 2. The remainder of this summary centers on the case where $y > c + \tau$. Households opt for DC use when their q and φ are such that they satisfy (A.61), which is also the expression shown in (12) in the statement of Proposition 2.

At Stage 3, the home lender buys data access at the fee shown in (A.68), while the away lenders have obtained free data access. At Stage 4, this leads to a loan market equilibrium in which all lenders offer revealed households the type-dependent loan rates:

$$R_k\left(q\right) = \frac{c + \tau}{q} \tag{A.69}$$

where we note the following: Table Prop2-S4 lays out that only households with $q > \frac{c+\tau}{y}$ are offered $R_k\left(q\right) = \frac{c+\tau}{q}$. However, since only households satisfying (A.60) have sorted into DC use - i.e., households with $q > \frac{\alpha\varphi + c + \tau}{y}$ - then all DC users necessarily satisfy $q > \frac{c+\tau}{y}$. Hence, any DC user that becomes revealed (and given $\theta = 1$ all DC users become revealed), will receive the offer $R_k\left(q\right) = \frac{c+\tau}{q}$ at Stage 4. Lastly, at Stage 5 all revealed households borrow from their home lender at these rates.

A.3 Proof of Proposition 3

Stage 5 is unchanged compared to the baseline: households borrow from the source (lender or DC issuer) that offers them the lowest loan rate. For Stages 1-4, we differentiate between the "High y" and "Low y" cases of Table 1.

In the "High y" case, lenders or the DC issuer offer a single loan rate to unrevealed borrowers at Stage 4. There are two on-island lenders for every island: the home lender and the DC issuer. Given Bertrand competition, the equilibrium loan rate for unrevealed borrowers is $R_u = \frac{c}{E[q|u]}$. This is the breakeven rate: for the home lender or the DC issuer, the expected profit per unrevealed borrower is 0.

Because the DC issuer hoards the payment data in this setup, all households are unrevealed to the lenders (i.e., all lenders are uninformed). At Stage 4, the DC issuer therefore optimally offers revealed borrowers the highest loan rate at which they are captive. It follows directly that this is $\frac{c}{E[q|u]} - \varepsilon$ with $\varepsilon \to 0$, because no revealed borrower can do better by switching to another lender. That is, the equilibrium loan rate for a revealed borrower is independent of its credit quality and is identical to that for the unrevealed.

Stage 3 is inactive in Section 5 (i.e., the DC issuer does not sell data).

At Stage 2, all households optimally choose cash. It can be seen from (A.24) that when $R_{dc} = R_{cash}$, a household prefers DC over cash if and only if $-\alpha\varphi\theta > 0$, which is never satisfied. Foreseeing that the Stage 4 loan rate is independent of q, households see no benefit to disclosure, while they do experience a privacy cost from using DC. Hence, DC is unused. Without DC use, DC design at Stage 1 is irrelevant. The DC issuer therefore makes no profit in the "High y" case, because no one uses DC, while loan rates on unrevealed households equal the DC issuer's breakeven rate.

In the "Low y" case, unrevealed borrowers cannot obtain credit, in line with Proposition 2. At Stage 4, any revealed borrower is only revealed to the DC issuer. Therefore, the DC issuer is a credit monopolist. Thus, the DC issuer will charge any revealed household a loan rate equal to y. This leaves no project value added to households. Foreseeing this at Stage 2, households therefore all choose cash. In the "Low y" case, no credit provision takes place at all in equilibrium, which implies that the DC issuer makes no profit.

A.4 Proof of Proposition 4

First, we consider W_h and then W_l.

Derivation of W_h:

The first term in (14): Since $y_h > 2(c+\tau)$, the first term in (14) becomes $\int_{\frac{1}{2}}^{1} 2(qy_h - c)\,dq = \frac{3}{4}y_h - c$, because all households are borrowers from their home lenders in the equilibrium of Proposition 1. That is, $f(q|\text{ "borrowing"}) = 2$ in (14). We also note that $\frac{3}{4}y_h - c$ is strictly positive given $y_h > 2(c+\tau)$.

The second term in (14): The second term in (14) follows from rewriting the indifference frontier (wherein (9) holds with equality) to integrate over $\varphi \in [0,2]$, which is $1 - \left(1 + \frac{\alpha\varphi}{c+\tau}\right) E[q|u]$.[3] Hence, the second term in (14) becomes $\int_0^2 (\alpha\varphi\theta)\left(1 - \left(1 + \frac{\alpha\varphi}{c+\tau}\right) E[q|u]\right) d\varphi$ which equals $2\alpha\theta\left[1 - \left(1 + \frac{4\alpha}{3(c+\tau)}\right) E[q|u]\right]$, and which we note is always a positive term given $E[q|u] \leq \frac{3}{4}$ and $\alpha < \frac{1}{6}(c+\tau)$, which jointly imply $\inf\left[1 - \left(1 + \frac{4\alpha}{3(c+\tau)}\right) E[q|u]\right] = \frac{1}{12} > 0$.

Putting terms together yields:

$$W_h = \frac{3}{4}y_h - c - 2\alpha\theta\left[1 - \left(1 + \frac{4\alpha}{3(c+\tau)}\right) E[q|u]\right] \tag{A.70}$$

Derivation of W_l:

The first term in (14): For W_l, the first term in (14) becomes an integral of $qy_l - c$ over revealed households (because these are the only ones to borrow in Proposition 2). The integrals for revealed households (i.e., θ times DC users) are shown in (A.62) and (A.63), and therefore the first term in (14) is $\theta \int_{\frac{c+\tau}{y_l}}^{1} (qy_l - c)\left(\frac{qy_l - (c+\tau)}{\alpha}\right) dq$ if $y_l < 2\alpha + c + \tau$; and $\theta\left(\int_{\frac{c+\tau}{y_l}}^{\frac{2\alpha+c+\tau}{y_l}} (qy_l - c)\left(\frac{qy_l-(c+\tau)}{\alpha}\right) dq + \int_{\frac{2\alpha+c+\tau}{y_l}}^{1} 2(qy_l - c)\,dq\right)$ if $y_l > 2\alpha + c + \tau$, which become:

$$W_l \text{ (first term)} = \begin{cases} \theta\frac{(y_l - c - \tau)^2(2y - 2c + \tau)}{6\alpha y_l} & \text{if } y_l < 2\alpha + c + \tau \\ \theta\frac{\left((y_l - c)^2 - \frac{4}{3}\alpha^2 - 2\alpha\tau - \tau^2\right)}{y_l} & \text{if } y_l > 2\alpha + c + \tau \end{cases} \tag{A.71}$$

The second term in (14): For W_l, for the second term in (14), we have that (A.62) and (A.63) represent the integrals for DC users as defined for integration over q, which we rewrite

[3] Rewriting (9) with equality to q gives the indifference frontier as $q = \left(1 + \frac{\alpha\varphi}{c+\tau}\right) E[q|u]$. When visualizing the indifference frontier on a plane where φ is on the horizontal and q is on the vertical axis (inverting Figure 2), the expression $1 - \left(1 + \frac{\alpha\varphi}{c+\tau}\right) E[q|u]$ then captures the mass of DC users (which lie above the indifference frontier in this visualization).

for integration over φ in the same manner as for W_h, but with the density function taken from setting (A.60) to equality. Note that for the $y_l < 2\alpha + c + \tau$ case, integration now only goes up to $\varphi = \frac{y_l-(c+\tau)}{\alpha}$, because beyond that point there are only cash users (i.e., from solving $1 = \frac{\alpha\varphi+(c+\tau)}{y_l}$ to φ). This gives $\int_0^{\frac{y_l-(c+\tau)}{\alpha}} (\alpha\varphi\theta)\left(1 - \frac{\alpha\varphi+(c+\tau)}{y_l}\right) d\varphi$ if $y_l < 2\alpha + c + \tau$; and $\int_0^2 (\alpha\varphi\theta)\left(1 - \frac{\alpha\varphi+(c+\tau)}{y_l}\right) d\varphi$ if $y_l > 2\alpha + c + \tau$, which become:

$$W_l \text{ (second term)} = \begin{cases} \theta\frac{1}{6\alpha y_l}\left(y_l - (c+\tau)\right)^3 & \text{if } y_l < 2\alpha + c + \tau \\ \theta\frac{2\alpha}{y_l}\left(y_l - (c+\tau) - \frac{4}{3}\alpha\right) & \text{if } y_l > 2\alpha + c + \tau \end{cases} \tag{A.72}$$

Putting terms together and simplifying, we can write

$$W_l = \begin{cases} \theta\frac{(y_l-(c+\tau))^2(y_l-c+2\tau)}{6\alpha y_l} & \text{if } y_l < 2\alpha + c + \tau \\ \theta\frac{\left((y_l-c)(y_l-c-2\alpha)+\frac{4}{3}\alpha^2-\tau^2\right)}{y_l} & \text{if } y_l > 2\alpha + c + \tau \end{cases} \tag{A.73}$$

where we note that the numerator in both expressions is always positive. This can be directly seen from $y_l > c$ (i.e., from (4)) for $\left(y_l - (c+\tau)\right)^2 (y_l - c + 2\tau)$. For $\left((y_l - c)(y_l - c - 2\alpha) + \frac{4}{3}\alpha^2 - \tau^2\right)$, we have that its infimum occurs at $y_l \to 2\alpha+c+\tau$ (the lower bound of y_l given $y_l > 2\alpha+c+\tau$), and this infimum is $2\alpha(2\alpha + 3\tau) > 0$.

A.5 Proof of Proposition 5

First, W_h in (16) is necessarily maximized at $\theta = 0$, as this sets the negative term representing aggregate privacy costs $\left(-2\alpha\theta\left[1-\left(1+\frac{4\alpha}{3(c+\tau)}\right)E\left[q|\,u\right]\right]\right)$ to 0. Hence, $\gamma = 1 \Rightarrow \overline{\theta} = 0$. Second, it follows directly from (17) that W_l is maximized at $\theta = 1$ and therefore $\gamma = 0 \Rightarrow \overline{\theta} = 1$. The interior solution, $\overline{\theta} \in (0,1)$, is found from $\frac{\partial E[W]}{\partial \theta} = 0$ in (15), which is solved to the expression shown in (18), based on the calculation below:

Let's define λ as the collection of constants that multiplies θ in Wl, as shown in Prop. 4. Then total expected welfare is given by

$$EW = \gamma\left(\frac{3y}{4} - c - \frac{2}{3}\alpha\theta\left(3 + \frac{(2\theta-1) + \sqrt{1-\theta-\frac{4}{3}\frac{\alpha^2}{(c+\tau)^2}(3-4\theta)\theta}}{2\theta\left(1-\frac{4}{3}\frac{\alpha^2}{(c+\tau)^2}\right)}\left(-3-\frac{4\alpha}{c+\tau}\right)\right)\right) + (1-\gamma)\theta*\lambda$$

$$(1-\gamma)\theta\lambda + \gamma\left(-c + \frac{3y}{4} - \frac{2}{3}\alpha\theta\left(3 + \frac{\left(-3-\frac{4\alpha}{c+\tau}\right)\left(-1+2\theta+\sqrt{1-\theta-\frac{4\alpha^2(3-4\theta)\theta}{3(c+\tau)^2}}\right)}{2\theta\left(1-\frac{4\alpha^2}{3(c+\tau)^2}\right)}\right)\right)$$

Then the interior solution for socially optimal θ can be found from

```
FullSimplify[Solve[D[EW, θ] == 0, θ]]
```

$$\left\{\left\{\theta \to \frac{1}{32\alpha^2}3\left(4\alpha^2 + (c+\tau)^2 - \left(\sqrt{\left((-12\alpha^2+(c+\tau)^2\right)\left(8c\alpha^2\gamma + 8\alpha^3\gamma - 3c^2(-1+\gamma)\lambda - 6c(-1+\gamma)\lambda\tau - 3(-1+\gamma)\lambda\tau^2 + 4\alpha^2((-1+\gamma)\lambda + 2\gamma\tau)\right)^2\left(16\alpha^4\gamma^2 - 48\alpha^3(-1+\gamma)\gamma\lambda + 9(-1+\gamma)^2\lambda^2(c+\tau)^2 - 12\alpha^2(-1+\gamma)\lambda(4c\gamma + (-1+\gamma)\lambda + 4\gamma\tau)\right)}\right)\Big/\left(16\alpha^4\gamma^2 - 48\alpha^3(-1+\gamma)\gamma\lambda + 9(-1+\gamma)^2\lambda^2(c+\tau)^2 - 12\alpha^2(-1+\gamma)\lambda(4c\gamma + (-1+\gamma)\lambda + 4\gamma\tau)\right)\right)\right\},$$

$$\left\{\theta \to \frac{1}{32\alpha^2}3\left(4\alpha^2 + (c+\tau)^2 + \left(\sqrt{\left((-12\alpha^2+(c+\tau)^2\right)\left(-8c\alpha^2\gamma - 8\alpha^3\gamma + 3c^2(-1+\gamma)\lambda + 6c(-1+\gamma)\lambda\tau + 3(-1+\gamma)\lambda\tau^2 + 4\alpha^2(\lambda - \gamma\lambda - 2\gamma\tau)\right)^2\left(16\alpha^4\gamma^2 - 48\alpha^3(-1+\gamma)\gamma\lambda + 9(-1+\gamma)^2\lambda^2(c+\tau)^2 - 12\alpha^2(-1+\gamma)\lambda(4c\gamma + (-1+\gamma)\lambda + 4\gamma\tau)\right)}\right)\Big/\left(16\alpha^4\gamma^2 - 48\alpha^3(-1+\gamma)\gamma\lambda + 9(-1+\gamma)^2\lambda^2(c+\tau)^2 - 12\alpha^2(-1+\gamma)\lambda(4c\gamma + (-1+\gamma)\lambda + 4\gamma\tau)\right)\right)\right\}\right\}$$

where we can find that

```
FullSimplify[1/(32 α^2) 3 (4 α^2 + (c + τ)^2)]
```

$$\frac{3}{32}\left(4 + \frac{(c+\tau)^2}{\alpha^2}\right)$$

has an infimum of

```
FullSimplify[3/32 (4 + (c + τ)^2/α^2) /. α → (c + τ)/6]
```

$$\frac{15}{4}$$

which is greater than 1. This means that the positive root solution is necessarily out of bounds (socially optimal θ>1 cannot be an interior solution). Hence:

$$\overline{\theta} = \frac{1}{32\alpha^2}3\left(4\alpha^2 + (c+\tau)^2 - \left(\sqrt{\left((-12\alpha^2+(c+\tau)^2\right)\left(8c\alpha^2\gamma + 8\alpha^3\gamma - 3c^2(-1+\gamma)\lambda - 6c(-1+\gamma)\lambda\tau - 3(-1+\gamma)\lambda\tau^2 + 4\alpha^2((-1+\gamma)\lambda + 2\gamma\tau)\right)^2\left(16\alpha^4\gamma^2 - 48\alpha^3(-1+\gamma)\gamma\lambda + 9(-1+\gamma)^2\lambda^2(c+\tau)^2 - 12\alpha^2(-1+\gamma)\lambda(4c\gamma + (-1+\gamma)\lambda + 4\gamma\tau)\right)}\right)\Big/\left(16\alpha^4\gamma^2 - 48\alpha^3(-1+\gamma)\gamma\lambda + 9(-1+\gamma)^2\lambda^2(c+\tau)^2 - 12\alpha^2(-1+\gamma)\lambda(4c\gamma + (-1+\gamma)\lambda + 4\gamma\tau)\right)\right)$$

Mathematica derivation of interior $\overline{\theta}$.

B Extensions

B.1 Expressions for the social planner with two instruments

This appendix records the main expressions for the two instrument case discussed in Section 7.1. We here show only the expressions themselves and not the underlying derivations, which are available in the *Mathematica* file included in this paper's supplemental files. Households' payoff function can be written as

$$u(q,\varphi) = q\max\{y-R,0\} - \alpha\varphi\Theta + JT + \text{transfers} \tag{B.1}$$

where J is an indicator variable that takes value 1 if the household chooses to use the DC and 0 if it chooses cash. $T>0$ is the tax ($T<0$ is the subsidy), the revenues from which are remitted back (costs of which are charged) to households lump-sum through the "transfers" term. E.g., a cash using household benefits from the presence of a tax, because DC users pay it, while the revenues are transferred back to all households. However, as transfers are independent of individual actions (given atomistic households), they do not affect household optimization, whereas the term JT does. The equilibrium sorting conditions for choosing DC over cash in Propositions 1 and 2, now become, respectively,

$$\varphi \leq \frac{1}{\alpha}\left((c+\tau)\left(\frac{q}{E[q|u]}-1\right)-\frac{T}{\theta}\right) \tag{B.2}$$

and

$$q \geq \frac{\alpha\varphi + c + \tau + T/\theta}{y_l} \tag{B.3}$$

Furthermore, the expression for $E[q|u]$ is

$$E[q|u] = \frac{3(c+\tau)^2(2\theta-1) + \sqrt{3}\sqrt{\begin{array}{c}(c+\tau)^2(3T(T+2\alpha\theta)(4\theta-3)) \\ +\theta\left(3(c+\tau)^2(1-\theta)+4\alpha^2\theta(4\theta-3)\right)\end{array}}}{2\theta^2\left(3(c+\tau)^2-4\alpha^2\right)-6T(T+2\alpha\theta)} \tag{B.4}$$

while expected welfare is given by $E[W] = \gamma W_h + (1-\gamma)W_l$ where

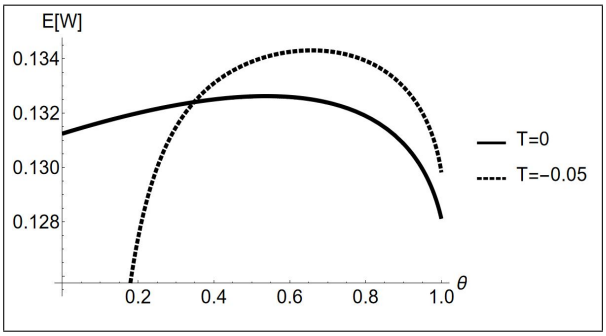
$$W_h = \frac{3}{4}y_h - c - 2\alpha\left(\theta(1-E[q|u]) - E[q|u]\frac{4\alpha\theta+3T}{3(c+\tau)}\right) \tag{B.5}$$

and for W_l the expression below is derived with the added condition that $y_l < 2\alpha+c+\tau+T$, which is sufficient for the integrals to be well-defined:

$$W_l = \frac{(T-\theta(y_l-\tau-c))^2(2T+\theta(y_l+2\tau-c))}{6\alpha\theta^2 y_l} \tag{B.6}$$

A numerical example of the use of two instruments is provided in the figure below. The parameterization (which satisfies the various constraints) is $c=1$, $\tau=0.1$, $\alpha=\frac{c+\tau}{9}=\frac{1+0.1}{9}$, $y_h=2.25$, $y_l=1.25$, and $\gamma=0.15$. In this figure, the solid line plots $E[W]$ relative to θ in the baseline, where $T=0$. Instead, the dashed line plots $E[W]$ relative to θ when

$T = -0.05$. In the baseline setup, the policy maker would ban the DC, setting $\overline{\theta} = 0$. With a second instrument, a higher expected welfare can be attained by loosening regulation in combination with a subsidy on DC use. The reason that the two instruments work as complements here is that the subsidy works against household underdisclosure in the bad state, which emanates from lender market power (with a high τ, loan rates are high and some positive NPV households will choose to use cash despite the fact that they then forego on receiving a loan in the bad state). As more households use DC in the bad state, the social value rises of allowing disclosure to occur (and therefore credit to flow) in that state, and therefore the subsidy goes hand-in-hand with looser regulation.



Numerical example with two instruments.

B.2 Alternative sources of credit quality data

In the baseline model, the DC is the unique source of credit quality information. We here consider two alternative sources: pre-existing public information and public information that is gathered through an alternative payment system technology.[4]

The first possibility is that there is pre-existing public information about credit quality, such as through credit registries or FICO scores. We assume that this information is freely available to lenders and we can incorporate this into our framework through a revelation draw that occurs before the rest of the game outlined in Section 3.5. I.e., before the game begins, some households become revealed and their credit quality is public information. Per our derivations in Appendix A, it follows directly that such households (provided these have positive NPV projects) always receive loan offers $R_k(q)$ as in (7). Nothing else changes in the game. Because this additional revelation occurs before household decisions, it has no impact on household payment choice: for the set of households that is not "pre-revealed", the model is identical to the baseline. Lender and DC issuer total profits are affected, but only because the pool of households that is not pre-revealed is smaller. With a renormalization, total profits remain the same as in the baseline too: if the probability of pre-revelation is $\xi \in (0,1)$, then the renormalization sets the total mass of households per home lender at $\frac{1}{\xi}$.

The second possibility is that, instead of cash, households have access to another alternative payment technology that creates public information about credit quality. Household payment choice then centers on DC in comparison to this alternative means of payment. This could, for example, be a data-sharing CBDC. Let the probability of household credit quality revelation be θ_{alt} when using this alternative, as compared to θ_{dc} when using the DC. We focus on the case where $\theta_{alt} < 1$ and $\theta_{dc} > \theta_{alt}$: the DC is more informationally intrusive than the alternative.

Here, a household chooses DC over the alternative if

$$q\left(y - \left(\theta_{dc} R_k(q) + (1-\theta_{dc}) R_u\right)\right) - \alpha\varphi\theta_{dc} > q\left(y - \left(\theta_{alt} R_k(q) + (1-\theta_{alt}) R_u\right)\right) - \alpha\varphi\theta_{alt} \tag{B.7}$$

which simplifies to

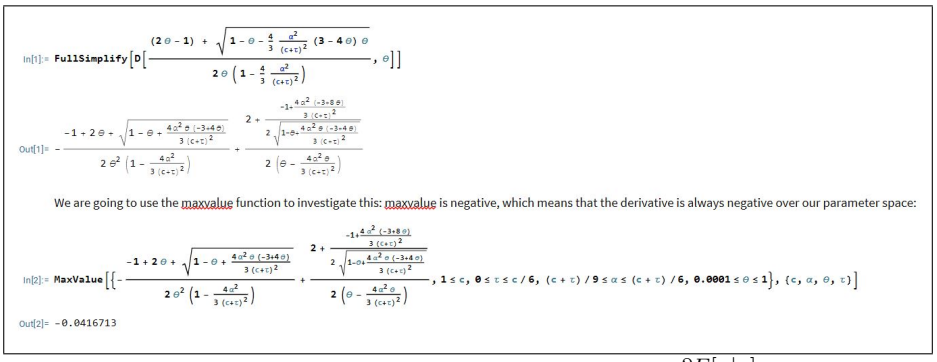
$$\alpha\varphi < q\left(R_u - R_k(q)\right) \tag{B.8}$$

and this is identical to (A.25). That is, the condition that previously identified household payment choice between DC and cash, now identifies household payment choice between DC and the alternative technology. Similarly, following through on other steps in Appendix A, nothing changes about the DC issuer optimization. For instance, (A.37) sees θ replaced with $\theta_{dc} - \theta_{alt}$ and therefore $\theta_{dc} = 1$ remains privately optimal. I.e., θ_{alt} is an added constant in the optimization to θ_{dc} that does not affect the optimality of $\theta_{dc} = 1$.

[4] We restrict attention to public information. Lender-specific data on a subset of households (such as depositors) would significantly complicate the model and lies outside the scope of the current analysis.

B.3 Intermediate y in the baseline model

This appendix analyzes the "Intermediate y" case of Table 1. From Sections 4.1 and 4.2, the interaction between y and model outcomes centers on whether the loan market for unrevealed borrowers functions. For $y > 2(c+\tau)$, there are always enough good borrowers with high privacy preferences among the cash users to ensure that the unrevealed obtain loans. For $y < \frac{4}{3}c$, the opposite is true. Instead, for $y \in \left(\frac{4}{3}c, 2(c+\tau)\right)$, Proposition 6 shows that either can be true, depending on parameters. Moreover, there can be a threshold $\widetilde{\theta}$, where $\theta > \widetilde{\theta}$ leads to an equilibrium as in Section 4.2 and $\theta < \widetilde{\theta}$ leads to an equilibrium as in Section 4.1. In particular, at $\theta = 0$, the loan market for the unrevealed functions for any $y > \frac{4}{3}c$. Hence, $\widetilde{\theta} \in (0,1)$ exists if at $\theta = 1$ the unrevealed do not obtain credit: this can happen because $\frac{\partial E[q|u]}{\partial \theta} < 0$, as can be shown from the expression for $E[q|u]$ found in (A.35):



Mathematica calculation showing that $\sup \frac{\partial E[q|u]}{\partial \theta} < 0$.

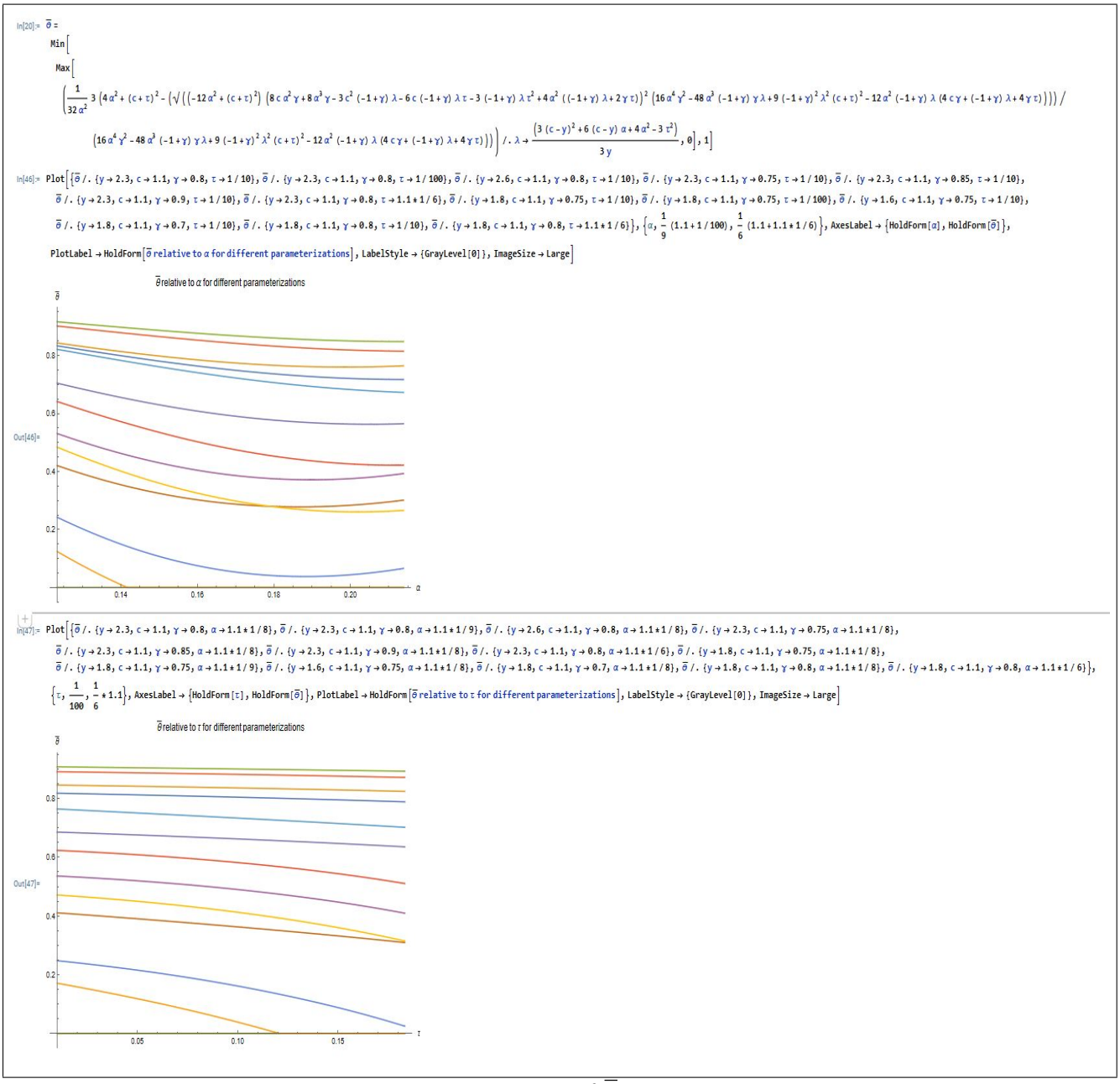
Proposition 6 *When $y \in \left(\frac{4}{3}c, 2(c+\tau)\right)$, then the equilibrium can be either as in Proposition 1 or as in Proposition 2, depending on parameters. For some parameterizations, there is a threshold value, $\widetilde{\theta}$, such that for $\theta < \widetilde{\theta}$ Proposition 1 prevails and for $\theta > \widetilde{\theta}$ Proposition 2. In such cases, Proposition 2 comes about, since the DC issuer continues to optimally set $\theta = 1$.*

Proof. First, from Proposition 2, for $y = \frac{4}{3}c$ the loan market is shut to the unrevealed even for $\theta = 0$. Then, for $y = \frac{4}{3}c + \varepsilon$ the same must be true for some $\theta > 0$, because $\frac{\partial E[q|u]}{\partial \theta} < 0$. This implies that there can be a threshold value of θ, $\widetilde{\theta}$, above which the outcome of Proposition 2 prevails and below which we get the outcome of Proposition 1. In cases where there is a $\widetilde{\theta} \in (0,1)$, the outcome will be as in Proposition 2, since the DC issuer continues to optimally set $\theta = 1$: this follows directly from the optimality of $\theta = 1$ in both Proposition 1 and Proposition 2. ■

We do not further investigate $\widetilde{\theta}$, but note that, for a given parameterization, $\widetilde{\theta}$ can be found numerically as a solution to $E\left[q|u\left(\widetilde{\theta}\right)\right] y = c$: this identifies the point where the home lender can expect zero profit from lending to the unrevealed market at the highest rate that borrowers are willing to accept, y. Moreover, the expression for $E\left[q|u\left(\widetilde{\theta}\right)\right]$ comes from replacing θ with $\widetilde{\theta}$ in (A.35).

B.4 Numerical comparative statics of $\overline{\theta}$ to α and τ

The figure below shows how socially optimal regulation (interior $\overline{\theta}$ from (18)) evolves relative to α and τ. The exercises vary α and τ, respectively, for given sets of other parameters, and draw out $\overline{\theta}$ on the vertical axis. In all cases, $\overline{\theta}$ declines as τ increases. However, $\overline{\theta}$ can be convex relative to α, and $\frac{\partial \overline{\theta}}{\partial \alpha}$ is non-monotonous for some parameter sets (i.e., $\overline{\theta}$ can be u-shaped in relation to α).



Mathematica numerical analysis of $\overline{\theta}$ with respect to α and τ.

B.5 Correlation of credit quality and privacy preferences

In the baseline model, the two dimensions of household heterogeneity, q and φ, are assumed to be independent of each other. This extension investigates correlation between household credit quality and privacy preferences. In the interest of tractability, we center attention on the cases of a perfectly positive and perfectly negative correlation of these features. In either case, the two dimensions of household heterogeneity collapse to a single dimension. In particular, $\varphi = 4q - 2$ represents a perfectly positive correlation, where $q = \frac{1}{2}$ implies $\varphi = 0$, while $q = 1$ implies $\varphi = 2$. Conversely, $\varphi = 4(1 - q)$ represents a perfectly negative correlation, where $q = \frac{1}{2}$ implies $\varphi = 2$, while $q = 1$ implies $\varphi = 0$. We report here on the main outcomes of the derivations, which are available in the *Mathematica* file included in this paper's supplemental files.

With a perfectly negative correlation the model collapses to a fully unraveling Lemons' Market, in which all households use DC. With a single dimension of households it is more challenging to generate a mechanism that "halts" cascades. This can only happen if high credit quality households, which have incentives to type differentiate, also care more about privacy, so that some of them sort into cash use, thereby cross-subsidizing lower quality households that choose to use cash. That is, when quality and care for privacy are positively correlated. In the case of a perfectly positive correlation, the indifference condition for choosing DC over cash becomes

$$q \geq \frac{E[q|u](c + \tau - 2\alpha)}{c + \tau - 4\alpha E[q|u]} \tag{B.9}$$

and solving for $E[q|u]$ using the same steps as before gives

$$E[q|u] = \frac{\sqrt{c+\tau}\sqrt{c+\tau+16\alpha} - (c+\tau)}{8\alpha} \tag{B.10}$$

Furthermore, $\theta = 1$ remains privately optimal for the DC issuer: although the mass of DC users, μ, does not depend on θ here, the mass of revealed borrowers is given by $m = \theta\mu$ which increases linearly as θ increases.

B.6 Partial disclosure

This extension considers how the model is affected when each household can decide for itself how much of its payment data it is willing to have observed. Instead of choosing one means of payment for consumption, the household can choose to consume a fraction ζ with DC and $1-\zeta$ with cash. We first consider

$$\max_{\zeta\in[0,1]} \left\{\theta\zeta\left(q\max\left\{y-R_k\left(q\right),0\right\}\right)+\left(1-\theta\zeta\right)\left(q\max\left\{y-R_u,0\right\}\right)-\alpha\varphi\theta\zeta\right\} \tag{B.11}$$

as the household optimization problem. Hence, $\theta\zeta$ now represents the probability of type revelation, which is jointly determined by how much disclosure the technology facilitates through θ and how intensely the household uses the technology for disclosure through ζ. In (B.11), the privacy cost remains proportional to the probability of revelation, like in our baseline model. In this extension, we focus on Table 1's "High y" case, because it is there that disclosure externalities matter. We assume in this extension that the DC issuer cannot make inferences directly from observing a household's ζ and households do not strategically choose ζ: incorporating such aspects is beyond the scope of this paper.

In the baseline model (in the "High y" case that is the focus of this extension), the equilibrium sorting condition is given by (9), which means that households choose DC when $\varphi\leq\frac{c+\tau}{\alpha}\left(\frac{q}{E[q|u]}-1\right)$ and choose cash otherwise. We note that the different objective function for households in this extension affects Stage 2 of the game. That is, the backward induction from Stage 5 to Stage 3 of the game is unaffected and remains as in the proof of Proposition 1. The household's optimization problem in (B.11) is linear in ζ. Therefore, optimal play is necessarily in the corners. A household obtains payoff $\theta q\left(y-R_k\left(q\right)\right)+\left(1-\theta\right)q\left(y-R_u\right)-\alpha\varphi\theta$ when $\zeta=1$, versus $q\left(y-R_u\right)$ when $\zeta=0$. Hence, the household chooses $\zeta=1$ if

$$\theta q\left(y-R_k\left(q\right)\right)+\left(1-\theta\right)q\left(y-R_u\right)-\alpha\varphi\theta>q\left(y-R_u\right) \tag{B.12}$$

and $\zeta=0$ otherwise, and this becomes

$$\alpha\varphi<q\left(R_u-R_k\left(q\right)\right) \tag{B.13}$$

which is identical to (A.24) and therefore all derivations are identical to proof of Proposition 1, as are the results, except that the household "chooses DC" is replaced with "chooses $\zeta=1$" and "chooses cash" is replaced with "chooses $\zeta=0$". Partial disclosure is never utilized because if a household wants to reveal its type (either intrinsically or because of cascades) it always wants to do so fully. The DC issuer continues to optimally set $\theta=1$ and therefore the baseline is fully replicated.

Quadratic cost variant

To bring about the possibility of optimal partial disclosure of households, we consider a second variant where privacy costs rise quadratically in relation to the probability of type disclosure. The household optimizes the same problem as in (B.11), except that $\alpha\varphi\theta\zeta$ is replaced with $\alpha\varphi\left(\theta\zeta\right)^2$. This setting does not readily lend itself to a full analytical solution (I.e., where the DC issuer internalizes each household's optimal ζ at Stage 1 and optimizes θ and data access fees accordingly). Nevertheless, we can show that there will be some

households that optimally mix payment instruments in this extension.

Again, the backward induction from Stage 5 to Stage 3 of the game is unaffected and remains as in the proof of Proposition 1. Replacing $R_k(q)$ and R_u, we can write the household optimization problem as

$$\max_{\zeta\in[0,1]}\left\{\theta\zeta q\left(y-\frac{c+\tau}{q}\right)+(1-\theta\zeta)\,q\left(y-\frac{c+\tau}{E[q|u]}\right)-\alpha\varphi(\theta\zeta)^2\right\} \tag{B.14}$$

and taking a first-order condition to solve for an interior solution gives

```
In[•]:= FullSimplify[Solve[D[θ*ζ*q (y - (c + τ)/q) + (1 - θ*ζ)*q (y - (c + τ)/Equ) - α*ϕ*(θ*ζ)^2, ζ] == 0, ζ]]
Out[•]= {{ζ → -((Equ - q) (c + τ))/(2 Equ α θ ϕ)}}
```

which we can rewrite to

$$\zeta = \frac{1}{2}\frac{c+\tau}{\alpha\theta\phi}\left(\frac{q}{Equ}-1\right)$$

Mathematica solution for optimal ζ.

$$\zeta=\frac{1}{2}\frac{c+\tau}{\alpha\varphi\theta}\left(\frac{q}{E[q|u]}-1\right) \tag{B.15}$$

Taking into account the corner solutions from the boundaries on $\zeta\in[0,1]$, we get optimal play as:

$$\zeta=\min\left\{\max\left\{\frac{1}{2}\frac{c+\tau}{\alpha\varphi\theta}\left(\frac{q}{E[q|u]}-1\right),0\right\},1\right\} \tag{B.16}$$

Hence, some households optimally set $\zeta\in(0,1)$ and thus mix means of payment. Intuitively, for such households the optimal ζ rises as q increases and as φ declines: households with better credit quality and lower privacy preferences use more DC in their payment mix.