Smart Lending

Mario Milone*

November 16, 2018

[Job Market Paper]

Abstract

This paper investigates the impact of data-based lending on financial intermediation. I show that a data-based screening technology can increase financial frictions. The use of data in the screening process reduces the acquisition of soft information by the lender which negatively impacts already constrained borrowers. Additionally, since borrowers that belong to groups that were less financed historically are under-represented in the data, the technological lender's screening efficiency differs in the cross-section of borrowers and is higher for borrowers with greater historical lending data. When traditional and technological lenders coexist, the borrowers about whom data can provide precise information raise funds from technological lenders while those with less informative historical data choose traditional lenders who can make up for the lack of hard data-based information by acquiring soft information. The intermediation cost to traditional borrowers is increased by the existence of technological lenders. I identify conditions under which traditional lenders benefit from restricting their own access to data processing technology when competing against the technological lender.

*Imperial College Business School and Université Paris Dauphine. I am deeply undebted to Gilles Chemla for his invaluable support, help, and patience. This paper has also benefited from comments from Denis Gromb, Christopher Hennessy, Christine Parlour, Ailsa Roell, Jérôme Dugast, Franklin Allen, Ansgar Walther, Olivier Scaillet, Dragana Cvijanovic, Tarun Ramadorai and seminar participants at Université Paris-Dauphine and Imperial College Business School.

1 Introduction

Information processing is entering a new age. Fast increases in the production and storage of data and in computing capacity are the two major ingredients for the big data revolution which quickly transforms the finance industry.[1] Today's top mortgage originator in the U.S. is Quicken, which uses automated algorithms and underwriting to perform lending decisions.[2] The rise of Fintech, big data, and machine learning has widely been recognized by the academic community, but *"there is no reason to think [...] that these innovations will automatically enhance stability or even access of service"* (Philippon, 2016).

This paper investigates how the intensive use of data in the lending process impacts financing decisions. I argue that technological and traditional lenders have very different attitudes toward screening, which is perhaps the most important aspect of capital allocation. The ability for financiers to assess the quality of investment opportunities, projects or mortgages, is of critical importance to a well functioning economy, and is the raison d'être of financial intermediaries. As an information intensive activity, screening is likely to be hugely impacted by the increasing use of data. Quicken's website "Rocket Mortgage" advertises a mortgage approval process of only a few minutes, thanks to an automated screening process of applicants. While the benefits of a more automated and intensively data-based screening technology such as speed and cost-efficiency may be easy to appreciate, its costs are far less understood. This paper shows that a data-based screening process is likely to aggravate financial frictions and intensify credit rationing.

Assessing borrowers credit-worthiness is a complex task that requires various sources of information. The literature typically distinguishes between two types of information: hard information, which is viewed as factual, quantitative, and easy to store and to substantiate, and soft information, which can only be acquired through human interaction and is more subjective and difficult to communicate and/or store (see e.g. Liberti and Petersen (2017), Berg, Puri, and Rocholl (2013), and Agarwal and Ben-David (2014)). Understanding the difference between these two types of information is highly relevant in the context of information processing, as not all types can be easily used by automated algorithms. While the definitions of soft and hard information varies across the literature, this paper defines a hard information as one that can be processed by a machine and used

[1] A 2017 report from IDC forecasts a ten-fold increase in data-creation over the next 10 years, from 16 Zetabytes in 2016 to more than 160 in 2025.

[2] Quicken has replaced Wells Fargo as top US retail mortgage lender in Q1 2018. Among the top ten U.S. mortgage originators, six of them can be considered Fintech firms in the sense that the use of automated processing of data and information is at the center of their business process.

by predictive algorithms, and a soft information as one that can only be acquired through human interaction.

This paper identifies two channels through which the use of data in the screening process can aggravate financial constraint created by the presence of moral hazard. First, the use of historical lending data based on hard information lowers the incentives for traditional lenders to acquire and process soft information in assessing the creditworthiness of borrowers. As a result, borrowers for whom soft information is an important dimension in determining their creditworthiness become more financially constrained when data is used in the screening process. Second, one of the key characteristics of data-based knowledge is that the precision of that knowledge reflects the availability and structure of the data used to construct it. In the context of lending, a data-based screening technology tends to be more precise and more efficient for applicants that are more represented in the data. As a result, borrowers that are already financially constrained due to the presence of moral hazard become even more so *compared* to the one that are less affected by the moral hazard problem, as the lender has a better ability to screen the latter. Importantly, this second channel does not result from possible existing biases present in the data. Rather, it is driven by the difference in the precision of the information extracted from the data. Since lending decisions depend on the efficiency of the screening process, the difference in knowledge precision inherent to a data-based approach tends to negatively affect groups of borrowers that are under-represented in existing data. These groups are in turn less likely to obtain credit, they have less data generated about them, which further increases the difference in the precision of screening across groups.

I also analyze an economy in which both traditional and technological lenders coexist. The traditional lender has both the ability to use historical lending data and to acquire soft information to screen borrowers. In contrast, the technological lender can only use data, but is more efficient at doing so. I show that when both traditional and technological lenders coexist, borrowers tend to separate between the two depending on their individual characteristics. Borrowers whose soft information is important in determining their creditworthiness prefer to seek financing from traditional lenders while the others are siphoned off by technological lenders. Interestingly, borrowers staying with the traditional lender may be better off if the latter does not make a heavy use of data in her screening process. Additionally this siphoning-off effect increases the intermediation costs of traditional lenders because the traditional lender faces a pool of borrowers whose average screening cost increases due to the entrance of the technological lender.

I examine these effects in a two periods model of firm financing featuring moral hazard and a data-based screening technology. Borrowers are heterogeneous in both their amount of net worth and the extend to which soft information matters in assessing their creditworthiness. In the first period, the traditional lender does not have access to historical lending data and only makes use of soft information to screen borrowers. The second period lending game is affected by the data generated during the first period. The existence of historical lending data affects the incentives for the traditional lender to exert costly effort for acquiring soft information which is detrimental to the borrowers for whom soft information matters. Additionally, the presence of moral hazard in the first period lending results in richer borrowers being more financed than poorer borrowers. As a consequence, in the second period, the lender has access to more historical data about rich borrowers which relaxes their financing constraint more compared to poor borrowers. In the second period I allow for a purely technological lender – one who only uses data ito screen borrowers – to coexist with the traditional lender. Because the technological lender is more efficient at extracting information from the data, she is able to siphon-off (from the traditional lender) borrowers for whom soft information is relatively less important in determining their creditworthiness. This separation result affects the average cost of screening for the traditional lender. Because she exerts more effort for borrowers whose soft information matters, the entrance of the technological lender in the credit market *increases* her average screening cost. While the technology efficiencies are assumed exogenous at first, I subsequently study the incentives for the traditional lender to restrict her use of data processing technology. A trade-off is revealed whereby increasing her use of data allows the traditional lender to compete with the technological lender and attract borrowers whose soft information is relatively less important. However, doing so increases credit-rationing for borrowers whose soft information is important. As a result, the traditional lender prefers to lower her technological efficiency when the population of borrowers contain relatively more of those whose soft information matters.

This paper is first and foremost related to the vast banking literature on screening and information. It has been repeatedly argued that one advantage of bank financing comes from banks ability to acquire and process soft information through lending relationships thereby providing superior screening and monitoring services (e.g., Allen (1990) and D. W. Diamond (1984)). Boot (2000) defines relationship banking as the process in which the lender gathers information (i) "beyond readily available public information", (ii) "over time through multiple interactions with the borrower",

and that (iii) "remains confidential (proprietary)". Petersen and R. Rajan (1994) provide empirical evidence showing that stronger relationships between lenders and borrowers are associated with increased availability of credit for small firms. Petersen and R. G. Rajan (1995) argue that competition in credit market lowers the value of lending relationship as it becomes harder for lenders to internalize the associated benefits. However, Petersen and R. Rajan (2002) provide empirical evidence that the increase in information technology allow lenders to provide credit to firms that are more distant, and with whom communication is more impersonal (see also Berger, Frame, and Miller (2005)).

The model captures that characteristic of bank lending by giving the traditional lender the unique ability to use soft information in the screening process. While the dynamic effects of such borrower-lender relationship have been the subject of many studies (e.g. Bolton and Scharfstein (1990), Dewatripont and Maskin (1995), and D. Diamond (1991)), this paper does not consider repeated relationships and the existence of possible long-term contracts between lenders and borrowers. While featuring a dynamic two periods setting, each period is considered as a static problem in the spirit of Innes (1990). Relatedly, I focus on the screening (e.g. Inderst and Mueller (2006)) rather than monitoring (e.g. (Townsend (1979) and Gale and Hellwig (1985)) problem faced by the lender. While contracting in every period is a static problem in our model, the interesting dynamic effects come from the use of a data-based screening technology by the lender, thereby contributing to the literature studying the effect of data and information on lending (e.g. Berg, Puri, and Rocholl (2016), Padilla and Pagano (1997), Padilla and Pagano (2000), Pagano and Jappelli (1993), and Brown, Jappelli, and Pagano (2009)).

Understanding the impact of information and data has become increasingly relevant in recent years with the rise in data availability and predictive algorithms. More specifically, Fuster et al. (2018) study the impact of the use of non-linear prediction techniques and their impact on mortgage lending and show that an increase in technology complexity may lead to discrimination in the mortgage lending market. Our paper relates to Fuster et al. (2018) by capturing an essential characteristic of data-based screening technology, specifically the fact that the structure of the data affects the performance of the assessment of borrowers creditworthiness in their cross-section (see also U. Rajan, Seru, and Vig (2015)). Berg, Burg, et al. (2018) show that digital information left by potential borrowers via the use of digital services – denoted digital footprint – provide relevant information for the purpose of assessing the creditworthiness of borrowers. Relatedly, Jagtiani and

Lemieux (2018) provide evidence that the use of alternative data sources improve the assessment of creditworthiness by providing different type of information compared to the traditional FICO scores and Paravisini and Schoar (2015) show that the use of credit score affect lenders incentives to exert effort. Our model encompasses these findings by allowing for a substitutability between the use of data and the use of soft information.

Because the paper considers the problem as being of informational in nature, it relates to the role of information in investment decisions (e.g. Cabrales, Gossner, and Serrano (2013), M. a. Meyer (1991), M. A. Meyer and Zwiebel (2007), and Shorrer (2015)), the role of information in the principal-agent relationships (e.g. Levitt and Snyder (1997) and Chaigneau, Edmans, and Gottlieb (2017)), and more importantly to the literature on learning and dynamic inattention (e.g. Ellis (2018), Mayskaya (2017), Matejka and McKay (2015), and Nimark and Sundaresan (2018)). In fact, one way to appreciate the problem of using hard and/or soft information from the lender's point of view is to consider them as two sources of information that the lender can choose from and focus more or less attention on.

The rest of the paper proceeds as follows. I first present the problem of the first period lending by the traditional lender where only soft information is used in the screening process. I then consider the second period lending in which the traditional lender has the ability to use historical lending data and I show the effect on the borrowers financing constraint. I then introduce the purely technological lender and study the coexistence of the two types of lenders. I show how borrowers separate between the lenders and that the average screening cost of the traditional lender increases. I then turn to the choice of technological efficiency for the traditional lender and identify conditions under which she prefers to restrict her use of data. I subsequently study the dynamics effect of the use of data on financial constraints and describe the dispersion effect – where the dynamic interaction of lending and screening increases the effect of moral hazard. I finally discuss the results and conclude.

2 The traditional lender's problem

This section studies the problem of a traditional lender. The traditional lender has the ability to use both soft information and historical lending data, when available, to screen borrowers. I first focus on the traditional lender's problem in the first period, where historical data is not available, and only soft information is used to screen borrowers. I then study the second period lending decision

where the traditional lender can use the first period lending data to help in the screening process. In the next section, a purely data-based lender, denoted technological lender will be introduced and the coexistence of both types of lenders will be studied.

2.1 Agents and Technologies

There is a representative lender and good and bad borrowers in equal proportion. Each borrower, indexed by k, has access to a project requiring an investment of 1. A good borrower's project returns R with probability q if the borrower behaves. If the borrower shirks, the project returns zero, but the borrower enjoys a private benefit B_k. In case of failure, the project returns zero. A bad borrower's project returns zero for sure, regardless of whether the borrower behaves or shirk. The returns for good and bad borrowers are illustrated in figure 1. The borrower's type – good or bad – is unobservable to *both* the representative lender and the borrower and we assume that good borrowers that behaves have positive NPV projects.

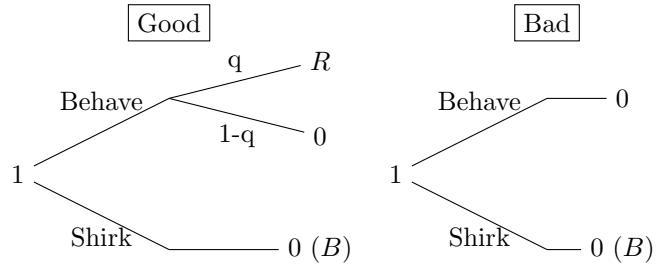

Figure 1: Projects returns

Assumption 1. *Good borrowers have positive NPV projects when they behave*

$$qR > 1$$

Assuming that both the lender and the borrower are unaware of the borrower's type ensures symmetry of information at all point in time. While strong, this assumption effectively removes adverse selection problems arising from information asymmetries. This allows to focus on the impact of moral hazard alone.

There are two periods $t = \{0, 1\}$. At the beginning of each period, N borrowers are born and live for one period, hence dismissing any possibility for long-term relationship or contracts. Each borrower is endowed with net worth $A_k \in (0, 1)$ and private benefit B, unique for all borrowers.

Additionally, each borrower k is characterized by a parameter γ_k that captures how important soft information matters in determining his credit-worthiness. The role of γ_k will become clear shortly.

A_k is uniformly distributed over $(0,1)$ and γ_k is distributed on $(0,\overline{\gamma})$ following the distribution $\Gamma(.)$. Finally, each borrower is associated with a vector of individual characteristics X_k that are predictive of his type. Net worth A_k, private benefit B, importance of soft information γ_k and characteristics X_k are perfectly observable by all agents.

Every period, borrowers apply to the lender for credit, who decides whether or not to finance each of them. Each borrower approaches the lender and asks for the lender to finance $(1-A_i)$ of the project's required investment. The borrower may choose to use all of his net worth, in which case $A_i = A_k$ or only part of it and $A_i < A_k$. The lender then screens the borrower. The screening outcome is modeled by an informative signal about the borrower's type, observed by the lender. Upon receiving the signal, the lender updates her belief about the success probability of the borrower and offers a contract specifying the repayment r_k to the lender if the project succeeds. The lender receives a binary signal $s_{t,k} \in \{0,1\}$ with precision $\tau_{t,k} \in \{\frac{1}{2},1\}$ such that

$$P[s_{t,k}=1|Good] = P[s_{t,k}=0|Bad] = \tau_{t,k}$$

As good and bad borrowers exist in equal proportion in the economy, and because borrowers do not observe their types, the unconditional probabilities that a good or bad borrower approaches the lender for financing are equal. Additionally, and without loss of generality, it is assumed that screening is necessary for financing in that the lender needs to acquire information before accepting to finance the borrower.

Assumption 2. *The prior belief about the borrower's type is such that*

$$P[Good] = P[Bad] = \frac{1}{2}$$

and screening is necessary for financing

$$qR < 2$$

Assumption 2 ensures that the lender refuses to finance borrower k unless she has access to an informative signal about his type. As a result, financing is always refused if the lender receives a signal $s_{t,k} = 0$. In the first period (period 0), the traditional lender does not have access to historical

lending data, and the signal precision only depends on how much soft information is acquired and processed. The first period signal precision follows the following functional form.

$$\tau_{0,k} = \frac{1}{2} + \gamma_k e_k \tag{1}$$

e_k denotes the effort exerted by the lender to acquire soft information. In the context of this model, soft information refers to the type of information that can be acquired through human interaction with the borrower. As all borrowers are born in the beginning of the period, it does not include information acquired over time (through long term relationship). γ_k captures the importance of soft information in assessing the credit-worthiness of borrower k. When soft information matters more (higher γ_k), acquiring soft information is more valuable and the signal precision is higher, given a level of effort. Acquiring and processing soft information is costly for the lender in the following way.

$$c(e_k) = \alpha \left(\frac{e_k^2}{2} \right) \tag{2}$$

For simplicity and tractability, I assume that the cost of effort is a quadratic function of the effort exerted by the lender. The results go through as long as the cost of effort is an increasing and convex function. The parameter α is a scaling factor ensuring that at the optimum level of the lender's effort, the signal precision is not greater than 1.

The timing is as follow, and illustrated in figure 2. At the beginning of the period, borrower k approaches the lender and ask her to finance $(1 - A_i)$ of the project financing cost. A_i is a choice variable for the borrower such that $A_i < A_k$ where A_k is the total wealth of borrower k. Then, the lender decides to exerts effort e_k at cost $c(e_k)$. After exerting the effort, the lender receives the signal $s_{0,k}$ with precision $\tau_{0,k}$. Upon receiving the signal, and if the lender decides to finance the project, she offers a contract r_k that maximizes her expected profits as will become clear shortly. Subsequently, if the borrower accepts the contract, he decides to behave or shirk, and finally returns realize and payoffs are distributed. The next subsection solves the optimal contract.

2.2 Optimal Contract

The optimal contract is solved by backward induction. From assumption 2, the lender only consider financing borrower k if she receives a signal $s_{0,k} = 1$, and the ex-ante probability to receive such a signal is $\frac{1}{2}$, regardless of effort e_k. Let us denote $p(e_k)$ the posterior belief about the probability

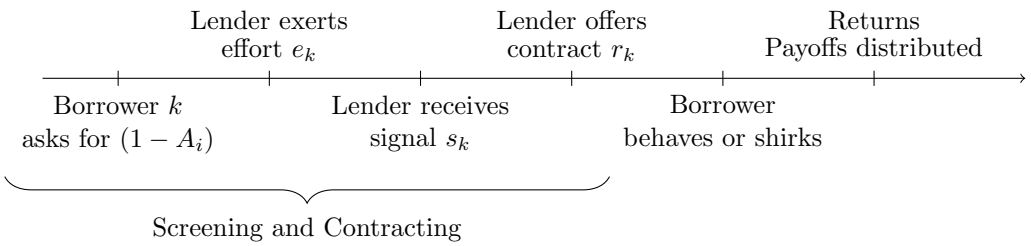

Figure 2: One period contracting timeline

of success of the borrower conditional on receiving $s_{o,k} = 1$ and that the borrower behaves. The borrower solves the following program.

$$\max_{A_i} \frac{1}{2} \left[p\left(e_{eq}\right) \left(R - r_k\right) - A_i \right] + A$$

$$\text{s.t. } p\left(e_{eq}\right) \left(R - r_k\right) \geq B \qquad (IC_B)$$

$$e_{eq} = \arg\max_{e} \frac{1}{2} \left[p\left(e_k\right) r_k - \left(1 - A_i\right) \right] - c\left(e_k\right) \qquad (IC_L)$$

$$p\left(e_{eq}\right) \left(R - r_k\right) \geq A_i \qquad (PC_B)$$

$$\frac{1}{2} \left[p\left(e_{eq}\right) r_k - \left(1 - A_i\right) \right] - c\left(e_{eq}\right) \geq 0 \qquad (PC_L)$$

$$A_i < A \qquad (LL_A)$$

$$r_k < R \qquad (LL_r)$$

The borrower chooses A_i to maximize his expected profit given effort e_{eq} exerted by the lender. Given A_i, the lender chooses the the level of effort e_{eq} that maximizes her profits minus the cost of effort, as described by the incentive compatibility constraint of the lender, equation (IC_L). If the lender receives the signal $s_{0,k} = 1$, she offers the contract r_k where r_k is the repayment to the lender if the project succeeds. r_k is such that the borrower is incentivized to behave, and needs to satisfy the incentive compatibility constraint of the borrower, inequation (IC_B). Equations (PC_B) and (PC_L) ensure that both parties want to participate, and equations (LL_A) and (LL_r) are limited liability constraints.

It is important to note that the incentive compatibility constraint of the borrower, inequation (IC_B), is a function of $p\left(e_{eq}\right)$, the posterior belief of the lender after receiving the signal $s_{0,k} = 1$. This is due to the fact that ex-ante, neither the lender nor the borrower know whether the borrower is good or bad. However, after the lender offers the contract r_k, the borrower can update his belief

about his own type as r_k is a sufficient statistic. His decision to behave or shirk is therefore based on the same updated belief than the lender. As the lender offers the contract, she chooses r_k such that the borrower receives just enough to behave.

$$r_k = R - \frac{B}{p\left(e_{eq}\right)}$$

The program can be rewritten (omitting the limited liability constraints for readability) as follows.

$$\begin{aligned} &\max_{A_i} \frac{1}{2}\left[B - A_i\right] + A_k \\ \text{s.t. } &e_{eq} = \arg\max_{e} \frac{1}{2}\left\{\left[p\left(e_k\right)R - 1\right] - \left[B - A_i\right]\right\} - c\left(e_k\right) \qquad (3) \\ &\frac{1}{2}\left[p\left(e_{eq}\right)R - 1\right] - c\left(e_{eq}\right) \geq \frac{1}{2}\left[B - A_i\right] \geq 0 \qquad (4) \end{aligned}$$

Inequation (4) combines the participation constraints of the lender and the borrower. Note that if $B < A_i$, the borrower is better off keeping his assets rather than investing them into the project while if A_i is too low the lender chooses not to participate. Hence, the borrower wants to choose A_i sufficiently low so that the constraint (PC_B) is satisfied and sufficiently high so that constraint (PC_L) is also satisfied. Given A_i, the lender chooses effort $e_{eq}\left(A_i\right)$ that satisfies the first order condition of equation (3). The borrower in turn chooses A_i to maximize his profits given the effort $e_{eq}\left(A_i\right)$ that the lender will exert. Lemma **??** lays down the optimal contract. It is important to note that while the lender exerts an effort that maximizes her profit, the borrower chooses A_i such that the lender is in zero-profits.

Lemma 1. Optimal contract without data

The optimal contract is such that the lender's effort satisfies

$$e_{eq} = \frac{1}{2\alpha} qR\gamma_k \qquad (5)$$

and the borrower's profits when the lender does not use historical lending data are

$$\pi_k = \frac{1}{2}\left[\frac{1}{2}qR - 1\right] + \frac{1}{2\alpha}\left(\frac{1}{2}qR\gamma_k\right)^2 \qquad (6)$$

Lemma 1 highlights that the borrower's profits are increasing in the importance of soft infor-

mation for assessing the borrower's creditworthiness. As γ_k increases, the lender exerts higher level of effort as it is more valuable to acquire soft information. In turn, screening is more efficient and the borrower's expected profits are higher. Plugging the optimal effort e_{eq} into the participation constraint (4) allows us to find the condition for financing to occur.

$$\pi_k(\gamma_k) \geq \frac{1}{2}(B - A_k) \tag{7}$$

Equation (7) makes clear that borrower k needs to finance enough of the project in order to obtain financing. If the borrower is not able to finance enough of the project's cost, it is too expensive for the lender to incentivize him to behave, and financing does not take place. This result is common in models of financing in the presence of moral hazard. Note that $\pi_k(\gamma_k)$ is increasing in γ_k. Therefore, for a given level of net worth A_k, borrower k needs to have a γ_k high enough so that it is profitable for the lender to exert effort in the screening process. On aggregate, only the fraction of borrowers with enough wealth A_k and/or high enough γ_k are able to access financing, due to the presence of moral hazard. Figure 3 illustrates the financing constraint resulting from moral hazard.

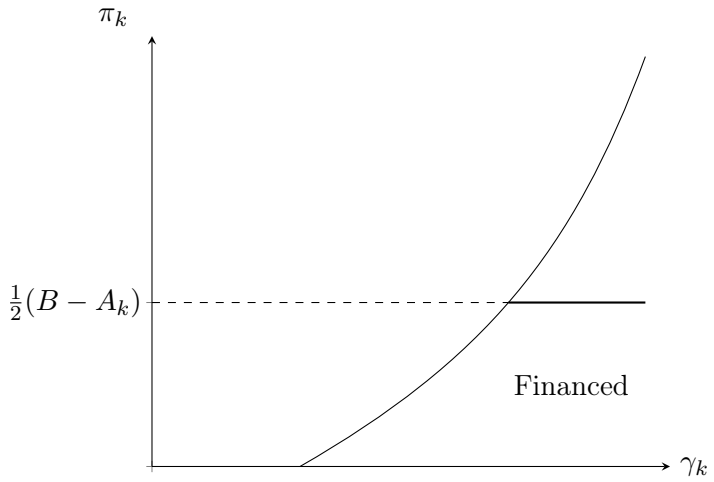

Figure 3: Financial constraint

The horizontal axis of figure 3 is the importance of soft information in assessing the creditworthiness of borrower k, γ_k. The increasing convex curve is the borrower's profits π_k under the optimal contract. The horizontal line represents the right hand side of the financial constraint (7). The higher the net worth A_k of borrower k, the lower the horizontal line. Financial constraint (7) states that all borrowers that are under the curve are able to obtain financing, as their combinations of (A_k, γ_k) satisfy the financial constraint arising from the presence of moral hazard. The solid hori-

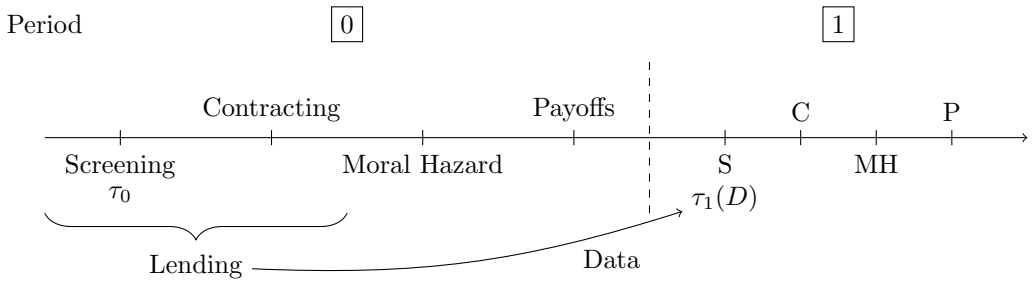

Figure 4: Timeline

zontal line on the right of the curve can be interpreted as the number of borrowers with a given net worth that are able to access financing. As the net worth of the borrower decreases, fewer borrowers are able to access financing (only those for whom soft information is informative enough). The slope of the curve represents the marginal effect of moral hazard on the financial constraint, as it captures how much more a borrower is constraint when his wealth marginally decreases.

We now turn to the second period lending to study how the usage of data in the screening process impacts the financial constraint of borrowers.

2.3 Screening with data

At the beginning of the second period, the lender has access to historical lending data consisting of all the borrowers that have been financed in the first period. The lender has lent to borrowers with observable characteristics X_k and has also observed whether those borrowers have defaulted or not, let us denote this information the output Y_k. As has been described previously, the observable characteristics X_k are predictive of the borrower type, good or bad. Using the period one data (X_k, Y_k) the lender can try to learn the mapping between X_k and Y_k in order to predict better the creditworthiness of the borrowers applying for credit in period 2. The way this is performed in practice is through the use of predictive algorithms. These could be logistic regressions or more complex algorithms from the field of machine learning. Such a learning problem, where both the inputs X_k and outcomes Y_k are used to learn the mapping between the two is typically tackled with the use of so called supervised predictive algorithms. For a more detailed explanation of the learning process, the reader is referred to section A of the appendix. Figure 4 illustrates the dynamic link between the first period lending decisions and the second period screening.

The model does not seek to capture the details of the algorithm used by the lender to infer the mapping between $X_k$and Y_k, as many of them are available and they are constantly evolving. Instead,

it seeks to capture fundamental characteristics of such a learning process in order to understand the implications for lending decisions. I assume that the outcome of learning using historical lending data depends on two factors. First, the lender can learn better the mapping between X_k and Y_k if more data is available. Second, the learning is of better quality if the technology used – performance of the algorithm – is more efficient. In other words, the lender can extract more information from historical lending data if (i) there is more data available and/or (ii) if the technology used is better able to extract useful information from the available data. It is important to note that the information given by past lending is not dependent on having the same borrowers seeking financing. Instead, it is assumed that all borrowers from the first period die before the beginning of the second period, and that newly born borrowers ask for financing in the second period. Nevertheless, the functional dependence between X_k and Y_k is similar enough between the first and the second period such that historical lending data is informative when screening new borrowers in the second period. Historical lending data is best thought of as capturing hard information that is useful in assessing the creditworthiness of the borrower.

Let us denote D the amount of data available to learn the mapping between X_k and Y_k and η the efficiency of the technology in use. The use of historical lending data modifies the precision of the second period $\tau_{1,k}$ and the cost of effort $c\left(e_{k)}\right)$ as follows.

$$\tau_{1,k} = \frac{1}{2} + \gamma_k e_k + \eta D \tag{8}$$

$$c(e_k) = \alpha \left(\frac{e_k^2}{2} + \eta D e_k \right) \tag{9}$$

Equations (8) and (9) capture two characteristics of the link between hard and soft information. First, it captures the fact that hard and soft information provide different types of information useful in assessing the borrower's creditworthiness. That is, even if one has access to hard information, it might still be useful to exert effort to acquire soft information. Second, equation (9) captures the substitutability between hard and soft information. If the lender already has access to historical lending data to assess the borrower's credit worthiness, the effort required to increase the precision of the screening signal will be higher because both sources of information aim at reducing the same uncertainty. The informational problem can equivalently be seen as having the lender accessing two sources of information, hard and soft. Both sources reduce the uncertainty about the borrower's type. If the lender already has access (costlessly) to the hard information source, the marginal value

of the soft information lowers – equivalently, the marginal cost of soft information increases. The program solved by the borrower is identical to the previous case, except for the change of functional forms. The optimal contract and effort is laid out in the next lemma.

Lemma 2. Optimal contract with data

The optimal contract is such that the lender's effort satisfies

$$e_{eq,D} = \frac{1}{2\alpha} qR\gamma_k - \eta^H D \tag{10}$$

and the borrower's profits when the lender uses historical lending data D are

$$\pi_{k,D} = \pi_k + \frac{1}{2}\eta D \left[qR\left(1-\gamma_k\right) + \alpha\eta D\right] \tag{11}$$

Financing is possible if and only if

$$\pi_{k,D}\left(\gamma_k, D\right) \geq \frac{1}{2}\left(B - A_k\right) \tag{12}$$

The new optimum effort exerted by the lender is now lowered by the fact that historical lending data is available to help the lender in screening the borrowers. When more data is available, the marginal benefit of soft information is lower – equivalently, the marginal cost of soft information is higher –, and the lender optimally reduces the acquisition of soft information. Equation (11) shows that the borrower's profits when the lender can use data equals to the profits when no data is used plus an additional term increasing in the amount of data and technology efficiency ηD and decreasing in the the importance of soft information γ_k. On the one hand having more data available to screen borrower k increases the signal precision and the borrower's profits. On the other hand, more data lowers the use of soft information by the lender, which negatively impacts borrowers for whom soft information is important in determining their creditworthiness. When γ_k is high enough, the fact that the lender has access to historical data is detrimental to the borrower, compared to the case where the lender does not have access to data. In other words, the availability of data discourages the lender's effort to use soft information, which negatively affects borrowers with high γ_k.

Lemma 3. *There exists $\tilde{\gamma}$ such that*

- *If $\gamma_k < \tilde{\gamma}$, the use of historical lending data* increases *the expected profits of borrower k.*
- *If $\gamma_k > \tilde{\gamma}$, the use of historical lending data* decreases *the expected profits of borrower k.*

where $\tilde{\gamma} = 1 + \alpha\frac{\eta D}{qR}$.

An important implication of the use of data by the traditional lender is that it increases the difference in financial constraints between richer and poorer borrowers, effectively increasing the impact of moral hazard on credit-rationing. To see that, one need to look at the slope of $\pi_{k,D}$ – with data – and how it differs from the one of π_k– without data –, because the financial constraint takes an identical form in both cases, as shown by inequation (12). The slope of the borrowers profits indicates the extend to which borrowers with low net worth are constrained *compared* to borrowers with high net worth. It can easily be shown that the slope of $\pi_{k,D}$ is lower than the one of π_k indicating that when the lender uses data when screening borrowers, there is a larger difference in the financial constraint of borrowers depending on their net worth. This leads to the next proposition.

Proposition 1. *The availability of historical lending data in the screening process accentuates the effect of moral hazard on credit-rationing.*

Figure 5 illustrates this result. Subfigure 5a shows the difference in credit-constraint between borrowers with different net worth. The lower the threshold on the horizontal axis, the higher the borrower's net worth A_k (as the threshold is $\frac{1}{2}(B - A_k)$). There are more borrowers with higher net worth that can access financing as the minimum γ_k that is required is lower. When the traditional lender uses data in the screening process (subfigure 5b), the slope of the borrowers profits lowers, and the profits are equal at the point $\tilde{\gamma}$. As the slope decreases, the difference between the credit constraints of rich and poor borrowers increases.

In the next section, I introduce a purely technological lender and subsequently study the outcome of the coexistence of both types of lenders.

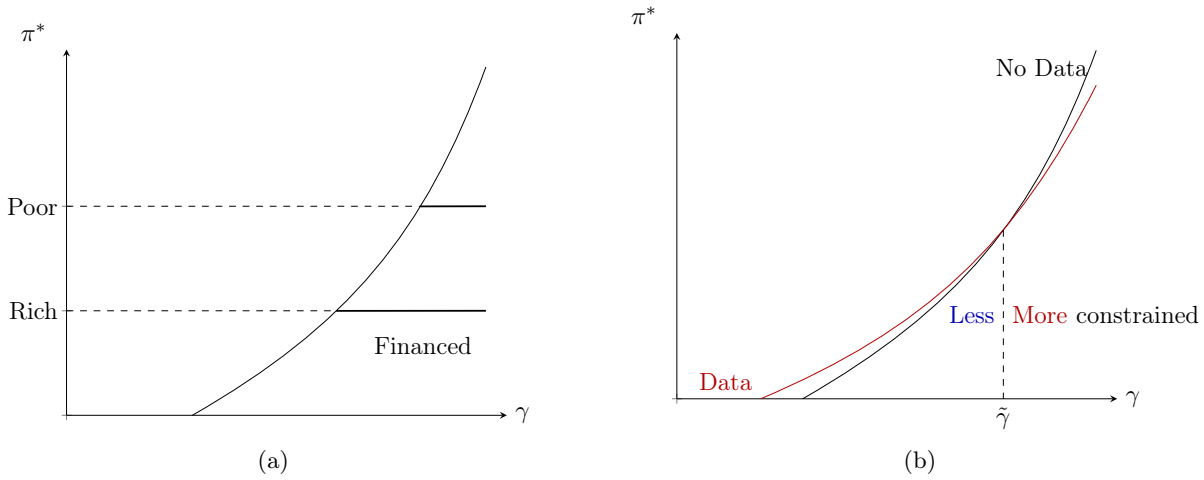

Figure 5: Effect of the use of historical lending data on credit-constraint.

3 Technological lender and coexistence with the traditional lender

3.1 Technological lender

We observe a rise in the presence of lenders whose screening process are mainly – if not only – based on the use of data. One can think of Quicken loans in the mortgage market. Most of these lenders only offers an online application process and promise much faster processing of the borrowers applications than that of traditional lenders. The main difference with traditional lenders is that these "technological lenders", as they are referred to in this paper, do not interact with the borrowers personally. Instead they make an extensive use of historical lending data to predict the quality of the applying borrowers in order to decide whether credit should be granted. In a way, these lenders make a greater use of hard information of borrowers, the type of information that can easily be transmitted through electronic systems and verified. That being said, the model acknowledges the fact that this hard information can be used to partly proxy what might be characterized as being soft information. For instance, assume that the lender somehow has access to the shopping pattern of the borrower applying for credit. That might inform the lender about the type of shopper the borrower is – cautious or not for instance. This substitutability between hard and soft information is captured in the functional form used in the traditional lender's problem.

I define a technological lender as one that can only make use of hard information, at zero cost. I additionally assume that the technological lender is better able to extract information from the available data. One may consider that the technological lender is specializing in the use of predictive

technologies and has an advantage on that area compared to the traditional lender, perhaps through ex-ante investment in the technology. Once the technology is ready to use, however, one can imagine that the technological lender only has to "push a button" to assess the creditworthiness of the borrower. While the difference of technological efficiencies between the traditional and technological lenders is assumed at first, I later consider the choice of the traditional lender in her use of data and identify conditions under which she might want to reduce her reliance on such technologies.

The precision of the signal received by the technological lender is

$$\tau_1^T = \frac{1}{2} + \eta^T D$$

where $\eta^T > \eta$ is the technological efficiency of the technological lender.

The optimal contract and the program solved by the borrower is identical as with the traditional lender albeit simpler as the technological lender does not have the ability to exert any effort. The signal precision only depends on the amount of historical data available, D, and η^T. Moral hazard is still present thereby financing is still constrained.

Lemma 4. *Borrower k can access financing through the technological lender if and only if*

$$\pi^T = \frac{1}{2}\left[qR\tau_1 - 1\right] \geq \frac{1}{2}\left(B - A_k\right)$$

Lemma 4 makes clear that the financial constraint of the technological lender is now independent of the importance of soft information, γ_k, in assessing the borrower's credit worthiness [3]. Figure 6 plots the financial constraint for the technological lender. Similarly to the case of the traditional lender, the borrower can only access financing if he can cover a large enough part of the project's cost. On the figure, all borrowers with a sufficiently large net worth (borrowers below the line π^T) can access financing.

3.2 Coexistence with the traditional lender

One of the question of interest when studying the impact of the use of data is to analyze how the entrance of a purely technological lender affects the traditional lender. To that end, I introduce

[3]Note that the fact that the financial constraint of the technological lender is independent of γ_k is not robust to the functional form used. One may intuitively expect that the borrower's profits might decrease in the important of soft information as the technological lender does not have the ability to use it. The subsequent results would go through even in that case however.

Figure 6: Financial constraint for the technological lender

both types of lenders and ask three questions. First, I study the choice of borrowers between asking financing to either of the lender. Second I show how the cost of intermediation for the traditional lender is affected by the entrance of the technological lender. Finally, I study the choice of the technological efficiency of the traditional lender under the assumption that both lenders compete for market share in the credit market.

Borrowers separation

When both types of lenders are available, the borrowers can choose between seeking financing to either of them. The choice of lender depends on the expected profits of the borrowers. As we already have derived the borrowers profits for both lenders in the previous sections, we can readily analyze the borrowers choices. Figure 7 illustrates the borrowers profits for both lenders.

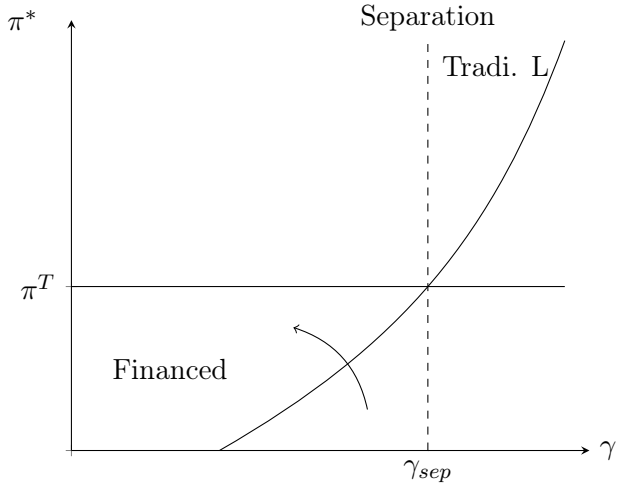

Figure 7: Borrowers choice of lender

As $\pi_{k,D}$ is increasing in γ_k, π^T is constant in γ_k, and $\eta^T > \eta$, there exists a threshold γ_{sep} such that $\pi_{k,D}(\gamma_{sep}) = \pi^T$. Depending on the important of soft information in assessing his creditworthiness γ_k, borrower k prefers to be financed by either the technological or traditional lender. This result is formally stated in the next proposition.

Proposition 2. *When both traditional and technological lenders coexist, there exists a threshold γ_{sep} such that*

- *If $\gamma_k < \gamma_{sep}$, borrower k prefers to seek financing from the technological lender.*
- *If $\gamma_k > \gamma_{sep}$, borrower k prefers to seek financing from the traditional lender.*

Proposition 2 has a very intuitive interpretation. The traditional lender has an comparative advantage over the technological lender as she can use soft information to assess the borrowers creditworthiness. This ability is valuable for borrowers whose soft information matters as it increases their expected profits. Therefore high γ_k borrowers prefer to seek financing from the traditional lender as her ability to discriminate them is higher. On the other hand borrowers whose soft information is not very important in determining their credit worthiness prefer to seek financing from the technological lender. Because the technological lender has a higher technological efficiency and because soft and hard information are substitutable to some extend, borrowers with lower γ_k are better off asking for financing from the technological lender. The entrance of the technological lender on the credit-market creates a *siphoning-off effect* on the traditional lender pools of borrowers, as illustrated by the curved arrow of figure (7).

Cost of traditional intermediation

Proposition 2 has consequences in terms of screening costs for the traditional lender. The optimal effort exerted by the traditional lender increases in γ_k since the marginal value of acquiring soft information is larger when γ_k is higher. As the screening cost is increasing in effort, it is costlier (but optimal) for the traditional lender to screen applicants with higher γ_k. Proposition 2 states that only borrowers with high γ_k stay with the traditional lender while the lower γ_k borrowers that used to be financed by the traditional lender shift away to the technological lender. As a result, the pool of borrowers faced by the traditional lender after the entrance of the technological lender exhibits a higher average γ_k. This leads to the next proposition.

Proposition 3. *The entrance of the technological lender increases the average screening cost of the traditional lender.*

It is interesting to interpret this result in light of Philippon (2016)'s paper. Philippon (2016) attempts to assess the potential impact of FinTech on the finance industry and documents the fact that while the financial sector has seen an entrance of new type of financial intermediaries, financial services remain expensive. While he argues that the cost of financial services can explain the entrance of new types of intermediaries, the result of proposition 3 suggests that the entrance of new intermediaries exhibiting lower costs may result in the increase in the cost of more traditional intermediaries, due to the fact that both lenders serve different types of borrowers, and the traditional lender is left with costly borrowers.

Choice of technological efficiency for the traditional lender

One assumption that has been made insofar is that the technological lender somehow has a superior ability to extract information from available historical lending data. In this section, I study the endogenous choice of the lender to modify her technological efficiency.

To do so, one needs to define the objective function of the lenders. In the current setup, both lenders are in perfect competition and earn zero-profit in expectation for each loan that is extended to borrowers. That section assumes that the lenders compete for market share in the credit market. One way to capture these incentives is to allow for the lenders to charge a fee ε for each borrower that receives financing. Note that ε can be thought of as being as small as desired, in line with the position of high competition between lenders. It turns out that the traditional lender faces a trade-off in her choice of technological efficiency η, as illustrated in figure 8.

Lemma (3) provides the basis for this trade-off. On the one hand, when the traditional lender increases her technology efficiency, she is able to attract borrowers for whom soft information is relatively less important. A shown in the subfigure 8b, increasing η shifts the separation threshold of the borrowers to the left, indicating that the traditional lender is able to capture a larger fraction of borrowers with average γ_k. On the other hand, increasing η results in more severe credit-rationing for the high γ_k borrowers. As seen on the figure, borrowers that lie between the red and the black curve for high values of γ_k do not have access to financing any longer if the traditional lender increases η.

The resulting incentives for the traditional lender to increase or decrease her technological ef-

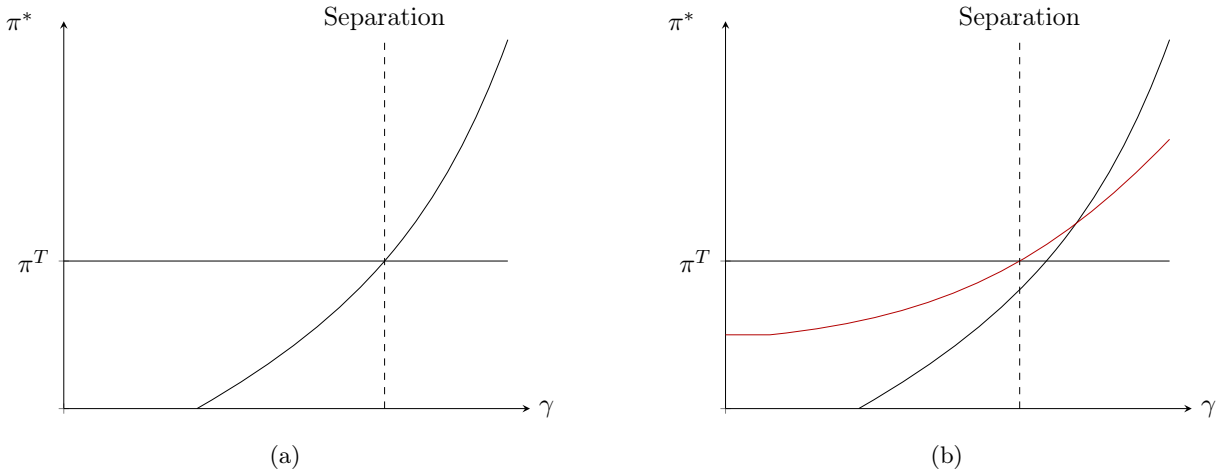

Figure 8: Effect of an increase in the technological efficiency (η) of the traditional lender.

ficiency when competing with a technological lender depends on the distribution of the borrowers population along γ_k. Recall that this distribution is denoted Γ. This leads to the next result.

Lemma 5. *The incentives for the traditional lender to choose her technological efficiency is as follows.*

- *If* Γ *is skewed towards high values of* γ_k*, the traditional lender prefers to lowers her technological efficiency* η*.*
- *If* Γ *is skewed towards low values of* γ_k*, the traditional lender prefers to increase her technological efficiency* η*.*

Lemma (5) characterizes the choice of the technological efficiency of the traditional lender as a function of the borrowers distribution along the γ_k dimension. In the case where there are many borrowers for whom soft information is important in determining their creditworthiness, it is more valuable for the traditional lender to decrease her technological efficiency because it decreases credit rationing for these borrowers thereby increasing the amount of borrowers to whom credit can be granted. Otherwise, if most borrowers do not rely much on soft information to be assessed, it becomes more profitable for the traditional lender to compete with the technological lender for these borrowers, and increase her technological efficiency.

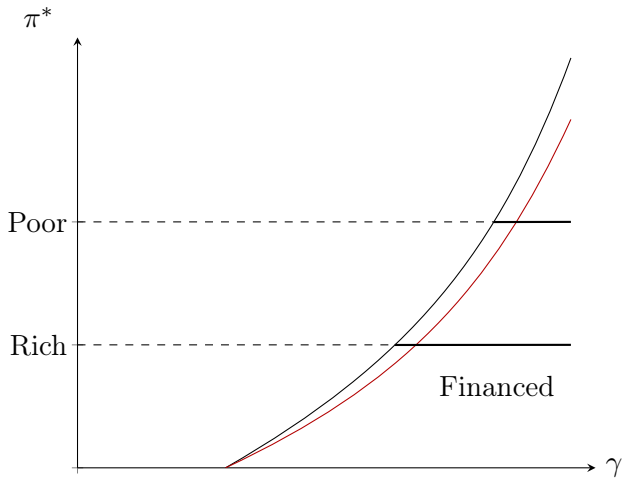

Figure 9: Dispersion effect

4 Dynamic implications of a data-based screening technology

In this section, I analyze further the dynamic impact of the use of a data-based screening technology. in the previous sections, I have analyzed the case in which the lender has access to an amount of historical data D informative of the borrower's type. It was then implicitly assumed that the same amount of information was available to the lender regardless of the borrower being assessed. However, the credit constraint of equation (7) makes clear that the presence of moral hazard constraints borrowers differently along their net worth. As illustrated in subfigure 5a, at the end of the first period lending, the lender gathers more data about rich borrowers compared to poor borrowers. As a result, the second period lending suffers from an imbalance in the structure of the data used in the screening process. This section aims to show that this dynamic effect between lending decision and data structure results in an increase of the effect of moral hazard. This effect is denoted the *dispersion effect*.

Let us first consider, as was previously assumed, that the lender has access to the same amount of informative data regarding borrowers of any net worth. In that case, the financing constraint is the one defined by equation (12) and the borrower's profits satisfy equation (11). The black curve of figure 9 plots the financial constraint in that case.

Let us now consider the fact that due to moral hazard, the first period lending results in an imbalanced historical lending data where richer borrowers are more represented than poorer borrowers.

In that case, the lender's optimal effort result in the following borrowers profits.

$$\pi_{k,D} = \pi_k + \frac{1}{2}\eta D\left(A_k\right)\left[qR\left(1-\gamma_k\right) + \alpha\eta D\left(A_k\right)\right]$$

where $D\left(A_k\right)$ captures the amount of informative historical data as a function of the borrower's net worth A_k.

The presence of moral hazard in the first period lending creates a positive relationship between the borrower's net worth and the amount of informative data available in the second period lending:

$$\frac{\partial D\left(A_k\right)}{A_k} > 0$$

The implication of that relationship is that poorer borrowers are now at a disadvantage compared to richer borrowers in the sense that the lender has a superior ability to screen richer borrowers using their hard information. As a consequence, compared to the case where the lender has the same amount of informative data to screen any borrower, the profit of a poor borrower is necessarily lower. Therefore, poor borrowers, already constrained from the presence of moral hazard, become even more so *compared* to richer borrower because the lender has *relatively* less informative data to assess their creditworthiness.

The red curve of figure 9 illustrates the impact of the data structure resulting from the first period lending on the financial constraint and leads to the result of the next proposition.

Proposition 4. ***Dispersion Effect***

The credit constraint resulting from the presence of moral hazard is increased *by the use of data in the screening process through the imbalance in historical lending data moral hazard creates.*

It is important to note that some conditions need to hold for the dispersion effect to arise. Recall that assessing the creditworthiness of the borrowers using historical lending data is performed through the use of predictive algorithm whereby the lender aims to find the mapping between the borrowers observable characteristics X_k and the outcome (success/default) Y_k. As is explained in more details in section A of the appendix, predictive algorithms are less efficient at assessing the creditworthiness of under-represented groups of borrowers under the assumption that the mapping between X_k and Y_k is not identical for all borrowers. In other words, it is necessary that the functional form (or the data generating process) f_k such that $Y_k = f_k\left(X_k\right)$ is not identical for all

borrowers.

5 Discussion and conclusion

The novelty of this paper is to investigate the effects of the use of data in the screening process, both through its interaction with the use of soft-information as well as dynamically. I first show that the use of data in the screening process lowers the lender's incentives to acquire costly soft information, which is detrimental to borrowers for whom soft information is important in determining their credit-worthiness. While the functional forms used in the current model may seem restrictive, they aim to capture important characteristics of the interaction between hard and soft information. Namely, they captures the fact the hard information can be a substitute to soft information, and that when lenders can use data, it becomes harder to improve the screening efficiency using soft information. Such results are related to the inattention literature, and one can view the lender as having a choice between using two sources of information, one denoted hard and another denoted soft. The fact that both sources are substitutable to some extend is important for the results of the paper. In fact, one can imagine that if both hard and soft information where to be complement, it might be that the traditional lender would always have a superior screening ability, ruling out the presence of purely technological lender, something we in fact observe.

I also show that the use of data in the screening process creates a dynamic effect arising from the interaction between lending and screening, denote the dispersion effect. Data-based technologies are fundamentally dependent on the structure of available observations which creates an inter-dependence between the lending decisions and the screening process. Such inter-dependence may lead to a screening technology that exhibits dispersion in its efficiency across borrowers. Namely, borrowers that are under-represented in the data are at a disadvantage in the screening process compared to other borrowers. It is important to note that the presence of moral hazard provides a fundamental endogenous reason as to why the data might be imbalanced, resulting in historical data where financially constrained borrowers become under-represented. This last channel is novel in that it does not rely on the fact that the data-based technology might exhibit biases (as in Fuster et al. (2018)), but arises because the predictions performances of such algorithms depend on the data that is provided to them. In that regard, it is close but different than a discrimination story whereby some individuals would be penalized because the technology provides biased estimates,

non-fundamentally justified by their differences.

In addition to shedding light on two possible channels through which the use of data may increase financial frictions, I find that when some lenders specialize in the use of data for screening borrowers (technological lender), the structure of the lending market changes in a way that affects the cost of intermediation of traditional lenders. This result hinges on the siphoning-off effect technological lenders have, leading only a certain type of borrowers to shift away from traditional lenders. Borrowers separate between the traditional and technological lender in such a way that the traditional lender is left with borrowers associated with higher screening costs – borrowers for whom soft information matters. This results in an increase of average screening costs borne by the traditional lender. While this result is not dependent on the dispersion effect of the data-based technology, it is strengthened when the technology exhibits such a characteristic.

Finally, I show that when a traditional lender competes with a technological lender, she may have incentives to limit the use of her data-based screening technology and specialize in her core competency that is the use of soft information. This is due to the fact that the traditional lender faces a trade-off whereby increasing its use of data allows to capture borrowers serviced by the technological lender, but also rations credit for borrowers whose soft information matters. Therefore, while the model assumes that traditional lenders are endowed with a less efficient technology, it provides reasons to believe that this can arise endogenously.

Overall the results of this paper are in line with the finding of Philippon (2016) showing that while we observe a rise in the presence of new types of lenders, the cost of intermediation does not seem to decrease, as would be expected by a typical competition argument. Instead, my model suggests that competition might incentivize traditional lenders to specialize in their soft information screening technology, even though it increases the average cost of intermediation.

References

Agarwal, Sumit and Itzhak Ben-David (2014). *Loan Prospecting and the Loss of Soft Information*. Working Paper 19945. National Bureau of Economic Research.

Allen, Franklin (1990). "The Market for Information and the Origin of Financial Intermediation". In: *Journal of Financial Intermediation* 1.1, pp. 3–30.

Berg, Tobias, Valentin Burg, et al. (2018). *On the Rise of FinTechs - Credit Scoring Using Digital Footprints*. Working Paper 24551. National Bureau of Economic Research.

Berg, Tobias, Manju Puri, and Jorg Rocholl (2013). *Loan Officer Incentives and the Limits of Hard Information*. Working Paper 19051. National Bureau of Economic Research.

— (2016). *Loan Officer Incentives, Internal Rating Models and Default Rates*. Working Paper ID 2022972. Rochester, NY: Social Science Research Network.

Berger, Allen N., W. Scott Frame, and Nathan H. Miller (2005). "Credit Scoring and the Availability, Price, and Risk of Small Business Credit". In: *Journal of Money, Credit and Banking* 37.2, pp. 191–222. JSTOR: 3838924.

Bolton, Patrick and David S. Scharfstein (1990). "A Theory of Predation Based on Agency Problems in Financial Contracting". In: *The American Economic Review* 80.1. ArticleType: research-article / Full publication date: Mar., 1990 / Copyright © 1990 American Economic Association, pp. 93–106. JSTOR: 2006736.

Boot, Arnoud W. A. (2000). "Relationship Banking: What Do We Know?" In: *Journal of Financial Intermediation* 9.1, pp. 7–25.

Brown, Martin, Tullio Jappelli, and Marco Pagano (2009). "Information Sharing and Credit: Firm-Level Evidence from Transition Countries". In: *Journal of Financial Intermediation* 18.2, pp. 151–172.

Cabrales, Antonio, Olivier Gossner, and Roberto Serrano (2013). "Entropy and the Value of Information for Investors". In: *American Economic Review* 103.1, pp. 360–377.

Chaigneau, Pierre, Alex Edmans, and Daniel Gottlieb (2017). *Does Improved Information Improve Incentives?* SSRN Scholarly Paper ID 2269380. Rochester, NY: Social Science Research Network.

Dewatripont, M. and E. Maskin (1995). "Credit and Efficiency in Centralized and Decentralized Economies". In: *The Review of Economic Studies* 62.4, pp. 541–555.

Diamond, Douglas (1991). "Monitoring and Reputation: The Choice between Bank Loans and Directly Placed Debt". In: *Journal of Political Economy* 99.4, pp. 689–721.

Diamond, Douglas W. (1984). "Financial Intermediation and Delegated Monitoring". In: *The Review of Economic Studies* 51.3, pp. 393–414. JSTOR: 2297430.

Ellis, Andrew (2018). "Foundations for Optimal Inattention". In: *Journal of Economic Theory* 173, pp. 56–94.

Fuster, Andreas et al. (2018). *Predictably Unequal? The Effects of Machine Learning on Credit Markets*. SSRN Scholarly Paper ID 3072038. Rochester, NY: Social Science Research Network.

Gale, Douglas and Martin Hellwig (1985). "Incentive-Compatible Debt Contracts: The One-Period Problem". In: *The Review of Economic Studies* 52.4, pp. 647–663. JSTOR: 2297737.

Inderst, Roman and Holger M. Mueller (2006). "Informed Lending and Security Design". In: *Journal of Finance* 61.5, pp. 2137–2162.

Innes, Robert D (1990). "Limited Liability and Incentive Contracting with Ex-Ante Action Choices". In: *Journal of Economic Theory* 52.1, pp. 45–67.

Jagtiani, Julapa and Catharine Lemieux (2018). *The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the Lendingclub Consumer Platform*. SSRN Scholarly Paper ID 3178461. Rochester, NY: Social Science Research Network.

Levitt, Steven D. and Christopher M. Snyder (1997). "Is No News Bad News? Information Transmission and the Role of "Early Warning" in the Principal-Agent Model". In: *The RAND Journal of Economics* 28.4, pp. 641–661. JSTOR: 2555780.

Liberti, Maria and Mitchell A. Petersen (2017). *Information: Hard and Soft*. Working Paper.

Matejka, Filip and Alisdair McKay (2015). "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model". In: *American Economic Review* 105.1, pp. 272–298.

Mayskaya, Tatiana (2017). *Dynamic Choice of Information Sources*. SSRN Scholarly Paper ID 2863605. Rochester, NY: Social Science Research Network.

Meyer, Margaret A. and Jeffrey Zwiebel (2007). *Learning and Self-Reinforcing Behavior*. Working Paper.

Meyer, Margaret a (1991). "Learning from Coarse Information: Biased Contests and Career Profiles". In: *The Review of Economic Studies* 58.1, pp. 15-41. JSTOR: 2298043.

Nimark, Kristoffer P. and Savitar Sundaresan (2018). *Inattention and Belief Polarization*. Working Paper.

Padilla, A. Jorge and Marco Pagano (1997). "Endogenous Communication Among Lenders and Entrepreneurial Incentives". In: *The Review of Financial Studies* 10.1, pp. 205–236.

— (2000). "Sharing Default Information as a Borrower Discipline Device". In: *European Economic Review* 44.10, pp. 1951–1980.

Pagano, Marco and Tullio Jappelli (1993). "Information Sharing in Credit Markets". In: *The Journal of Finance* 48.5, pp. 1693–1718.

Paravisini, Daniel and Antoinette Schoar (2015). *The Incentive Effect of Scores: Randomized Evidence from Credit Committees*. Working Paper 19303. National Bureau of Economic Research.

Petersen, Mitchell A. and Raghuram Rajan (1994). "The Benefits of Lending Relationships: Evidence from Small Business Data". In: *The Journal of Finance* 49.1, pp. 3–37.

— (2002). "Does Distance Still Matter? The Information Revolution in Small Business Lending". In: *The Journal of Finance* 57.6, pp. 2533–2570.

Petersen, Mitchell A. and Raghuram G. Rajan (1995). "The Effect of Credit Market Competition on Lending Relationships". In: *The Quarterly Journal of Economics* 110.2, pp. 407–443.

Philippon, Thomas (2016). *The FinTech Opportunity*. Working Paper 22476. National Bureau of Economic Research.

Rajan, Uday, Amit Seru, and Vikrant Vig (2015). "The Failure of Models That Predict Failure: Distance, Incentives, and Defaults". In: *Journal of Financial Economics* 115.2, pp. 237–260.

Shorrer, Ran I. (2015). *Entropy and the Value of Informatio for Investors: The Prior-Free Implications*. Working Paper.

Townsend, Robert M (1979). "Optimal Contracts and Competitive Markets with Costly State Verification". In: *Journal of Economic Theory* 21.2, pp. 265–293.

A The technological lender's problem

This section describes the statistical problem faced by a lender, and lays out results specific to the lender's learning problem driving for the subsequent economic findings. The goal of the technological lender is to use past lending data in order to assess the creditworthiness of future borrowers applying for credit. The general form of this learning problem is to approximate a data generation process for prediction purposes and can be stated as follows. Consider a data generating process mapping covariates x into a value y such that

$$y = f(x) + \varepsilon$$

where ε is a random variable with mean zero and variance σ^2. In the context of lending, the covariates x refer to all the observable characteristics of the borrower applying for credit and y is whether or not the borrower will be successful in the project he is undertaking or not. To learn the function f, the lender has access to historical observations about different borrowers that have been lent to in the past. Her goal is to find a function $\hat{f}$ that approximates the true f using this historical data. If she is able to find an $\hat{f}$ that is close to the true f, the lender is able to efficiently screen any *new* borrower applying for credit. I stress the fact that the lender is not learning about an unobservable state regarding a specific borrower through multiple interactions with that borrower over time, but is instead trying to find a general mapping between any borrower that would be true for any borrower, even if it is the very first time that borrower applies for credit. That being said, it does not prevent the mapping to depend on the history of the borrower, as the covariates x can include information such has past credit card records.

The learn the mapping f, the lender needs a way, an algorithm, also called a "learner", to find an approximation $\hat{f}$. Typically, one defines a loss function $L(\hat{y}, y)$ that captures the cost of predicting $\hat{y} = \hat{f}(x)$ when the true value is y. Because y is a non-deterministic function of x, the learner's objective is to find the *optimal prediction model*, which is the unique function f^* that minimizes the expected loss $E_y\left[L(\hat{y}, y)\right]$ for *every possible* value x. The subscript y denotes that the expectation is taken with respect to all possible values of y given x.[4] To find f^*, the learner has access to a set d of covariates X and values Y, typically called a training set, corresponding in our setting of the past lending data. Given d, a learner produces a model f_d which maps every x into a value

[4]This definition is independent of the distribution of ε, as the expectation takes into account the probability of y, so it remains valid even if, say, ε is skewed.

$f_d(x) = y_d$. Because every training set d is different, the *same* learner typically produces different prediction models f_d for each of them. A "good" learner is able to generate a prediction model whose predicted values are *close* to the optimal predictions *regardless* of the set d used to train it. In other other words, a good predictive model is one that performs well *out-of-sample*. Regardless of the sample used to train it, a good learner should generate a model that provides accurate predictions on different samples.

A learner might fail at finding such a model for two main reasons. First, it might generate models whose predictions differ depending on the training set d. We say that the learner exhibits *variance*. Second, it might generate models whose average predictions over multiple training sets d are far from the optimal prediction. We say that the learner exhibits *bias*. It turns out that it is difficult for a learner to exhibit both low variance and low bias, as variance increases and bias decreases in the model complexity. A model that predicts for every x the average of all y in a training set d has low variance and high bias, as the predictions do not vary greatly across training sets while being far from the true values. A model that predict for every x the value y associated with the closest observation in the training d has high variance and low bias, as the predictions are very dependent on the training set, but close to the true value when averaged across several training sets.[5] The main challenge of machine learning is to devise learners that optimally balance this bias-variance trade-off.[6] Two common techniques to solve this problem are regularization and resampling. Regularization aims at penalizing the use of complex functional forms so that simpler ones are preferred even if they do not minimize the expected loss in-sample. This is performed by directly adding a penalization term within the objective function. In resampling techniques – such as bootstrapping or bagging – several training sets are randomly sampled within the full training data and the final model uses the average predictions.

The intuition behind the economic findings of the paper is to consider how an imbalance in the training set can affect the performances of a statistical learner. To analyze this possibility, consider a training set d with N observations, and two partitions (d_1, d_2) of this training set with (N_1, N_2) observations. We denote (X_1, X_2) the sets of covariates and (Y_1, Y_2) their corresponding

[5] One might therefore wonder why machine learning algorithms do not just generate a large number of models with high variance and low bias using multiple training sets, and use the average prediction as the final model, ideally generating a model very close to the optimal one with low variance and low bias. This idea is the basis of many machine learning methods such as bootstrapping and bagging in order to reduce variance while preserving low bias.

[6] It is worth noting the bias-variance problem vanishes when one considers that a learner has access to an infinite amount of data, as a one can use an infinite amount of training sample to find the optimal model exhibiting low variance and low bias. Similarly, the bias-variance trade-off becomes irrelevant for parametric methods as the functional form determines the amount of bias and variance the learner exhibits.

values. (x_1, x_2) and (y_1, y_2) are elements of (X_1, X_2) and (Y_1, Y_2). We allow the two partitions to be generated by different data-generating process f_1 and f_2. The statistical learner, unaware of any subgroup, is trying to find the function f_d mapping x and y that minimizes the expectation of a given loss function L on the full training set d. The learner searches in the set of all possible functions $\{\hat{f}\}$ and solves

$$f_d = \arg\min_{\{\hat{f}\}} E_d \left[L\left(\hat{f}(x), y \right) \right] \tag{13}$$

Because the expectation operator is linear, equation (13) reads

$$f_d = \arg\min_{\hat{f}} \left\{ \frac{N_1}{N} E_{d_1} \left[L\left(\hat{f}\left(x_1\right), y_1 \right) \right] + \frac{N_2}{N} E_{d_2} \left[L\left(\hat{f}\left(x_2\right), y_2 \right) \right] \right\} \tag{14}$$

Equation (14) makes clear that minimizing the expected loss function on the whole sample is equivalent to minimizing the average of the expected loss functions on each subgroup, *weighted by their proportional sample size.* This observation leads to the result of Lemma 6.

Lemma 6. *Consider a training set d of* constant *size N containing two partitions (d_1, d_2) of sizes (N_1, N_2) following data generating processes (f_1, f_2).*

If $f_1 \neq f_2$, the predictive performance of a learner increases on the first group and decreases on the second group as $\frac{N_1}{N_2}$ increases.

Lemma 6 has several intuitive interpretations. First, in terms of information. If one subsample increases in size, the learner has relatively more information about that subsample and generates a model that performs better on that subgroup. Second, equation (14) can be understood as follows. A learner minimizing the total expected loss does a better job by minimizing the expected loss on the larger group rather than the one on the smaller group as the former carries a larger weight in the objective function. Lemma 6 states that these intuitions are correct if the two groups are generated by different data generating processes. Intuitively, the results are correct if the information available on one group is not fully informative for predictions on the second group. If both groups follow exactly the same data generating processes, that would not be true.

A very important point is that this lemma applies to *non-parametric* methods such as machine learning techniques that are recently getting widely adopted. Parametric techniques are not subject to the bias-variance trade off described above as they use a fixed model complexity. As instance, an OLS regression find only the best *linear* fit of the data. While both methods are aiming at

minimizing a loss function, the problems are fundamentally different as a non-parametric method has the ability to modify the complexity of the model used to fit the data, while a parametric method does not.

In the rest of the paper, we will assume that the conditions specified in Lemma 6 are satisfied on the problem considered. While it may seem arbitrary to assume so, one needs only to assume that the problem faced by the lender – assess the creditworthiness of borrowers – is sufficiently complex and that the probability that all borrowers are affected in the exact same way by fundamentals is sufficiently small. In fact such an assumption is perfectly in line with the observed rise in the use of machine learning techniques as they are specifically designed to find solutions in a complex world where the statistician is not able or prefers not to specify functional forms, leaving that task to the learning algorithm.

B Proofs

Lemma 1

First, note that $p(e_k)$ is the probability of success of the borrower conditional on receiving the signal $s_{0,k} = 1$ and conditional on the borrower behaving. Using Bayesian updating, we have that

$$p(e_k) = q\tau_0(e_k) = q\left(\frac{1}{2} + \gamma_k e_k\right)$$

Given A_i, the lender exerts effort e_{eq} that satisfy the following first order condition (where subscripts denote derivatives)

$$\frac{1}{2} p_e(e_{eq}) R = C_e(e_{eq})$$

Using the functional forms, we have

$$e_{eq} = \frac{1}{2\alpha} qR\gamma_k$$

The borrower chooses the optimal A_i^* that maximizes his profits given that the lender will exert

the effort derived above

$$\begin{aligned} &\max_{A_i} \frac{1}{2}\left[B - A_i\right] + A_k \\ \text{s.t. } &e_{eq} = \frac{1}{2} qR\gamma_k \\ &\frac{1}{2}\left[p\left(e_{eq}\right)R - 1\right] - c\left(e_{eq}\right) \geq \frac{1}{2}\left[B - A_i\right] \geq 0 \end{aligned}$$

It is clear that the borrower chooses the maximum value of A_i such that

$$\frac{1}{2}\left[B - A_i^*\right] = \frac{1}{2}\left[p\left(e_{eq}\right)R - 1\right] - c\left(e_{eq}\right)$$

The borrower's profit when financed by the traditional lender is therefore

$$\pi_k = \frac{1}{2}\left[p\left(e_{eq}\tau_{1,k}\left(e_{eq}\right)\right)R - 1\right] - c\left(e_{eq}\right)$$

Replacing $p\left(e_{eq}\right)$ and $c\left(e_{eq}\right)$ with the lender's effort derived above, we obtain

$$\begin{aligned} \pi_k &= \frac{1}{2}\left[qR\tau_{0,k}\left(e_{eq}\right) - 1\right] - \frac{e_{eq}^2}{2} \\ &= \frac{1}{2}\left[\frac{1}{2}qR - 1\right] + \frac{1}{\alpha}\left(\frac{1}{2}qR\gamma_k\right)^2 - \frac{1}{2\alpha}\left(\frac{1}{2}qR\gamma_k\right)^2 \\ &= \frac{1}{2}\left[\frac{1}{2}qR - 1\right] + \frac{1}{2\alpha}\left(\frac{1}{2}qR\gamma_k\right)^2 \end{aligned}$$

π_k is strictly increasing and convex in γ_k as

$$\frac{\partial \pi_k}{\partial \gamma_k} = \frac{1}{\alpha}\left(\frac{1}{2}qR\right)^2 \gamma_k = \frac{1}{2\alpha} qRe_{eq}$$

Lemma 2

When the traditional lender has access to historical lending data, the borrower's problem stays identical, only the functional forms change. The effort exerted by the lender satisfy

$$\frac{1}{2} p_e\left(e_{eq}\right)R = C_e\left(e_{eq}\right)$$

which gives

$$\frac{1}{2}qR\gamma_k = \alpha\left(2e_{eq} + \eta D\right)$$
$$e_{eq} = \frac{1}{2\alpha}qR\gamma_k - \eta D$$

And the borrower's profits are

$$\begin{aligned}
\pi_{k,D} &= \frac{1}{2}\left[p\left(e_{eq}\right)R - 1\right] - c\left(e_{eq}\right) \\
&= \frac{1}{2}\left[qR\tau_{1,k}\left(e_{eq}\right) - 1\right] - \alpha\left(\frac{e_{eq}^2}{2} + \eta D e_{eq}\right) \\
&= \frac{1}{2}\left[qR\left(\frac{1}{2} + \gamma_k e_{eq} + \eta D\right) - 1\right] - \alpha\left(\frac{e_{eq}^2}{2} + \eta D e_{eq}\right) \\
&= \frac{1}{2}\left[\frac{1}{2}qR - 1\right] + \frac{1}{2}qR\left(\gamma_k\left(\frac{1}{2\alpha}qR\gamma_k - \eta D\right) + \eta D\right) - \alpha\left(\frac{\left(\frac{1}{2\alpha}qR\gamma_k - \eta D\right)^2}{2} + \eta D\left(\frac{1}{2\alpha}qR\gamma_k - \eta D\right)\right) \\
&= \frac{1}{2}\left[\frac{1}{2}qR - 1\right] + \frac{1}{\alpha}\left(\frac{1}{2}qR\gamma_k\right)^2 - \frac{1}{2}qR\gamma_k\eta D + \frac{1}{2}qR\eta D - \frac{1}{2\alpha}\left(\frac{1}{2}qR\gamma_k\right)^2 - \frac{\alpha}{2}\left(\eta D\right)^2 + \frac{1}{2}qR\gamma_k\eta D - \frac{1}{2}qR\gamma_k\eta D \\
&= \frac{1}{2}\left[\frac{1}{2}qR - 1\right] + \frac{1}{2\alpha}\left(\frac{1}{2}qR\gamma_k\right)^2 + \frac{1}{2}qR\eta D\left(1 - \gamma_k\right) + \frac{\alpha}{2}\left(\eta D\right)^2 \\
&= \pi_k + \frac{1}{2}\eta D\left[qR\left(1 - \gamma_k\right) + \alpha\eta D\right]
\end{aligned}$$

As in the case without data, π_k is increasing and convex in γ_k as

$$\frac{\partial \pi_k}{\partial \gamma_k} = \frac{1}{\alpha}\left(\frac{1}{2}qR\right)^2\gamma_k - \frac{1}{2}qR\eta D = \frac{1}{2}qRe_{eq}$$

However, the slope is lower when the lender uses data as effort decreases in ηD.

Lemma 3

The borrower's profits when the lender does not have access to data is π_k and the profits when the lender does not have access to data is $\pi_k + \frac{1}{2}\eta D\left[qR\left(1 - \gamma_k\right) + \alpha\eta D\right]$. The borrower has higher

expected profits if the lender does not have access to data if

$$\begin{aligned} qR\left(1-\gamma_k\right)+\alpha\eta D &< 0 \\ qR\left(\gamma_k-1\right) &> \alpha\eta D \\ \gamma_k &> 1+\alpha\frac{\eta D}{qR}\equiv\tilde{\gamma} \end{aligned}$$

Note that the optimal signal precision when the lender uses data is

$$\tau_{1,k}=\frac{1}{2}+\frac{1}{2\alpha}qR\gamma_k^2+\eta D\left(1-\gamma_k\right)$$

at $\gamma_k=1$, the signal precision is $\frac{1}{2}+\frac{1}{2\alpha}qR$ which is lower than 1 if α is high enough. Also, at $\gamma_k=\tilde{\gamma}$, the precision is $\frac{1}{2}+\frac{1}{2\alpha}qR-\frac{\alpha}{2}\frac{(\eta D)^2}{qR}$.