

Payments and Mechanism Design*

Thorsten Koepl

Queen's University

Kingston, Ontario

thor@qed.econ.queensu.ca

Cyril Monnet

European Central Bank

Frankfurt, Germany

cyril.monnet@ecb.int

Ted Temzelides

University of Pittsburgh

Pittsburgh, PA 15260

tedt@pitt.edu

First version: 2004

December 13, 2006

*We thank audiences at CEMFI, The Central Bank of Portugal, Cornell University, the Federal Reserve Banks of Cleveland and New York, the European Central Bank, and Aleks Berentsen for comments and suggestions on an earlier draft. The views expressed in the paper are not necessarily those of the ECB. The first and third author gratefully acknowledge support from SSHRC and the NSF through grants 410-2006-0481 and SES 0517862 respectively.

Abstract

We use *mechanism design* in order to study efficient arrangements when the ability of agents to perform certain welfare-improving transactions is subject to random and unobservable shocks. We study implementation via a *payment system* that involves assigning balances to participants and optimally adjusting these balances given their histories of transactions. Our analysis has several implications for the design of payment systems. Efficiency requires that, in order to overcome informational frictions, transactions that are subject to high monitoring costs need to be subsidized. Optimal settlement frequency should balance the fixed cost of settlement against the costs arising from providing intertemporal incentives. Settlement costs must be borne by those participants for whom certain participation constraints are slack.

1 Introduction

One of the features of the economy that Walrasian models abstract concerns the institutions through which transactions for goods and services take place: the *payment systems (PS)*. This results in a need for new models in order to study efficient ways to design transactions and the corresponding properties of PS. This paper builds such a model using mechanism design. Our approach involves two main ingredients. First, it is explicitly dynamic, emphasizing the role of *intertemporal incentives*. This is important since actual payment systems almost always involve repeated interactions between the system and its participants. A second distinguishing feature of our analysis is that, unlike most of the literature in mechanism design, since we model the entire payment system, we must take into consideration *general equilibrium* effects.

Our approach emphasizes the role of private information in a way related to the dynamic contracting literature.¹ We use a dynamic framework since some of the questions that optimal payment system design poses are inherently dynamic and, therefore, hard to capture within the existing literature, which is almost exclusively static.² For example, the PS might need to make use of intertemporal incentives in order to explore the agents' willingness to participate. In addition, PS design is subject to a private information problem to the extent that participants' abilities to perform certain transactions are not directly observable. For example, within a retail PS, the ability of a consumer to eventually make a credit-card payment might not be observable. Similarly, within a wholesale system, a bank might have private information about its (in)ability to meet certain obligations.

We employ a version of the model of exchange developed by Kiyotaki and Wright (KW 1989, 1993) in order to study monetary exchange. In contrast to the monetary theory approach, however, our model involves a cash-less environment. The KW model is appropriate for our objective for several

¹See, for example, Green (1987). Other classic references include Spear and Srivastava (1987) and Atkeson and Lucas (1993). Our analysis also relates to recent work by Kocherlakota (2005), who extends the model of Mirrlees (1971). The payment system in our model plays an analogous role to that of the tax authority in Kocherlakota (2005): it explores intertemporal incentives in order to implement efficient outcomes under private information.

²See Kahn and Roberds (1998, 2001) for two prominent papers in this spirit. Kahn (2006) provides an excellent summary of the current literature and outlines some of the main open questions in payment systems research.

reasons. First, it involves an explicit role for transactions, and it naturally incorporates frictions such as private information and lack of commitment. Second, the random matching shocks that agents are subject to in this model are a tractable way of modelling random needs for liquidity. Such needs are an important feature of actual PS, as they capture the fact that PS participants need to transfer resources to their participants, and that these needs are random. Third, the model is consistent with the fact that actual transactions are bilateral and, frequently, subject to private information. Finally, this setup naturally lends itself to mechanism design.

Our methodological contribution consists of developing a model that we hope will prove useful in the design of actual PS. One case of particular interest involves when the private information friction is more severe in some transactions than in others. For example, in some wholesale PS, banks interact mostly, but not exclusively, through a local network of other banks. Such networks might have detailed information about their participants. The same information, however, might not be available to the system if a bank needs to transact with a bank outside its network.³ We will assume that each transaction involves a “producer” and a “consumer.” This is only an interpretation, as what is important is that a transaction involves a cost for one party (the producer) and a gain for the other party (the consumer), and there are several ways to rationalize such costs and gains. In the presence of private information about the ability of agents to produce/consume, in order for the efficient volume of transactions to take place, the rules of the game (the PS) must provide the right incentives to the players (the participants in the PS). The study of the optimal way to provide such incentives is the main focus of this paper. For example, if some of transactions are not subject to a private information problem, say because the ability of the agents to transact is monitored, the PS might adjust the terms of trade in such transactions in order to make certain non-monitored transactions to incentive compatible. We model this feature by assuming that with probability α a transaction is monitored, in which case the ability of the two parties to perform the transaction is observable to the PS. With probability $1 - \alpha$ a transaction is not monitored; i.e., the ability of the parties to transact is not observable.⁴

³Alternatively, in a retail PS, say, an American consumer’s credit card history might be readily available in the US but not necessarily in Europe. In wholesale PS, in which participants are themselves banks, such networks might be run by large correspondent banks.

⁴For an alternative interpretation we can suppose that the set of agents is divided

We use the model to study two types of issues of relevance to actual PS design. The first involves the structure of balance adjustments within the trading cycle. The second involves the optimal frequency of settlement and the study of efficient ways to share settlement costs among PS participants.⁵ We demonstrate that, in the presence of a positive volume of non-monitored transactions, an optimal PS will need to “tax” monitored transactions in order to “subsidize” the former at the expense of the latter. When settlement is subject to fixed operational costs, optimal settlement frequency balances such costs against the constraints arising from the need to provide intertemporal incentives. Finally, we argue that settlement costs must be borne by those participants for whom certain participation constraints are slack. Thus, our model provides a novel prescription for the optimal sharing of costs associated with the operation of a PS (an example of which could be *interchange fees* for credit-card use), across PS participants. More precisely, our analysis suggests that either such fees should be entirely borne by non-consumers, or that their efficient allocation is indeterminate.

The paper proceeds as follows. Section 2 introduces the model. Sections 3 and 4 discuss the two applications to optimal balance adjustments and to the optimal frequency of settlement, respectively. Section 5 offers a brief conclusion and discusses some of the many possible directions for future research.

2 A Dynamic Model of Payments

Time, t , is discrete and measured over the positive integers. There is a $[0, 1]$ continuum of infinitely lived agents. The common discount factor is $\beta \in (0, 1)$. We assume a periodic pattern of length $n + 1$, in which each *transactions stage*, consisting of n bilateral rounds, is followed by a centralized round which we shall term the *settlement stage*.⁶ We describe the

into two symmetric *networks*. Each agent needs to transact within his network with probability α , and with another participant outside his network with probability $1 - \alpha$. In this interpretation, within-network transactions are monitored, while during inter-network transactions the ability to transact is private information of each of the trading partners.

⁵In a sequel paper, Koepl, Monnet, and Temzelides (KMT, 2006), we demonstrate that settlement is a necessary feature of optimal payment systems.

⁶The continuum assumption precludes aggregate risk. Issues related to optimal PS design in the presence of aggregate risk are of interest, but beyond the scope of this paper. Lagos and Wright (2006) introduced related periodic trading patterns in monetary models.

transactions stage and the settlement stage in turns.

During the transactions stage, agents are matched bilaterally. Randomness is captured by assuming that in each period an agent needs to transact with the agent he is matched with as a producer or as consumer, each with probability γ . Thus, in each period during the transaction stage, an agent is in a *trade meeting* with probability 2γ . Agents cannot pre-commit to produce in such meetings. With probability $(1 - 2\gamma)$, an agent is in a no-trade meeting. Production in the transactions stage is perfectly divisible and the produced goods are non-storable. Producing q units implies dis-utility $-e(q)$, while consumption of q units gives utility $u(q)$. We assume that $e'(q) > 0$, $e''(q) \geq 0$, $\lim_{q \rightarrow 0} e'(q) = 0$, and $\lim_{q \rightarrow \infty} e'(q) = \infty$. In addition, $u'(q) > 0$, $u''(q) \leq 0$, $\lim_{q \rightarrow 0} u'(q) = \infty$, and $\lim_{q \rightarrow \infty} u'(q) = 0$. Thus, there exists a unique q^* such that $u'(q^*) = e'(q^*)$. The quantity q^* gives the efficient level of output, in the sense that it uniquely maximizes the joint surplus created in a transaction. Since we will concentrate on this quantity, in order to simplify notation, we will hence denote $u(q^*)$ by u , and $e(q^*)$ by e .

The information structure during the transactions stage is as follows. Whether a trade meeting has occurred is always observable to the two agents in the meeting. This is also publicly observable with probability α . We interpret such meetings as “within-network” meetings that are subject to monitoring. With probability $1 - \alpha$ whether the meeting is a trade meeting or not is not publicly observable. We interpret such meetings as “inter-network” meetings that are not subject to monitoring. While the opportunity to trade is not observable in non-monitored meetings, we assume that, should they take place, production and consumption are always verifiable.

During the settlement stage, each agent can produce and consume a general non-storable good. No other commodity can be produced or consumed during such stages. Producing ℓ units of the general good implies dis-utility $-\ell$, while consuming ℓ units gives utility ℓ . The settlement stage is frictionless, in the sense that, just as in Walrasian markets, agents interact in a centralized fashion and there are no informational frictions.

Starting with the transactions stage, the timing of events within each period is as follows. First, agents are bilaterally matched. In a monitored meeting, the producer is instructed to produce a quantity, q , for the consumer. The system then makes adjustments to the histories of the two parties in order to include the most recent transaction (more about this later). The transaction protocol we will assume for non-monitored meetings obtains the following interpretation. Each participant in the system has access to both

a “card” and a “card-reading machine.” Consumers can choose to identify themselves to the PS by having their card read by the producer’s machine. In that case, the system becomes aware that the two agents are in a match, as well as of the identities of the producer and that of the consumer. Finally, after production and consumption take place, the PS update the agents’ histories accordingly.

More formally, in non-monitored meetings, the two agents in a match each choose a number in $\{0, 1\}$. If either chooses 0, nothing further happens (this, for example, will always be the case in no-trade meetings) and the system treats both agents as if they are in a no-trade meeting. If both choose 1, the potential producer commits to produce q for the potential consumer. In addition, in that case the agents’ identities become known to the system who can make adjustments on how the respective agents will be treated in the future. Finally, consumption takes place. No such reporting needs to occur in monitored meetings, since in such meetings the PS observes everything.

We will concentrate on implementing the full information first-best allocation, in which the efficient level of production, q^* , takes place in *all* trade meetings, both monitored and non-monitored.⁷ We term this allocation (ex ante) *efficient*. The difficulty, of course, lies in that the system cannot always verify whether a trade meeting has taken place. For example, consider a distinguished agent who, say, for the k -th time in a row, participates in a non-monitored meeting and reports that he could not produce since he had k consecutive non-trade meetings (say, he reports “0” k consecutive times). Given the information structure, the system can verify that the agent did not produce in any of the last k meetings. What the system cannot verify, however, is whether the agent had an opportunity to produce and simply declined (by refusing to identify himself as a producer), or whether he did not encounter any trade meetings (an event of probability $[(1 - \alpha)(1 - \gamma)]^k$).

An allocation specifies, possibly as a function of the entire history, the quantity produced (consumed) in each transaction as well as in the settlement round. An allocation is *incentive feasible (IFA)* if it respects certain incentive, ex ante participation, and resource feasibility constraints. Throughout, we will restrict attention to IFAs that are stationary and symmetric. We want to characterize conditions under which the *efficient* allocation is an IFA.

⁷Since the utility from consuming any amount of the general commodity equals the resulting disutility to the producer, the efficient amount of the general good produced in each settlement round is indeterminate.

Clearly, since this allocation is efficient even in the absence of informational frictions, it is the best allocation that can possibly be decentralized in this environment. Next, we demonstrate how this can be accomplished via a PS. Decentralization involves assigning balances to individual participants and specifying rules for how these balances are updated in order to satisfy incentive and participation constraints.

2.1 Decentralization

Here we discuss how the efficient allocation can be decentralized. We will assume that the PS assigns *balances* to participants. In addition, the PS specifies rules for how the balances are updated given the histories of reports regarding transactions. As in Koepl, Monnet, and Temzelides (2006), we will decentralize the settlement stage as a periodic competitive market. More precisely, during settlement participants can trade their balances for the general good. Those that are “low” can increase their balances by producing, while those with high balances end up as consumers. The price, p , at which balances are traded, is determined by market clearing conditions. We make two additional assumptions. The first is that the PS has the ability to permanently exclude from the economy all agents that refuse to participate in settlement. Second, we assume that operating the settlement round is associated with an aggregate (average) resource cost $\delta > 0$. The presence of this cost will lead to an optimal settlement frequency in our model. First, we will assume that this cost is borne equally by all agents through production in the settlement stage, independent of the type of transactions that they need to settle. Later we will investigate the most efficient way to finance δ across PS participants.

In any given period, t , an agent’s history is summarized by two variables (d_t, h_t) . The real number d_t indicates the agent’s *balance* in period t . Intuitively, this balance summarizes a participant’s transaction history in non-monitored transactions. In addition, the PS keeps a record of all monitored transactions. The indicator variable $h_t \in \{0, 1\}$ summarizes the information on whether an agent has “defaulted” in the past ($h = 0$), either by not producing in a monitored transaction or by not participating in a settlement stage. Such an agent is permanently excluded from the economy and receives a future payoff that is normalized to 0.⁸

⁸In the card interpretation of the previous section, the agent’s card is de-activated.

In each of the first n periods of the transactions stage, participants engage in bilateral transactions (both monitored and non-monitored). In what follows, we analyze a generic period, t , and work backwards, first considering the participant's problem in the settlement stage, and then moving on to the transactions stage.

Let $V(d, h, p)$ denote the value function of a participant that exits the transaction stage with balance d , given that the anticipated price in the following settlement stage is p . Let $v(\hat{d}, \hat{h}, \Psi)$ denote the value of a participant who exits the settlement stage with balance $\hat{d}$, given that the resulting distribution of balances is denoted by Ψ .⁹ Given p and Ψ , participants at the beginning of the settlement stage solve the following:

$$V(d, h, p) = \max_{\hat{h}} \left\{ \hat{h} \left\{ -\delta + \max_{\ell, \hat{d}} [-\ell + \beta E v(\hat{d}, \hat{h}, \Psi) \text{ s.t. } p\hat{d} = pd + \ell] \right\} + (1 - \hat{h})0 \right\}. \quad (1)$$

Thus, at the end of each transactions stage, each participant chooses whether to settle at the resulting price, p , or default. We now turn to the problem faced by the participants during each round of the transactions stage. Recall that during this stage, participants that are in non-monitored meetings make reports to the PS about the type of the meeting that they are in. The PS makes a recommendation for the size of the transaction that it wishes to take place in each transaction. In addition, it makes adjustments balances depending on the participants' reports and on the resulting transactions. Both the transaction size and the adjustments also depend on the participants' histories as summarized by (d, h) and, in principle, on the distribution of balances, Ψ . Not taking into account balances, a participant can be in a consumption, a production, or a no-trade meeting. In addition, the meeting is either monitored or non-monitored. This results in six possible cases for each transactions round. The vector of policy rules $\{\bar{L}_t, \bar{K}_t, \bar{B}_t, \bar{q}_t; L_t, K_t, B_t, q_t\}$ determines the respective balance adjustments and the prescribed transaction size in a trade meeting within the network, $\bar{q}_t$, or outside the network, q_t . More precisely, $L_t(K_t)$ is the adjustment for a participant who consumes (produces), while B_t is the adjustment for a participant who does not transact in a non-monitored meeting. The variables $\bar{L}_t$, $\bar{K}_t$, and $\bar{B}_t$ are defined analogously for monitored meetings. Recall that balances are represented by

Since we concentrate on the efficient allocation where q^* is produced in all trade meetings and in which no agent defaults, we will ignore the index h in what follows.

⁹We describe $E[v(\hat{d}, \hat{h}, \Psi)]$ in detail below.

real numbers not restricted in sign, while production of goods during trade meetings is restricted to be positive. After each transactions round, knowing their new balances, participants choose whether to stay in the economy and enter the next round. Similarly, at the end of the transactions stage, agents choose whether to settle or default.

We will concentrate on arrangements that satisfy certain incentive and participation constraints. In each round, t , of the transaction stage, *incentive compatibility (IC) constraints* require that the following inequalities hold for non-monitored meetings:

$$u(q) + (d_t + L_t, p) \geq V(d_t + B_t, p), \quad (2)$$

$$-e(q) + V(d_t + K_t, p) \geq V(d_t + B_t, p). \quad (3)$$

The two constraints ensure that participants truthfully report consumption (production) opportunities during non-monitored meetings. Notice that no such truth-telling constraints are needed for monitored transactions. In addition, *participation (PC) constraints* require that producers, consumers, and participants in no-trade meetings, respectively, are better off staying in the system; i.e.,

$$u(q) + V(d_t + L_t, p) \geq 0, \quad (4)$$

$$-e(q) + V(d_t + K_t, p) \geq 0, \quad (5)$$

$$V(d_t + B_t, p) \geq 0. \quad (6)$$

Similarly, for monitored transactions we need

$$u(q) + V(d_t + \bar{L}_t, p) \geq 0, \quad (7)$$

$$-e(q) + V(d_t + \bar{K}_t, p) \geq 0, \quad (8)$$

$$V(d_t + \bar{B}_t, p) \geq 0. \quad (9)$$

We are now ready to formally define a Payment System.

Definition: A *Payment System* is an array $S = \{\bar{L}_t, \bar{K}_t, \bar{B}_t, \bar{q}_t; L_t, K_t, B_t, q_t\}$. S is *incentive feasible* if it satisfies the incentive and participation constraints. S is *simple* if balance adjustments do not depend on the participants' current balances, and it is *repeated* if (adjusted for discounting) balance adjustments do not depend on the time, t . An incentive feasible S is *optimal* if it decentralizes the efficient allocation.

In what follows, we restrict attention *Simple Repeated Payments Systems* (SRPS). This is mainly for simplicity. Clearly, this restriction is without loss of generality in the cases where a SRPS decentralizes the efficient allocation. Given any incentive feasible PS, the value function at the end of the settlement stage, $E[v(d, \Psi)]$, for a participant with balance d is given by

$$\begin{aligned}
& E[v_p(d, \Psi)] \\
= & \alpha [\gamma[u(q) + V(d + \bar{L}(d, d', p))] + \gamma[-e(q) + V(d + \bar{K}(d, d', p))] \\
& + (1 - 2\gamma)[V(d + \bar{B}(d, d', p))] + (1 - \alpha) \int_{d'} \{\gamma[u(q) + V(d + L(d, d', p))] \\
& + \gamma[e(q) + V(d + K(d, d', p))] + (1 - 2\gamma)V(d + B(d, d', p))\} d\Psi]. \quad (10)
\end{aligned}$$

Note that the balance adjustments are in general functions of the participant's own balance, d , as well as of the balance of his trading partner, d' , h' . We next investigate properties of optimal SRPS. Our first application concerns the structure of optimal balance adjustments within such systems in the case where some transactions are subject to monitoring, while others are not. We then discuss issues related to the optimal frequency and financing of settlement.

3 Optimal Balance Adjustments

In the previous sections we introduced an environment in which certain transactions are subject to a private information problem and discussed how a payment system can decentralize efficient allocations given incentive and resource constraints. In this section we use this setup in order to study the nature of optimal balance adjustment by the PS. We are particularly interested in investigating whether the PS can use resources from monitored transactions in order to manipulate binding incentive constraints during non-monitored transactions. In what follows, we will restrict attention to simple, repeated PS. However, it can be demonstrated that our results apply more generally to all simple PS, even those that are non-repeated.

The intuition behind the interaction between monitored and non-monitored transactions is simple. An efficient PS will use balance adjustments that effectively “tax” monitored transactions and “subsidize” certain non-monitored transactions in order to support the efficient overall volume of transactions.

For simplicity, since we do not deal with questions related to the frequency of settlement, for the remainder of this section we will assume that $\delta = 0$.

First, consider the benchmark case where $\alpha = 1$. In that case, all transactions are monitored and a “gift-giving” regime can be supported as long as $\beta u \geq e$. In other words, the efficient volume of transactions can be supported through the threat of exclusion if a participant refuses to comply. In that case, settlement is not a necessary feature of an optimal PS. At the other extreme, suppose that $\alpha = 0$. In that case, all transactions are non-monitored and frequent settlement is a necessary feature of an optimal PS.¹⁰ We summarize this discussion in the following the following.

Proposition 1 *Suppose $\alpha = 1$ and $n = \infty$. Then an optimal PS exists if and only if $\beta u \geq e$. Suppose $\alpha = 0$. An optimal PS exists if and only if $\beta^n u \geq e$.*

In other words, if $\beta^n u < e$, the existence of some monitored transactions is necessary in order to support the efficient outcome. A straightforward Corollary of the above Proposition is that when $n > 1$ and $\alpha \in (0, 1)$, for an optimal PS to exist, the balance adjustments associated with monitored transactions must be such that the incentive constraints in the non-monitored transactions are somehow relaxed. Our next step is to allow for such adjustments and show that the use of “within-network incentives” in order to subsidize some non-monitored transactions can decentralize the efficient allocation for a wider range of parameter values.¹¹ More precisely, we have the following.

Proposition 2 *An optimal PS exists if and only if $\beta^n u \geq \beta^{n-1}e + (1 - \alpha)(1 - \beta^{n-1})e$.*

Proof. Market clearing requires that

$$\alpha [\gamma p_t \bar{K} + \gamma p_t \bar{L} + (1 - 2\gamma) p_t \bar{B}] + (1 - \alpha) [\gamma p_t K + \gamma p_t L + (1 - 2\gamma) p_t B] = 0. \quad (11)$$

The PS must satisfy both IC and PC; i.e.,

$$p_t K - e \geq p_t B, \quad (12)$$

$$u + p_t L \geq p_t B, \quad (13)$$

$$p_t L + \frac{\beta}{1 - \beta} \gamma (u - e) + p_t X_{\min} \left[\frac{1}{\beta} + \cdots + \frac{1}{\beta^{n-1}} \right] \geq 0. \quad (14)$$

¹⁰See KMT (2006) for details.

¹¹It turns out that in the case where $n = 1$, the efficient allocation is incentive feasible for any $\alpha \in [0, 1]$, if and only if $\beta u \geq e$.

Hence, it is necessary to set $\bar{K}$, $\bar{L}$, and $\bar{B}$ strictly negative in order to cross-subsidize non-monitored transactions. From the PC of having n -times the minimum adjustment, $X_{\min}$, we obtain

$$p_t \bar{K} - e + \frac{\beta}{1-\beta} \gamma(u-e) + p_t X_{\min} \left[\frac{1}{\beta} + \cdots + \frac{1}{\beta^{n-1}} \right] \geq 0, \quad (15)$$

$$p_t \bar{L} + \frac{\beta}{1-\beta} \gamma(u-e) + p_t X_{\min} \left[\frac{1}{\beta} + \cdots + \frac{1}{\beta^{n-1}} \right] \geq 0, \quad (16)$$

$$p_t \bar{B} + \frac{\beta}{1-\beta} \gamma(u-e) + p_t X_{\min} \left[\frac{1}{\beta} + \cdots + \frac{1}{\beta^{n-1}} \right] \geq 0. \quad (17)$$

Note that these are the only constraints that need to be satisfied in monitored transactions. Multiplying the first and the second constraint by γ and the third constraint by $(1-2\gamma)$ and adding, we obtain

$$\gamma p_t \bar{K} + \gamma p_t \bar{L} + (1-2\gamma) p_t \bar{B} \geq \gamma e - \frac{\beta}{1-\beta} \gamma(u-e) - p_t X_{\min} \left[\frac{1}{\beta} + \cdots + \frac{1}{\beta^{n-1}} \right]. \quad (18)$$

Hence for non-monitored transactions, we have

$$\gamma p_t K + \gamma p_t L + (1-2\gamma) p_t B \leq -\frac{\alpha}{(1-\alpha)} \left[\gamma e - \frac{\beta}{1-\beta} \gamma(u-e) - p_t X_{\min} \frac{1}{\beta^{n-1}} \left(\frac{1-\beta^{n-1}}{1-\beta} \right) \right]. \quad (19)$$

For such transactions, the PS must satisfy both IC and PC; i.e.,

$$p_t K \geq p_t B + e, \quad (20)$$

$$u + p_t L \geq p_t B, \quad (21)$$

$$p_t L \geq -\frac{\beta}{1-\beta} \gamma(u-e) - p_t X_{\min} \left[\frac{1}{\beta} + \cdots + \frac{1}{\beta^{n-1}} \right]. \quad (22)$$

Setting

$$p_t K = p_t L + e, \text{ and} \quad (23)$$

$$p_t L = p_t B \quad (24)$$

both IC are satisfied. In addition,

$$\gamma p_t K + \gamma p_t L + (1-2\gamma) p_t B \geq -\frac{\beta}{1-\beta} \gamma(u-e) - p_t X_{\min} \frac{1}{\beta^{n-1}} \left(\frac{1-\beta^{n-1}}{1-\beta} \right) + \gamma e. \quad (25)$$

Thus, the efficient allocation can be decentralized if and only if

$$\begin{aligned}
& -\frac{\alpha}{(1-\alpha)} \left[\gamma e - \frac{\beta}{1-\beta} \gamma(u-e) - p_t X_{\min} \frac{1}{\beta^{n-1}} \left(\frac{1-\beta^{n-1}}{1-\beta} \right) \right] \\
& \geq \gamma p_t K + \gamma p_t L + (1-2\gamma) p_t B \\
& \geq \gamma e - \frac{\beta}{1-\beta} \gamma(u-e) - p_t X_{\min} \frac{1}{\beta^{n-1}} \left(\frac{1-\beta^{n-1}}{1-\beta} \right), \text{ or,} \\
p_t X_{\min} & \geq \gamma \frac{\beta^{n-1}}{(1-\beta^{n-1})} (e - \beta u). \tag{26}
\end{aligned}$$

In order to maximize $p_t X_{\min}$, the PS must attempt to extract this amount by “taxing” all participants in monitored transactions; i.e., the PS must set $p_t \bar{K} = p_t \bar{L} = p_t \bar{B} = p_t L = p_t B = p_t X_{\min}$, so that

$$p_t X_{\min} = -(1-\alpha) \gamma e. \tag{27}$$

Thus, the PS can decentralize the efficient allocation if and only if

$$\begin{aligned}
p_t X_{\min} & = -(1-\alpha) \gamma e \geq \gamma \frac{\beta^{n-1}}{(1-\beta^{n-1})} (e - \beta u), \text{ or if and only if} \\
\beta^n u & \geq \beta^{n-1} e + (1-\alpha) (1-\beta^{n-1}) e. \tag{28}
\end{aligned}$$

■

In summary, the presence of some transactions that are not subject to private information frictions allows the PS to support the efficient volume of transactions for a wider range of parameters. In order for this to be accomplished for the widest possible range of parameter values, the system needs to “tax” all monitored transactions and use the obtained revenue in order to relax the incentive compatibility constraints of potential producers in non-monitored transactions.

Such adjustments offer a policy recommendation for actual PS. One can argue that transactions among banks that belong to different networks or among consumers and firms that belong to different debit card networks are subject to a higher degree of private information than within network transactions. Monitoring such transactions might be more costly. Since incentive constraints require that “inter-network transactions” are subsidized, any such costs must be borne during “within network transactions.”¹² In the next section we abstract from monitoring in order to study issues related to the optimal frequency of settlement and to the financing of settlement costs.

¹²This recommendation should not be taken literally, as it ignores all operational costs

4 Settlement

In the case where settlement is not subject to any operational costs, our model implies that a PS that decentralizes an efficient allocation for the widest range of parameter values will have settlement occurring after each transaction ($n = 1$).¹³ In actual PS, settlement occurs only periodically, partly since it is associated with certain operational costs. Here we use the value of δ to represent such fixed costs. We will first assume that δ is equally shared by all participants in the PS. Later, we study the best way to allocate costs across participants.

In this section we abstract from questions related to cross-subsidization of transactions. We, therefore, study the benchmark case in which all transactions are non-monitored ($\alpha = 0$). We will also assume that $\delta \leq \gamma(\beta u - e)$. In other words, settlement is not so costly that we can immediately rule out that it ever takes place. More restrictively, we assume that δ is small enough so that settlement must occur sometimes as part of the optimal arrangement. We assume that the fixed costs are covered by production of the general good during settlement rounds.

First, we demonstrate the feasibility of the efficient allocation when settling every n periods. In other words, we characterize the values of n for which the PS can decentralize an allocation in which the efficient quantity, q^* , is produced in each transaction, given a settlement cost $\delta > 0$. We have the following incentive constraints for producers and for consumers and non-traders, respectively

$$p_t K_s \beta^{n-s} = p_t B_s \beta^{n-s} + e, \quad (29)$$

$$B_s = L_s. \quad (30)$$

The most binding participation constraint, that of a consumer that consumed n times in a row, gives

$$p_t L \left(\frac{\beta}{1-\beta} \right) \left(\frac{1-\beta^n}{\beta^n} \right) \geq -\frac{\beta}{1-\beta} \gamma(u-e) + \delta \left(\frac{\beta^n}{1-\beta^n} \right). \quad (31)$$

other than those associated with incentive issues. Accounting for other costs might well require that inter-network transactions are also taxed. Our recommendation is simply that incentive constraints should not be ignored when the PS chooses balance adjustments.

¹³This is the reminiscent of a *Real-Time-Gross-Settlement (RTGS)* system in the context of our model.

Finally, market clearing requires

$$\gamma \sum_{s=1}^n K_s + (1 - \gamma) \sum_{s=1}^n L_s = \frac{-\delta}{p_t}. \quad (32)$$

The participation constraint takes into account the settlement cost, which is incurred every n periods. In addition, market clearing specifies that participants must pay the cost δ when settling their balances. As δ is a real cost, its value is expressed in units of balance adjustments. Finally, the PS provides incentives in the least costly way by making *all* incentive constraints exactly binding.

Substituting the incentive constraints in the market clearing condition we obtain

$$pL = -\delta \left(\frac{1 - \beta}{\beta} \right) \left(\frac{\beta^n}{1 - \beta^n} \right) - \gamma e. \quad (33)$$

Substituting this into the PC, we obtain

$$\beta^n u - e \geq \frac{\delta}{\gamma} \left(\frac{1 - \beta}{\beta} \right) \left(\frac{\beta^n}{1 - \beta^n} \right). \quad (34)$$

Hence, any n satisfying the above condition allows the PS to decentralize the efficient transaction level, q^* . By assumption, this is the case for at least some $n \geq 1$. We now turn to the question of the optimal settlement frequency. An optimal PS needs to minimize the incurred costs of settlement while, at the same time, decentralizing the efficient allocation. Thus, the PS must choose the longest length of a transaction cycle compatible with optimality, given the costs δ as expressed by the constraint (34). Note that if n is large enough, we have

$$\beta^n u(q^*) - e(q^*) < 0. \quad (35)$$

In particular, we have that $\frac{\beta^n}{1 - \beta^n} \rightarrow 0$ as $n \rightarrow \infty$. Hence, for large enough n , the participation constraint of a consumer that consumed n times in a row will be violated. In other words, there exists a maximal n such that constraint (34) is satisfied. This gives the settlement frequency that maximizes welfare among the participants of the PS. We summarize this in the following.

Proposition 3 *Suppose that $\beta^n u^* - e^* \geq \frac{\delta}{\gamma} \left(\frac{1 - \beta}{\beta} \right) \left(\frac{\beta^n}{1 - \beta^n} \right)$. Then an optimal PS exists. The optimal settlement frequency, n^* , is the maximum n for which this condition holds.*

4.1 Financing Settlement Costs

A constraint on the PS is that the operational fixed cost, δ , must be entirely financed by PS participants. In our analysis so far, we assumed that δ is shared equally by all participants. Here, we will briefly discuss certain implications of the model when we view the division of δ across participants as a policy variable. This is an important issue and actual PS differ substantially on their policies regarding the financing of operational costs, such as *inter-change fees*. Indeed, there is ongoing discussion about whether it should be mainly the consumers or the producers that finance such costs.

We perform a comparative statics exercise for the case where there is a slight increase in δ from its current value (which we normalize to zero). Since we will not discuss issues related to settlement frequency here, we set $n = 1$. We assume an economy in which initially $\delta = 0$. We let (K, B, L) represent the balance adjustments associated with an optimal PS for this economy. Furthermore, we assume that all ICs and PCs are slack when $\delta = 0$, that is

$$p_t K - e + \frac{\beta}{1 - \beta} \gamma(u - e) > p_t B + \frac{\beta}{1 - \beta} \gamma(u - e), \quad (36)$$

$$u + p_t L + \frac{\beta}{1 - \beta} \gamma(u - e) > p_t B + \frac{\beta}{1 - \beta} \gamma(u - e), \quad (37)$$

$$p_t L + \frac{\beta}{1 - \beta} \gamma(u - e) \geq 0, \quad (38)$$

$$\gamma K + \gamma L + (1 - 2\gamma) B = 0. \quad (39)$$

Next, suppose that δ increases slightly to become strictly positive. We will discuss the efficient way of financing this increase in operational costs. Let δ_K be the fraction of δ paid by producers, and similarly for δ_L and δ_B . A fully financed PS must satisfy the constraint that $\delta = \gamma\delta_K + \gamma\delta_L + (1 - 2\gamma)\delta_B$. A PS $\{K, L, B; \delta_K, \delta_B, \delta_L\}$ is optimal if and only if it satisfies the following.

$$p_t (K - \delta_K) - e + \frac{\beta}{1 - \beta} \gamma(u - e) \geq p_t (B - \delta_B) + \frac{\beta}{1 - \beta} \gamma(u - e) \quad (40)$$

$$\geq p_t (L - \delta_L) + \frac{\beta}{1 - \beta} \gamma(u - e) \geq 0 \quad (41)$$

$$\gamma K + \gamma L + (1 - 2\gamma) B = 0, \quad (42)$$

$$\gamma\delta_K + \gamma\delta_L + (1 - 2\gamma)\delta_B = \delta. \quad (43)$$

Starting from $\delta = \delta_K = \delta_B = \delta_L = 0$ and an initially optimal PS, the question is how should the increase in δ be shared across consumers,

producers, and no-traders. Suppose that we start increasing δ_K , δ_B , and δ_L in some way. Initially, all IC constraints will still hold as strict inequalities. In addition, the participation constraint of the consumer will either bind or not bind. If we can finance the entire δ while this constraint is still not binding, then the ordering of δ_K , δ_B , and δ_L is indeterminate. In other words, there is a continuum of possible ways to finance δ , and we cannot determine which side will finance most of the cost. If the participation constraint for the consumer is binding, however, the PS must stop increasing δ_L further, while continuing increasing δ_K and δ_B until the entire δ is financed. Thus, in general, $\delta_L \leq \delta_B \leq \delta_K$. We summarize this in the following.

Proposition 4 *In an optimal PS either $\delta_L < \delta_B, \delta_K$ or the order of $\delta_L, \delta_B, \delta_K$ is indeterminate.*

The intuition behind this proposition is simple. If the participation constraint binds for some participants, any increase in the costs imposed on them will lead to default. Thus, any such increases must be borne by those participants whose participation constraints are slack. In our model, the binding constraint is that of the consumer. Thus, if that constraint holds as an equality, the PS must impose the “tax increase” on non-consuming participants. This suggests that consumers must pay a lower fraction of such costs in countries where they are more credit constrained, such as the US. This intuition, which is different from the standard argument that relies on which side of the market is subject to more competition, seems to fit well with actual observations.

5 Discussion and Extensions

Under what conditions can a privately operated payment system support the efficient volume of transactions in the presence of private information? We studied optimal payment systems in a dynamic model in which the ability of agents to perform certain welfare improving transactions is subject to random and unobservable shocks. We examined the interplay between within-network (monitored) and inter-network (non-monitored) transactions. We demonstrated that an optimal PS will use the information about the former in order to induce the efficient volume of the latter type of transactions. Finally, we discussed issues related to optimal settlement frequency and to

the best way to finance settlement when this activity is subject to a fixed operational cost.

Our model could be used to further investigate several other issues related to payments. An open question is whether more complicated payment systems than the ones considered in this analysis could implement the first-best under less restrictive conditions. Although we have not studied this issue formally, we strongly suspect that the answer is no, in general. We are currently extending our analysis in order to study properties of payment systems when the first-best cannot be supported. This introduces an additional margin since the PS might now optimally choose to further reduce the settlement frequency by accommodating a decrease in the volume of transactions (a reduction in q).

Since the payment system can be thought of as being formed by the grand coalition of agents, there is nothing in our model that suggests that payment services should be publicly provided. Given that we deal with dynamic incentives, we could investigate the time consistency of various payment system policies. This adds another dimension to the debate regarding the public versus private provision of payment services since optimal dynamic schemes might require some commitment.

References

- [1] Atkeson A. and R. Lucas Jr. (1993) “On Efficient Distribution with Private Information,” *Review of Economic Studies*, vol. 59 (July)
- [2] BIS (2003) “Glossary of Terms Used in Payment and Settlement Systems”, March.
- [3] Fujiki H., Green E.J., and A. Yamazaki (1999) “Sharing the Risk of Settlement Failure,” Federal Reserve Bank of Minneapolis Working Paper 594D
- [4] Green E.J. (1987) “Lending and the Smoothing of Uninsurable Income,” in *Contractual Arrangements for Intertemporal Trade*, Edward C. Prescott and Neil Wallace, Eds. Minneapolis
- [5] Green, E.J., and R.H. Porter (1984) “Non-Cooperative Collusion Under Imperfect Price Information,” *Econometrica* 52 (1), p. 87-100
- [6] Kahn C.M., and W. Roberds (1998) “Payment System Settlement and Bank Incentives,” *The Review of Financial Studies* 11(4), p. 845-870
- [7] Kahn C.M., and W. Roberds (2001) “Real-Time Gross Settlement and the Costs of Immediacy,” *Journal of Monetary Economics*
- [8] Kahn C.M. (2006): “Why Pay?,” Plenary talk, *Conference on the Economics of Payments II*, Federal Reserve Bank of New York
- [9] Kiyotaki N., and R. Wright (1989) “On Money as a Medium of Exchange,” *Journal of Political Economy* 97, p. 927-954
- [10] Kiyotaki N., and R. Wright (1993) “A Search-Theoretic Approach to Monetary Economics,” *American Economic Review* 83, p. 63-77
- [11] Kocherlakota N. (2005) “Zero Expected Wealth Taxes: A Mirrless Approach to Dynamic Optimal Taxation”, *Econometrica* 73(5), p. 1587-1621
- [12] Koepl T., Monnet C., and T. Temzelides (2006) “A Dynamic Model of Settlement,” *Journal of Economic Theory*, forthcoming

- [13] Lagos R. and R. Wright (2006) “A Unified Framework for Monetary Theory and Policy Analysis,” *Journal of Political Economy*
- [14] Mirrlees J. (1971) “An Exploration in the Theory of Optimal Income Taxation,” *Review of Economic Studies* 38, p. 175-208
- [15] Spear S., and S. Srivastava (1987) “On Repeated Moral Hazard with Discounting,” *Review of Economic Studies* 54(4)