Queen’s Economics Department Working Paper No. 1124

Payments and Mechanism Design

Thor Koeppl
Queen’s University

Cyril Monnet
European Central Bank

Ted Temzelides
University of Pittsburg

Department of Economics
Queen’s University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

1-2007

Payments and Mechanism Design[*]

Thorsten Koeppl
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

March 27, 2007

[*] We thank audiences at CEMFI, The Central Bank of Portugal, Cornell, the Federal Reserve Banks of Cleveland and New York, the European Central Bank, Penn, Wharton, and Aleks Berentsen for comments and suggestions on earlier drafts of this and related work. 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 to study efficient intertemporal payment arrangements when the ability of agents to perform certain welfare-improving transactions is subject to random and unobservable shocks. Efficiency is achieved via a *payment system* that assigns balances to participants, adjusts them based on the histories of transactions, and periodically resets them through settlement. Our analysis has several implications for the design of actual payment systems. Efficiency requires that, in order to overcome informational frictions, agents participating in transactions that do not involve monitoring frictions subsidize those that are subject to such frictions. Optimal settlement frequency should balance liquidity costs from settlement against the need to provide intertemporal incentives. Settlement costs must be borne by agents for whom the incentives to participate in the system are highest. Finally, an increase in settlement costs implies that, in order to counter a higher exposure to default, the frequency of settlement must increase and, at the same time, the volume of transactions must decrease.

Keywords: Payment Systems, Frequency of Settlement, Liquidity Costs, Subsidization across Transactions

1 Introduction

One of the features of the economy that Walrasian models abstract from concerns the institutions through which payments for goods and services take place: the *payment system (PS)*. This results in the need for a framework that can guide policy makers in the efficient design of such systems. This paper builds such a framework using mechanism design.

Our approach involves three 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.[1] For example, the PS might need to make use of intertemporal incentives in order to explore the agents' willingness to participate and carry out transactions efficiently. Second, we emphasize the role of private information.[2] Actual PS design is subject to a private information problem since whether or not participants can perform certain transactions is not directly observable. For example, within a retail PS, the ability of a consumer to make a credit-card payment might not be observable. Similarly, within a wholesale system, banks might have private information about their ability to meet certain payment obligations. The third distinguishing feature of our analysis is that, unlike most of the literature in mechanism design, since we model the entire payment system, rather than a particular transaction or participant in isolation, we must take into consideration general equilibrium effects.

We employ a version of the search model that Kiyotaki and Wright (1989, 1993) developed in order to study monetary exchange. In contrast to the monetary theory approach, however, our model involves a "cashless" environment. This framework is appropriate for

[1] The existing literature on PS is almost exclusively static (see Kahn and Roberds (1998, 2001) for two prominent papers). Kahn (2006) provides an excellent summary of the current literature and outlines some of the main open questions in PS research.

[2] Hence, our work is related to the dynamic contracting literature (see Green (1987), and Spear and Srivastava (1987), among others). Our analysis also relates to recent work by Kocherlakota (2005), who extends the model of Mirrlees (1971) to a dynamic economy. 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 decentralize efficient allocations under private information.

our objective for several 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. This is important in actual PS where participants are subject to random needs for making payments to one another. 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.

We will assume that each transaction involves an agent that enjoys an instantaneous benefit: the "consumer," and an agent that suffers an instantaneous cost: the "producer."[3] In the presence of private information about the ability of agents to produce or consume, the rules imposed by the PS must provide the right incentives for its participants to be part of the arrangement and to reveal their information truthfully. The PS in our model accomplishes this by assigning individual "balances" and by specifying rules on how these balances are updated given the participants' trading histories. Furthermore, the PS requires that balances are periodically "settled".[4] This implies that participants are required to periodically "reset" their balances through centralized trading in what we will model as a Walrasian market. Optimal PS design involves providing incentives so that the efficient volume of transactions is carried out.

We use our framework to study two issues pertinent to the design of actual PS. The first concerns the structure of optimal balance adjustments in transactions between settlement periods. Our model shows that an optimal PS must shift the costs of providing incentives from the transactions stage to the settlement stage. Producers in the transactions stage

[3] There are several ways to rationalize such costs and gains. In a wholesale PS, for example, a *payer* bank might need to incur a cost in order to make a payment, while the *payee* enjoys a direct benefit from receiving the payment. In the context of a retail PS, the instantaneous benefit is usually enjoyed by a consumer who receives a good or service, while the cost is borne by its supplier. To fix ideas, we will hence refer to the two agents in the transaction as the "consumer" and the "producer," respectively.

[4] In a sequel to this paper, Koeppl, Monnet, and Temzelides (2006) demonstrate that settlement is a necessary feature of optimal payment systems.

are rewarded through balance increases, while consumers are penalized through balance decreases. These rewards and costs materialize when participants equalize their balances during settlement. One case of particular interest arises when some of the transactions are subject to monitoring, in which case the ability of the two parties to perform a transaction is observable to the PS.[5] In this case, we demonstrate that an optimal PS will "tax" monitored transactions in order to "subsidize" the costs of providing incentives in transactions that are subject to private information.

The second issue we study involves the frequency of settlement when there are operational costs associated with the settlement stage. Optimal settlement frequency implies evaluating such operational costs against the constraints arising from the need to provide intertemporal incentives. We discuss the optimal allocation of settlement costs. In our model, such costs must be borne by those participants for whom some participation constraints are slack. More generally, our approach 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. In contrast to the standard arguments that emphasize the role of competition, our approach highlights the role of default risk.

Settlement frequency is a key variable in actual PS design. High settlement frequency, such as in real time gross settlement systems, implies high liquidity costs. On the other hand, less frequent settlement, such as in net settlement systems, might lead to large net exposures and to an increased probability of default by participants that have built high negative balances. Currently, most actual large-value PS involve immediate settlement. As a response to increased liquidity costs, several PS offer short-term credit facilities, at the expense of potentially re-introducing default risk. We argue that optimal PS design should explore the trade-off between minimizing liquidity costs and dealing with default exposures. A strength

[5] For example, in some wholesale PS, banks interact mostly, but not exclusively, through a local network. Such networks, which are often run by large correspondent banks, might have detailed information about their participants. The same information, however, might not be available to the PS when a bank transacts outside its network. Similarly, in the context of retail PS, a consumer's credit card history might be readily available within a credit card network, but not necessarily within other networks.

of our approach is that these two costs have a clear counterpart in the model. The liquidity costs in our model are represented by the fixed costs of settling transactions. The default exposure arises from the possibility of a long sequence of unfavorable balance adjustments. This allows us to study this trade-off directly from fundamentals. We demonstrate that, in an optimal PS, an increase in liquidity costs results in an increase in the frequency of settlement and a decreased volume of transactions.

The paper proceeds as follows. Section 2 introduces the model and discusses the concept of a PS. Sections 3 and 4 discuss optimal balance adjustments, the optimal frequency of settlement, and the distribution of settlement costs. Section 5 offers a brief conclusion and discusses some of the many possible directions for future research. The Appendix introduces a continuous-time extension of the model and contains most of the proofs.

2 A Dynamic Model of Payments

2.1 The Environment

Time, t, is discrete and measured over the natural numbers. 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, in which n *transactions stages*, each consisting of exactly one bilateral meeting for every agent, is followed by a round of centralized trading (termed *settlement stage*) at the end of the last period.[6] Discounting applies after each period, except between the last transactions stage and the settlement stage. We describe the transactions stage and the settlement stage in turns.

During the transactions stage, agents are randomly matched bilaterally in each period. Randomness in the transactions needs is captured by assuming that in each period an agent can trade with the agent he is matched with as a producer or as a consumer, each with

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

probability γ. Thus, in each period during the transactions 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 disutility $-c(q)$, while consumption of q units gives utility $u(q)$ where $q \in \mathbb{R}_+$. We assume that $c'(q) > 0$, $c''(q) \geq 0$, $\lim_{q\to 0} c'(q) = 0$, and $\lim_{q\to\infty} c'(q) = \infty$. In addition, $u'(q) > 0$, $u''(q) \leq 0$, $\lim_{q\to 0}\ u'(q) = \infty$, and $\lim_{q\to\infty} u'(q) = 0$. Thus, there exists a unique q^* such that $u'(q^*) = c'(q^*)$. 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 concentrate on this quantity for most of the paper, we will simplify notation by letting u denote $u(q^*)$ and c denote $c(q^*)$.[7]

The information structure during the transactions stage is as follows. Whether or not a trade meeting has occurred is always observable to the two agents in the meeting. This, together with the identities of the two agents, is also publicly observable with probability α. On the other hand, with probability $1-\alpha$, neither the identities of the two agents, nor the type of their meeting is observable by anyone outside the meeting.[8] 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 good can be produced or consumed during this stage. Producing ℓ units of the general good implies disutility $-\ell$, while consuming ℓ units gives utility ℓ. Thus trading this good does not directly increase welfare. The settlement stage is frictionless in the sense

[7] We reiterate that what is important is that each transaction involves an agent that enjoys an instantaneous benefit and an agent that suffers an instantaneous cost. In a wholesale system these costs and benefits will typically be associated with costly payments of size q made by one participant to the other. The functions c and u provide a reduced-form way of capturing such costs and benefits.

[8] This information structure can be interpreted as the result of the agents being divided 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" meetings are monitored, while during "inter-network" meetings the ability to transact is private information of the trading partners.

that, just as in Walrasian markets, agents interact in a centralized fashion, and there are no informational frictions.

2.2 Payment Systems

An allocation for the above environment specifies the quantity produced and consumed during bilateral transactions, as well as the production and consumption of the general good for each agent during settlement. In general, allocations may well exhibit history-dependence. In this paper we largely study whether efficient allocations can be decentralized through a Payment System. A PS keeps a record of past transactions by assigning *balances* to its participants. In addition, the PS instructs participants to produce or consume certain quantities in trade meetings and specifies rules for how the balances are updated given the participants' current transactions. Finally, during the settlement stage, participants trade their balances against the general good in order to achieve a particular starting balance for the next transactions stage. Agents with low balances can then increase their balances by producing, while those with high balances can reduce them by purchasing consumption of the general good. Since we model the settlement stage as a Walrasian market, the price, p, at which balances are traded, is determined by market clearing conditions. We make three additional assumptions. First, we restrict attention to allocations that are stationary and symmetric across agents.[9] Second, we assume that the PS can permanently exclude from all future transactions agents that do not produce or consume the prescribed amount during monitored transactions or do not settle their balances. Finally, we rule out short-sales of balances in the Walrasian market. Later we will assume that each settlement stage involves an aggregate (average) resource cost $\delta > 0$ and study the implications of this cost for optimal settlement frequency. For simplicity, in this section we first formulate the general framework assuming that $\delta = 0$.

Formally, the PS keeps a record of all transactions. For any agent, in any given period,

[9] This is without loss of generality when, as is the case in most of the paper, the full-information-first-best allocation is decentralized.

t, this record is summarized by his *balance*, $d_t \in \mathbb{R}$. First, consider the problem of an agent in the settlement stage at time t. Let $V_t(d_t, p_t)$ denote his value function if he exits the last bilateral meeting in the transactions stage with balance d_t, given that the anticipated price in the following settlement stage is p_t. Let $E_t v_{t+1}(\widehat{d}_t, \Psi_{t+1})$ denote the expected future value of an agent who exits the settlement stage with balance $\widehat{d}_t$, given that the resulting distribution of balances is denoted by Ψ_{t+1}. The problem of the agent is then given by

$$V_t(d_t, p_t) = \max_{\ell_t, \widehat{d}_t} -\ell + \beta E_t v_{t+1}(\widehat{d}_t, \Psi_{t+1}) \tag{1}$$

subject to

$$p_t \widehat{d}_t = p_t d_t + \ell_t \tag{2}$$

$$\widehat{d}_t \geq \min\{0, d_t\}. \tag{3}$$

In this problem we have imposed the constraint that agents cannot short sell their balances in the Walrasian market. Note that, given a price level p_t, V_t is linear in balances, d_t.[10] While it is not necessary for the results, linearity greatly simplifies our analysis.

We now turn to the problem faced by the participants during the transactions stage. In each period, agents are bilaterally matched. In non-monitored meetings, the PS receives a joint report from the agents in the match. Formally, the two agents, say i and j, each report a number $\eta_i, \eta_j \in \{0, 1\}$. The agents' identities become known to the system if and only if $\eta_i \eta_j = 1$. If either agent chooses 0, so that $\eta_i \eta_j = 0$, the agents' identities are not revealed and the PS instructs them to exchange nothing. If both choose 1, they identify themselves to the PS as being in a trade meeting. The potential producer is then instructed to produce $q_t(d_t, d_t', \Psi_t)$ for the potential consumer. Note that q_t can depend on both participants' balances as well as on the overall distribution of balances. In a monitored meeting, no such reporting needs to occur since in that case the type of the meeting is observable to the PS. Hence, the PS instructs people directly to trade a quantity $\bar{q}_t(d_t, d_t', \Psi_t)$. Again, this quantity can depend on the balances of both participants as well as on the overall distribution.

Upon observing the reports, production, and consumption by every agent, the PS ad-

[10] This follows from an argument similar to that in Lagos and Wright (2005).

justs their balances. An *adjustment,* $X_t(d_t, d'_t, \Psi_t)$, is added to an agent's current balance, conditional on his current trading history. Recall that an agent 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 adjustments for each transactions round which we denote by $\{L_t, K_t, B_t, \bar{L}_t, \bar{K}_t, \bar{B}_t\}$. 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. Balances are represented by real numbers not restricted in sign, while production of goods during trade meetings is restricted to be positive.[11] Agents may decide to leave the PS at any point. In that case, we assume that they cannot be re-admitted and that they receive a permanent future payoff that is normalized to zero. We can now formally define a PS.[12]

Definition 1. *A* Payment System *is an array*

$$S_t(d_t, d'_t, \Psi_t) = \{L_t, K_t, B_t, q_t, \bar{L}_t, \bar{K}_t, \bar{B}_t, \bar{q}_t\}, \textit{ for all } t.$$

We restrict attention to PS that are *incentive feasible.* We term a PS incentive feasible if *(i)* all agents have an incentive to participate in each transaction as well as in the settlement stage, *(ii)* all agents in non-monitored transactions truthfully reveal their type of meeting they are in, and *(iii)* the market clears in each settlement stage. The last requirement implies

[11] To further clarify the information-related problem, the difficulty is that whether or not a trade meeting has taken place is not always observable to the PS. Thus, any arrangement must rely on reports by the agents about the type of the meeting that has taken place. The transaction protocol we have formulated for non-monitored meetings obtains the following interpretation. Each participant in the system has access to both a "card" and a "card-reading machine." Agents can choose to identify themselves to the PS by having their card read by their partner's machine ($\eta_i = \eta_j = 1$). In that case, the system becomes aware that the two agents are in a trading meeting. The balances of both parties are then updated given their reports and the production/consumption that has taken place.

[12] Strictly speaking, a PS must also condition on whether an agent chooses to participate in the system or not. For simplicity, we leave this indicator variable out of the formal definition of a PS.

that, in each t that corresponds to a settlement stage,

$$\int_{d_t} (\widehat{d_t} - d_t) d\Psi_t = 0. \tag{4}$$

This, in turn, immediately implies that $\int_{d_t} l_t d\Psi_t = 0$.[13] To formulate incentive compatibility (IC) and participation constraints (PC), we first describe the value functions of participants during the transactions stage. Recall that there are n bilateral meetings (one in each period) between settlement stages. To ease notation, we denote the current-period immediate return in period t by $f(X_t)$, where $X_t \in \{L_t, K_t, B_t, \bar{L}_t, \bar{K}_t, \bar{B}_t\}$. Thus, $f(L_t) = u(q_t)$, $f(K_t) = c(q_t)$, and $f(B_t) = 0$. The adjustments for monitored meetings, $\bar{L}_t$, $\bar{K}_t$, and $\bar{B}_t$, are defined similarly. If the last settlement stage occurred in period t, the value function during each round s, $s = 1, \ldots, n-1$, of the current transactions stage is given by

$$\begin{aligned} & E_{t+s-1}[v_{t+s}(d_{t+s-1}, \Psi_{t+s-1})] \\ = & \int_{d'_{t+s}} E[f(X_{t+s}) + \beta E_{t+s}[v_{t+s+1}(d_{t+s-1} + X_{t+s}, \Psi_{t+s})]] d\Psi_{t+s-1}, \end{aligned} \tag{5}$$

where E denotes the expectation over the type of meeting the agent will be in during the current period.[14] For the last period of the transactions stage, $t+n$, we have

$$\begin{aligned} & E_{t+n-1}[v_{t+n}(d_{t+n-1}, \Psi_{t+n-1})] \\ = & \int_{d'_{t+n}} E[f(X_{t+n}) + V(d_{t+n-1} + \bar{X}_{t+n}, p_{t+n})] d\Psi_{t+n-1}. \end{aligned} \tag{6}$$

Since consumption and production are verifiable, agents can only misreport during a non-monitored transaction by claiming that they are in a no-trade meeting. Assuming that the last settlement stage occurred at time t, allocations are IC during each round $s = 1, \ldots, n-1$ of the current transactions stage if in all non-monitored transactions we have

$$\begin{aligned} & f(X_{t+s}) + \beta E_{t+s}[v_{t+s+1}(d_{t+s-1} + X_{t+s}, \Psi_{t+s})] \geq \\ & \qquad f(B_{t+s}) + \beta E_{t+s}[v_{t+s+1}(d_{t+s-1} + B_{t+s}, \Psi_{t+s})], \end{aligned} \tag{7}$$

[13] As discussed earlier, the settlement round does not generate welfare. It redistributes welfare in a way that results in a "re-initialization" of the agents' histories. For a detailed discussion on this issue, we refer to Koeppl, Monnet and Temzelides (2006).

[14] For example, $\alpha\gamma$ gives the probability that the agent is engaged in a monitored production meeting. To ease notation, we have suppressed the dependence of the PS on $(d_{t+s}, d'_{t+s}, \Psi_{t+s})$.

where $X_{t+s} \in \{K_{t+s}, L_{t+s}, B_{t+s}\}$. Similarly, for $s = n$, IC requires that for all X_{t+n},

$$f(X_{t+n}) + V(d_{t+n-1} + X_{t+n}, p_{t+n}) \geq f(B_{t+n}) + V(d_{t+n-1} + B_{t+n}, p_{t+n}). \tag{8}$$

Finally, PC require that agents have an incentive to remain in the PS during both the transactions and the settlement stages. Thus, if the last settlement stage occurred at t, we require

$$f(X_{t+s}) + \beta E_{t+s}[v_{t+s+1}(d_{t+s-1} + X_{t+s}, \Psi_{t+s})] \geq 0, \tag{9}$$

for all X_{t+s}, and all $s = 1, \ldots, n-1$. Finally, for $s = n$ we require

$$f(X_{t+n}) + V(d_{t+n-1} + X_{t+n}, p_{t+n}) \geq 0, \tag{10}$$

for all X_{t+n}. The absence of short-sales implies that settlement involves a positive expected lifetime payoff, independent of any balance adjustments. Otherwise, agents would default on their obligations. Formally,

$$V(d_{t+n-1} + X_{t+n}, p_{t+n}) \geq 0. \tag{11}$$

In most of the paper, we will concentrate on implementing the full-information-first-best allocation, in which the efficient transaction level q^* is exchanged in *all* trade meetings, both monitored and non-monitored.[15] We have the following.

Definition 2. *A PS is* optimal *if it is incentive feasible and if it decentralizes the efficient level of production, q^*, in all trade meetings.*

To conclude this section, it is useful to specify two particular types of PS that we will use extensively in what follows. A PS is *simple* if balance adjustments do not depend on the agents' current balances. A PS is *simple* and *repeated* (SRPS) if, in addition, it satisfies

$$\begin{aligned} X_{t+s} &= \frac{X_{t+n}}{\beta^{n-s}}, \text{ and} && (12) \\ X_{t+kn} &= X, && (13) \end{aligned}$$

[15] Linear utility in the settlement stage implies that the utility from consuming any amount of the general good equals the disutility to the producer. Hence, the efficient amount of the general good production is indeterminate.

where $X \in \{L, K, B, \bar{L}, \bar{K}, \bar{B}\}$, $s = 1, \ldots, n$, and $k \in \mathbb{N}$. In the above expressions, t represents the date of a settlement round. In words, adjusting for discounting, a repeated PS imposes the same balance adjustments in each period of the transactions cycle. To ease notation, we will drop the time index whenever possible. SRPS are convenient since the linearity of V implies that, in any optimal SRPS, the incentive constraints for all $s = 1, \ldots, n-1$ are fulfilled whenever those for $s = n$ hold. In what follows, we restrict attention to SRPS.

3 Optimal Balance Adjustments

In the previous section, we described an environment in which certain transactions are subject to a private information problem. We introduced a payment system as a way to decentralize incentive feasible allocations for that environment. In this section, we use this setup in order to analyze some of the properties of optimal PS. We are particularly interested in investigating how the PS can use monitored transactions in order to alleviate the incentive problem in non-monitored transactions. To this end, we will first take the length of the transactions stage, n, as given, and assume that there are no settlement costs ($\delta = 0$). In the next section we study the issue of optimal settlement frequency when $\delta > 0$.

3.1 Perfect Monitoring

It is instructive to first discuss the case in which all transactions are perfectly monitored; i.e., $\alpha = 1$. This special case is convenient as it implies that there are no IC constraints. Thus, incentive feasible allocations need to satisfy only market clearing in the settlement stage and, of course, the PC. We will consider two optimal PS for this environment. First, suppose that the PS sets all balance adjustments permanently equal to zero, and imposes a "gift-giving" game in which agents are induced to produce in all trade meetings under the threat of permanent exclusion in the case of a deviation. Since this PS does not rely on any balance adjustments, there is never a need to trade balances in the settlement stage.

Furthermore, agents have an incentive to participate if and only if

$$\frac{\beta}{1-\beta}\gamma(u-c) \geq c, \tag{14}$$

or, if the future expected discounted utility from staying in the PS is greater than the current cost of producing the efficient quantity, c. Provided that the above participation constraint holds, this PS is optimal.

Second, consider a PS which uses balance adjustments in order to shift all costs related to incentive provision to the settlement stage. Denote the minimum balance adjustment in any given period by $X_t^{\min} = \min\{\bar{K}_t, \bar{L}_t, \bar{B}_t\}$. Normalize the required starting balance after the settlement stage in period t to $\widehat{d}_t = 0$.[16] Agents that do not leave the settlement stage with $\widehat{d}_t = 0$, as well as those that do not exchange q^* in a trade meeting, are permanently excluded from the PS. Hence, in equilibrium, the distribution of balances, Ψ_t, at the end of each settlement stage is degenerate, for all t. The only potentially binding PC are then given by

$$V(\widehat{d}_t + \sum_{s=1}^{n} X_{t+s}^{\min}, p_{t+n}) \geq 0, \tag{15}$$

and

$$f(X_{t+n}) + V(\widehat{d}_t + \sum_{s=1}^{n-1} X_{t+s}^{\min} + X_{t+n}, p_{t+n}) \geq 0. \tag{16}$$

The first constraint implies that an agent chooses to settle and remain in the PS even under the *worst* possible history of adjustments. The second constraint is the PC for agents in the last transaction round, conditional on having had the worst balance adjustment until this round. Such agents still need an incentive to carry out the transaction as they can always avoid the cost $f\left(X_{t+n}\right)$ by leaving the PS prior to the settlement stage. Finally, the PS must satisfy the market clearing condition during settlement (4) with $\widehat{d}_{t+n} = 0$; i.e.,

$$\gamma p_t \sum_{s=1}^{n} \bar{K}_{t+s} + \gamma p_t \sum_{s=1}^{n} \bar{L}_{t+s} + (1-2\gamma) p_t \sum_{s=1}^{n} \bar{B}_{t+s} = 0. \tag{17}$$

[16] This implies that the minimum adjustment will be negative, while agents that produce are rewarded with a positive adjustment.

When we restrict the PS to be simple and repeated, we can use (12) and (13) to simplify this market clearing condition to

$$\gamma p_t \bar{K} + \gamma p_t \bar{L} + (1-2\gamma) p_t \bar{B} = 0. \tag{18}$$

Let p denote the (constant) equilibrium price in the settlement round. We then have the following.

Proposition 3. *Suppose that $\alpha = 1$. If $\frac{\beta}{1-\beta}\gamma(u-c) \geq c$, the PS with no balance adjustments ($\bar{K}_t = \bar{L}_t = \bar{B}_t = 0$, for all t) is optimal. If $\beta^n u \geq c$, the simple repeated PS with balance adjustments $p\bar{K} = p\bar{L} + c = p\bar{B} + c$ is optimal.*

All proofs are relegated to the Appendix. Notice that the condition that $\beta^n u \geq c$ requires that settlement is sufficiently frequent. Thus, if n is large, the second PS is no longer optimal since it is not incentive feasible. It is easy to check that as long as

$$\gamma \geq \frac{\beta^n}{1-\beta^n}\frac{(1-\beta)}{\beta}, \tag{19}$$

the first PS is optimal for the widest set of parameter values. In other words, provided that this inequality holds, if an optimal PS exists, the first PS is optimal. Of course, there exist parameters for which both the above PS are optimal. The main difference between the two is that the first PS does not make use of the settlement stage, while the second one collects all costs from incentive provision during the transactions stage and periodically resets the participants' balances through settlement.

3.2 Cross-subsidizing Transactions

Before we analyze the possibility of cross-subsidization, it is useful to first consider the other extreme case in which there are no monitored transactions ($\alpha = 0$). In this case, the PS relies on reports by participants about whether or not they are in a trade meeting. By the linearity of V in d, the relevant IC in any period s, $s = 1, \ldots, n$, during the transactions stage are given by

$$-c(q_{t+s}) + \beta^{n-s} p_{t+n} K_{t+s} \geq \beta^{n-s} p_{t+n} B_{t+s}, \text{ and} \tag{20}$$

$$u(q_{t+s}) + \beta^{n-s} p_{t+n} L_{t+s} \geq \beta^{n-s} p_{t+n} B_{t+s}. \tag{21}$$

In this case, the SRPS that uses the balance adjustments described in the second part of Proposition 3 is optimal if and only if $\beta^n u \geq c$.[17] Of course, as long as $\beta^n u \geq c$, this PS also remains optimal for any $\alpha \in [0, 1]$. Indeed, even if $\alpha > 0$, the PS can always treat all transactions as if they are non-monitored, and make balance adjustments as in Proposition 3.

The above PS, however, does not exploit the fact that, as long as $\alpha > 0$, some transactions are monitored. The question then becomes whether an optimal PS exists in this case for a wider range of parameters (that is, even if $\beta^n u < c$). Such a PS would "tax" some of the surplus created in monitored transactions in order to relax the incentive constraints by "subsidizing" non-monitored transactions. The following Proposition asserts that this is indeed possible.

Proposition 4. *Assume that $\beta^n u < c$. There exists an optimal simple repeated PS if and only if*

$$\frac{\beta}{1-\beta}\gamma(u-c) - c \geq (1-\alpha)\gamma\left(\frac{\beta}{\beta^n}\frac{1-\beta^n}{1-\beta}\right)c. \tag{22}$$

For any parameter values satisfying the above inequality, the PS with balance adjustments $p\bar{K} = p\bar{L} = p\bar{B} = pB = pL = -(1-\alpha)\gamma c$ and $pK = pB + c$ is optimal.

The intuition behind this result is straightforward. First, recall that cross-subsidization is only useful when $\beta^n u < c$ (otherwise we already know that an optimal PS exists). This, together with (22), implies that (19) holds. Thus, without loss of generality, we employ the PS that uses a gift-giving game during monitored transactions. The left hand side of inequality (22) is the maximum total tax revenue that an optimal PS can extract when it employs a gift-giving game. The right hand side gives the total subsidy required in a

[17] See Koeppl, Monnet and Temzelides (2006) for a formal analysis of this special case.

settlement stage for the incentive constraints to be satisfied for non-monitored producers. In other words, the above PS imposes a uniform lump-sum tax on all participants and then uses the proceeds in order to make the incentive constraints for non-monitored producers hold. Such a PS is optimal as long as the available taxes are high enough to subsidize the cost of providing incentives to non-monitored producers.

We remark that the restriction to SRPS likely involves some loss of generality as such a PS cannot extract *all* possible surplus. Indeed, a PS that would condition on the entire history of all transactions since the last settlement stage could potentially redistribute surplus more efficiently. Finally, note that, since we assume that $\beta^n u < c$, condition (22) cannot be met if $\alpha = 0$, in which case there are no monitored transactions to finance the incentive subsidy.

It is worth pointing out that, as the above analysis suggests, balances in our model are distinct from currency in at least one important way. When two agents transact using currency, the amount of money that the seller receives is equal to the amount that the buyer offers. Here, however, in order for non-monitored transactions to be subsidized, it is necessary that *both* the buyer and the seller in monitored transactions receive a negative balance adjustment. The ability to implement such a policy, which is analogous to a transaction-specific tax or subsidy, distinguishes the PS in our model from a monetary authority imposing an inflation tax.

4 Optimal Settlement Frequency

4.1 Settlement Costs

In actual PS, settlement occurs only periodically due to both operational costs associated with the activity itself, as well as liquidity costs that settlement imposes on PS participants.[18] On the other hand, it is also recognized that infrequent settlement allows for the buildup

[18] While the former cost is the natural point of discussion in retail PS, the latter is a key issue for wholesale PS, in particular, for what practitioners call Real-Time-Gross-Settlement (RTGS) versus Deferred-Net-Settlement (DNS).

of overly large adverse balances by some participants. This trade-off raises the question how to optimally balance such costs against the higher default exposure associated with less frequent settlement. Our model captures the fact that infrequent settlement increases the potential for default. However, the analysis so far has abstracted from settlement costs and assumed the length of the transactions stage as exogenously given. For the remainder of the paper, in order to investigate this trade-off, we assume that there is a fixed cost, $\delta > 0$, that is incurred in each period in which settlement takes place. The existence of this cost is important if one wants to study the frequency of settlement as a policy variable chosen by the PS. Indeed, in the absence of such costs, our model implies that an optimal PS will have settlement occurring after each transaction; i.e., $n = 1$.

To streamline the analysis, for the remainder of the paper we abstract from questions related to cross-subsidization of transactions and study the benchmark case in which there is no monitoring ($\alpha = 0$). We will also assume that δ is small enough, so that settlement must periodically occur as part of an optimal PS. Finally, we assume that δ is covered by production of the general good during settlement.

Our first goal is to characterize the values of n for which an optimal PS exists in the presence of the fixed settlement cost $\delta > 0$. Building on our earlier findings, we consider a PS that sets balance adjustments such that $pK = pB + c = pL + c$. This satisfies all IC and PC (see equation (20) and (21)). The PS must also recover the cost δ. Hence, market clearing during the settlement stage is now given by

$$\gamma \sum_{s=1}^{n} K_{t+s} + \gamma \sum_{s=1}^{n} L_{t+s} + (1-\gamma) \sum_{s=1}^{n} B_{t+s} = -\frac{\delta}{p_{t+n}}. \tag{23}$$

Assuming that participants share the settlement costs equally[19], the value function, V, of an agent prior to entering the settlement stage is now given by

$$V(d_t, p_t) = d_t p_t + \frac{\beta}{1-\beta}\gamma(u-c) + \delta\left(\frac{\beta^n}{1-\beta^n}\right). \tag{24}$$

We have again assumed, without loss of generality, that the PS sets the desired balance at the end of the settlement stage to $\widehat{d_t} = 0$. Like before, the first two terms of the value

[19] As we show in the next section, this is without loss of generality.

function give the value of the agent's current balances and the future expected utility from participating in the PS, respectively. The last term gives the present value of all future settlement costs.[20]

Like before, it is sufficient to check that the PC under the worst possible balance adjustments holds during the last period of the transactions stage. For the above PS, this constraint is given by

$$pL\left(\frac{\beta}{1-\beta}\right)\left(\frac{1-\beta^n}{\beta^n}\right) \geq -\frac{\beta}{1-\beta}\gamma(u-c)+\delta\left(\frac{\beta^n}{1-\beta^n}\right). \tag{25}$$

One can use the market clearing condition to solve for the implied adjustment for consumers, L, which is given by

$$pL = -\delta\left(\frac{1-\beta}{\beta}\right)\left(\frac{\beta^n}{1-\beta^n}\right)-\gamma c. \tag{26}$$

Substituting this into the PC, we obtain

$$\beta^n u - c \geq \frac{\delta}{\gamma}\left(\frac{1-\beta}{\beta}\right)\left(\frac{\beta^n}{1-\beta^n}\right). \tag{27}$$

Hence, any n satisfying this condition allows the above PS to decentralize the efficient transaction level, q^*. We then have the following.

Proposition 5. *Assume that $\alpha = 0$. There exists an optimal simple repeated PS if and only if*

$$\beta^n u - c \geq \frac{\delta}{\gamma}\left(\frac{1-\beta}{\beta}\right)\left(\frac{\beta^n}{1-\beta^n}\right). \tag{28}$$

For any parameter values satisfying the above inequality, the PS with balance adjustments $pK = pB + c = pL + c$ is optimal. The optimal settlement frequency, n^*, is the maximum n for which this condition holds.

Conditional on decentralizing an efficient allocation, an optimal PS should minimize the incurred costs from settlement. In other words, the PS must choose the longest length of a transactions stage that is compatible with optimality, given the costs δ as expressed by (27). Such an n exists as long as $\delta \leq \gamma(\beta u - c)$. Note also that if n is large enough, we have that

$$\beta^n u < c. \tag{29}$$

[20] The current settlement cost is contained in the agent's balance, d_t.

Thus, if settlement is sufficiently infrequent (n is sufficiently large), the PC of an agent that consumed n times in a row will be violated. In other words, there exists a maximum n such that constraint (27) is satisfied.

4.2 Financing Settlement Costs

Throughout, we are assuming that the PS is self-financed. This implies that the fixed cost, δ, associated with its operation must be entirely financed by PS participants. In our analysis so far, we assumed that δ is shared equally by all participants. Actual PS, however, differ substantially on their policies regarding the financing of such costs.[21]

Here we study the division of δ across participants as a policy variable by performing a comparative statics exercise for the case where there is a small increase in δ from its current value, which we normalize to zero. For simplicity, we also set $n = 1$. Denote the value of the cost paid by each producer by δ_K. Consumers and non-traders pay δ_L and δ_B, respectively. Clearly, since we assume that the PS is self-financed, we have

$$\gamma\delta_K + \gamma\delta_L + (1 - 2\gamma)\,\delta_B = \delta. \tag{30}$$

The IC (20) and (21) are then given by

$$\begin{aligned} pK - \delta_K - c &\geq pB - \delta_B, \text{ and} && (31) \\ u + pL - \delta_L &\geq pB - \delta_B. && (32) \end{aligned}$$

Assuming, as before, that $\widehat{d_t} = 0$, and since the expected settlement cost in any future period is given by δ, the PC (16) is given by

$$f(X) + pX - \delta_X \geq -\frac{\beta}{1-\beta}\left[\gamma(u - c) - \delta\right]. \tag{33}$$

[21] While we associate δ with settlement costs here, our analysis applies to other costs associated with the operation of a PS. One example is credit card fees in retail PS. There is an ongoing debate about whether consumers or stores should be responsible for such fees. Another example is cost recovery in wholesale PS involving banks. The discussion there involves whether the "payee" or "payer" should pay, as well as whether the fees should be fixed or volume-based.

This PC implies that if an optimal PS exists, one can always construct an optimal PS by setting balances such that $pK - \delta_K - c = pB - \delta_B = pL - \delta_L$. Hence, we have the following.

Proposition 6. *Suppose that an optimal simple repeated PS exists. Then, there exists an optimal PS in which settlement costs are shared equally across all agents; i.e.,* $\delta_L = \delta_B = \delta_K = \delta$.

The intuition for this result is straightforward. The right-hand side in (33) gives the value of an agent who stays in the PS after settlement. This value is independent of any history of past transactions. An optimal PS can always set balance adjustments so as to make all PC exactly binding.[22] In that case, when $\delta = \delta_K = \delta_B = \delta_L = 0$, the PS sets $pK = pB + c = pL + c$. If δ increases, such a PS would increase the costs to all agents so as to keep the incentives to participate the same across all agents.

The above argument continues to hold even if $n > 1$, as long as $\alpha = 0$. Of course, other cost allocations can also be consistent with an optimal PS. However, it is worth mentioning that when $\alpha > 0$, it is the PC for agents that have consumed n-times in a row that will bind first. In that case, any optimal PS will levy a higher share of the settlement costs on non-consumers; i.e., $\delta_L = \delta_B < \delta_K$. The rationale for this policy prescription is different from the standard argument that competitive forces drive the allocation of PS costs. Rather, the above argument suggests that, in order to reduce their incentive to default during the settlement stage, it is the agents who are most constrained that must pay the lowest share of the costs.[23]

[22] This is not a necessary feature of optimal PS. However, if an optimal PS exists, one can always construct another optimal PS that satisfies this property. In this sense, this property characterizes PS that are optimal for the widest range of parameter values.

[23] This observation is consistent with the fact that interchange fees for credit cards are more likely to be passed to consumers in countries where consumer credit-card debt and default rates on such debt are relatively low.

4.3 Settlement Frequency and Transaction Size

It is recognized that in actual PS frequent settlement involves high liquidity costs. Less frequent settlement, on the other hand, allows for the netting of exposures by different participants. At the same time, this could lead to default problems since it allows for exposures to become too large. Recognizing this trade-off, actual PS often make provisions, such as offering short-term credit facilities, in order to economize on liquidity costs. A strength of our approach is that the trade-off between liquidity costs and default exposure has a counterpart in the context of our model. The liquidity costs are represented by the shared fixed costs associated with settling transactions. The exposure arises from the possibility of building large negative balances if settlement is too infrequent.

So far in our analysis we termed a PS optimal if it decentralizes the efficient level of production, q^*, in all transactions. However, in the presence of settlement costs, an optimal PS must explore the trade-off between reducing the size of transactions versus lengthening the transactions stage. We now turn to the more general problem of determining *jointly* the efficient settlement frequency and the efficient transaction size.

We assume that $\delta \leq \gamma(\beta u - c)$. These costs must be small enough, so that it is optimal for settlement to occur eventually. Based on our earlier result, since $\alpha = 0$, the fixed cost δ is shared equally across all participants and is covered by production of the general good in the settlement stage. In the Appendix we use a continuous-time formulation of our model in order to set up the joint choice of settlement frequency and transaction size by the PS as a planning problem. In order for the constraint set of this problem to be convex, it is sufficient that the cost function $c(q)$ is log-convex. Under these assumptions, we derive the following.

Proposition 7. *Assume that $c(q)$ is log-convex. Any optimal simple PS implies that $q_t = \hat{q} < q^*$, for all t. Furthermore, as δ increases, the optimal transaction size, $\hat{q}$, as well as the optimal length of the settlement cycle, $\hat{T}$, decrease.*

The first part of the Proposition confirms our intuition. Proposition 5 already established that (given that q^* can be decentralized) the PS must optimally reduce settlement frequency as much as possible in order to economize on settlement costs. The above Proposition asserts that it is also optimal to economize further on settlement costs by reducing the intensive margin (i.e., the transaction size) below its first-best level.[24] This is equivalent to reducing participants' exposures prior to the settlement stage, by imposing a tighter cap on the total amount of goods produced during bilateral transactions.

The second part of the Proposition is somewhat more surprising. It asserts that as a response to an increase in settlement costs, an optimal PS must adjust *both* $\hat{q}$ and $\hat{T}$ in the same direction. In other words, an optimal PS must reduce the volume of balance adjustments that need to be settled in two complementary ways: by shortening the length of the transactions stage, and by reducing the transactions size. The explanation for this is as follows. The binding constraint on the PS is the PC of an agent that has consumed the most during the transactions stage and, as a result, has to settle a large negative balance. An increase in δ makes it more likely that this agent's participation constraint will be violated. Hence, in order to avoid default, the PS must decrease the potential exposure of this agent by reducing his negative balance adjustments. This involves reducing both the quantity produced and the time between settlement periods.

5 Conclusion

Under what conditions can a PS decentralize the efficient volume of intertemporal transactions in the presence of private information? We studied this question in a dynamic model in which the ability of agents to perform at least some welfare improving transactions is subject to random and unobservable shocks. In particular, we examined the interplay between monitored and non-monitored transactions. In general, the optimal PS will tax the first type of transactions in order to subsidize incentive provision for the latter type. We also discussed key issues for PS design related to optimal settlement frequency and cost recovery.

Our framework is certainly not limited to these questions and can be used to further

[24] One can show that an optimal PS involves a constant level of production across transactions.

investigate several other issues related to payments. One open question is whether more complicated payment systems than the ones considered here could decentralize efficient outcomes under less restrictive conditions. For the no-monitoring case ($\alpha = 0$), restricting attention to simple, repeated payment systems is without loss of generality. Due to linearity in the settlement round, whenever the IC in the last period of the transaction stage hold as equalities under such a PS, all other IC in earlier rounds also hold as equalities.

When some transactions are monitored, a PS that is simple but not repeated still cannot improve on SRPS. Under any simple PS, only the total adjustments accumulated for the next settlement stage matter. Thus, a SRPS can decentralize a constant production level in all transactions that lead to at least as high total adjustments as any other simple PS. However, a PS where balance adjustments are history-dependent within a transactions cycle can likely improve on a SRPS. Hence, we would expect more conditional balance adjustments in PS with better information about participants' transactions.

An important current debate concerns the public versus private provision of payment services. Given that our framework deals with dynamic incentives, we could investigate the time consistency of various payment system policies since optimal dynamic schemes might require some commitment. Finally, the existence of different competing payment networks and of tiered structures in PS, as well as extending the model to incorporate aggregate risk, outlines a whole range of interesting issues that our framework can potentially address.

6 Appendix

Proof of Proposition 3

For the first part of the proof, note that zero balance adjustments imply that there is no trade in the settlement stage. It can then be easily verified that, when $q_t = q^*$, for all t, the PC for consumers, producers, and non-traders are the same for every period. The PC for producers is fulfilled if and only if equation (14) holds.

For the second part of the proof, we first show that the PC (15) and (16) imply that all other PC are satisfied. First, note that the value function $V(d_t, p_t)$ is linear in p_t and given by

$$V(d_t, p_t) = \frac{\beta}{1-\beta}(u-c) + p_t(d_t - \widehat{d_t}). \tag{34}$$

Having normalized $\widehat{d_t} = 0$ for all t, the PC at $t+n-1$ gives

$$\begin{aligned}
& f(X_{t+n-1}) + \beta E_{t+n-1}[v_{t+n}(d_{t+n-2} + X_{t+n-1}, \Psi_{t+n-1})] \geq \\
& f(X_{t+n-1}) + \beta E_{t+n-1}[v_{t+n}(\sum_{s=1}^{n-2} X_{t+s}^{\min} + X_{t+n-1}, \Psi_{t+n-1})] = \\
& f(X_{t+n-1}) + \beta \left[\gamma(u-c) + E_{t+n-1}[V(\sum_{s=1}^{n-2} X_{t+s}^{\min} + X_{t+n-1} + X_{t+n}, p_{t+n})] \right] = \\
& f(X_{t+n-1}) + \beta \left[\frac{1}{1-\beta}\gamma(u-c) + p_{t+n}\left(\sum_{s=1}^{n-2} X_{t+s}^{\min} + X_{t+n-1} \right) + p_{t+n}E[X_{t+n}] \right] = \\
& f(X_{t+n-1}) + \beta p_t X_{t+n-1} + \frac{\beta}{1-\beta}\gamma(u-c) + \beta p_{t+n}\left(\sum_{s=1}^{n-2} X_{t+s}^{\min} \right) + \beta p_{t+n}E[X_{t+n}] = \\
& f(X_{t+n}) + p_{t+n}X_{t+n} + \frac{\beta}{1-\beta}\gamma(u-c) + p_{t+n}\left(\sum_{s=2}^{n-1} X_{t+s}^{\min} \right) + \beta p_{t+n}E[X_{t+n}] \geq \\
& f(X_{t+n}) + p_{t+n}X_{t+n} + \frac{\beta}{1-\beta}\gamma(u-c) + p_{t+n}\left(\sum_{s=1}^{n-1} X_{t+s}^{\min} \right) = \\
& f(X_{t+n}) + V(\sum_{s=1}^{n-1} X_{t+s}^{\min} + X_{t+n}, p_{t+n}),
\end{aligned}$$

which is just the PC for adjustment X in the last transactions round. The last inequality follows since $X^{\min} \leq 0$, and market clearing implies $E[X_{t+n}] = 0$. Hence, the PC at $t+n-1$ hold provided that they hold for $t+n$. By induction, it follows that they also hold for any $t+s$, $s = 1, \ldots, n-2$.

Next, suppose that $\beta^n u \geq c$, and let $p_t\bar{K} = p_t\bar{L} + c = p_t\bar{B} + c$. Market clearing implies that $p_t\bar{K} = (1-\gamma)c$. It then follows that both PC (15) and (16) are satisfied. The PS is thus incentive feasible. Since it decentralizes q^*, it is also optimal.

Proof of Proposition 4

First note that condition (22) is equivalent to

$$\frac{c}{\gamma}(1-\beta^n)\left[\alpha\gamma - \frac{\beta^n}{(1-\beta^n)}\frac{1-\beta}{\beta}\right] \geq c - \beta^n u. \tag{35}$$

Define a PS as in the statement of the Proposition. It is clear that $K \neq X_{\min}$. In addition, all IC are fulfilled and the PC for producers holds since $pK \geq pB + c$. Market clearing requires that

$$\alpha p\bar{K} + (1-\alpha)\left(pB + \gamma c\right) = 0, \tag{36}$$

or that

$$p\bar{K} = -(1-\alpha)\gamma c. \tag{37}$$

The necessary and sufficient condition for the PS to decentralize the efficient allocation in the transactions stage is, thus, given by

$$p\bar{K}\left(1 + \frac{1}{\beta} + \cdots + \frac{1}{\beta^{n-1}}\right) - c \geq -\frac{\beta}{1-\beta}\gamma(u-c). \tag{38}$$

All other PC hold if this PC involving the worst balance adjustment is fulfilled. Hence, we obtain a single condition in terms of parameters which is given by

$$(1-\alpha)\gamma c\frac{\beta}{1-\beta}\frac{1-\beta^n}{\beta^n} - c \geq -\frac{\beta}{1-\beta}\gamma(u-c). \tag{39}$$

Rewriting this expression we obtain

$$\beta^n u \geq c + c\beta^n\frac{1-\beta}{\beta}\frac{1}{\gamma} - \alpha c(1-\beta^n), \tag{40}$$

or

$$\frac{c}{\gamma}(1-\beta^n)\left[\alpha\gamma - \frac{\beta^n}{(1-\beta^n)}\frac{1-\beta}{\beta}\right] \geq c - \beta^n u. \tag{41}$$

This is condition (35).

For the converse, consider again the PS specified in the statement of the Proposition. Under this PS, only the PC for monitored producers is binding. Suppose that one increases $\bar{K}$ by any amount $\Delta > 0$, and, at the same time, lowers all other balance adjustments by

$-\frac{\alpha\gamma}{1-\alpha\gamma}\Delta$, so as to satisfy the market clearing condition. This relaxes the constraint (38) if and only if

$$\alpha\gamma < \frac{\beta^n}{\beta}\frac{1-\beta}{1-\beta^n}. \tag{42}$$

First, suppose that this condition is satisfied. The PS can then set $pK = pB + c$, $p\bar{K} = p\bar{L} = p\bar{B} = pB = pL$, and $p\bar{K}$ such that

$$pB\frac{\beta}{\beta^n}\frac{1-\beta^n}{1-\beta} = -\frac{\beta}{1-\beta}\gamma(u-c), \text{ and} \tag{43}$$

$$\alpha\gamma p\bar{K} + (1-\alpha)\gamma c + (1-\alpha\gamma)pB = 0. \tag{44}$$

This PS is optimal for the widest possible range of parameters. Since $B = X_{\min}$, in order to satisfy all PC, it must be the case that $p\bar{K} - c \geq pB$, or,

$$\frac{\beta^n}{1-\beta^n}\gamma(u-c) \geq \gamma c. \tag{45}$$

But this implies that $\beta^n u \geq c$, a contradiction. Hence, it must be the case that $\alpha\gamma > \frac{\beta^n}{\beta}\frac{1-\beta}{1-\beta^n}$. Suppose now that there exists an optimal PS while condition (35) is not satisfied. We have just shown that in that case the PS with $p\bar{K} = p\bar{L} = p\bar{B} = pB = pL = -(1-\alpha)\gamma c$ and $pK = pB + c$ is optimal for the widest possible range of parameters. But then, as condition (35) is violated, the PC for non-monitored producers, equation (38), cannot be satisfied, a contradiction.

Proof of Proposition 7

In order to demonstrate Proposition 7, we find it convenient to use differential calculus. To this end, we develop a continuous-time version of the model. We assume that consumption and production opportunities follow a Poisson process with arrival rate γ. The (continuous) rate of time preference is now denoted by ρ. The fixed cost, δ, is incurred whenever the transaction process stops and settlement occurs. This occurs after a deterministic interval of length T. As before, we denote balance adjustments by $(K(t), L(t), B(t))$. All other assumptions remain the same as in the text.

We let the random time before the next arrival of a trading opportunity be denoted by τ. In that case, τ has a distribution function given by

$$F(t) = \Pr(\tau \leq t) = 1 - \Pr(\tau > t) = 1 - e^{-\gamma t}. \tag{46}$$

Hence, the time until the next arrival of a trading opportunity is an exponentially distributed random variable with distribution function $F(t) = 1 - e^{-\gamma t}$.

In order to determine, for any given q, the expected future payoff for an agent at the end of the settlement stage, we first consider a PS that employs a gift-giving game as described in the main body of the paper. Denote this expected utility by V_0. It is straightforward to show that an optimal PS involves a constant level of transactions. First, assume that there are no settlement costs. Since both consumption and production opportunities are independent, arrive at rate γ, and have the same continuation value, we have

$$\begin{aligned} V_0 &= \int_0^\infty e^{-\rho t}(u(q) - c(q) + V_0) d(1 - e^{-\gamma t}) \\ &= \frac{\gamma}{\gamma + \rho}(u(q) - c(q) + V_0), \end{aligned} \tag{47}$$

which yields

$$V_0 = \frac{\gamma}{\rho}(u(q) - c(q)). \tag{48}$$

This is analogous to the lifetime utility under a PS that employs a gift-giving game in the discrete-time version of the model presented in the text. In the absence of settlement costs, equation (48) also gives the life-time expected payoff of an incentive feasible PS that decentralizes transactions of size q.

When costly settlement occurs after each time length T, it involves an aggregate (average) fixed cost δ. Hence, the net present value of the settlement costs is given by

$$\sum_{n=1}^{\infty} e^{-n\rho T}\delta = \delta \frac{e^{-\rho T}}{1 - e^{-\rho T}}. \tag{49}$$

This implies that the continuous-time version of the value function, V_0, is given by

$$V_0 = \frac{\gamma}{\rho}(u(q) - c(q)) - \delta \frac{e^{-\rho T}}{1 - e^{-\rho T}}. \tag{50}$$

As before, we define the PS adjustments conditional on the agents' reports by

$$pK_t - c(q) = pL_t = pB_t, \tag{51}$$

for all t. Also, since the PS is repeated, we have that adjustments, X, satisfy

$$X_{nT+t} = Xe^{\rho(T-t)}, \tag{52}$$

for all $t \in [nT; (n+1)T]$, where n is an integer. As in the discrete-time case, such a PS implies that all IC are fulfilled. In addition, this PS satisfies all PC for the largest set of parameter values. Next, we derive the market clearing condition for the settlement stage. This is accomplished by approximating total balance adjustments in an interval of length T. First, note that the probability of having exactly n arrivals of trading opportunities in the interval $[0, t]$ is given by

$$P[N_t = n] = e^{-\gamma t} \frac{(\gamma t)^n}{n!}. \tag{53}$$

For small Δ, we then have that

$$P[N_\Delta = 1] \approx \gamma \Delta, \tag{54}$$

where $P[N_\Delta > 1] = o(\Delta)$. Next, define $\Delta = \frac{T}{m}$, where $m \in [0, T]$ is an integer. The total adjustment for producers over an interval of length T is then approximately given by

$$\begin{aligned}
&\gamma\Delta K_\Delta + \cdots + \gamma\Delta K_{(m-1)\Delta} + \gamma\Delta K_{m\Delta} \\
&\quad = \gamma\Delta K \left[e^{\rho(T-\Delta)} + \cdots + e^{\rho(T-(m-1)\Delta)} + e^{\rho(T-m\Delta)}\right] \\
&\quad = \gamma\Delta K e^{\rho T} \left[\frac{1 - \left(e^{-\rho\Delta}\right)^m}{1 - e^{-\rho\Delta}} - 1\right] \\
&\quad = \gamma K \left[\frac{\Delta e^{\rho(T-\Delta)} - \Delta}{1 - e^{-\rho\Delta}}\right].
\end{aligned} \tag{55}$$

As $\Delta \to 0$, an agent will receive either none or one opportunity to trade during a time length Δ. In that case, using L'Hôpital's rule,[25] the expected total adjustments to producers are

[25] Note that both the numerator and the denominator in this expression go to zero. In addition, we have $\lim_{\Delta\to 0} \frac{f'(x)}{g'(x)} = \lim_{\Delta\to 0} \frac{-\Delta\rho e^{\rho(T-\Delta)} + e^{\rho(T-\Delta)} - 1}{\rho e^{\rho\Delta}} = \lim_{\Delta\to 0} -\Delta e^{\rho T} + \frac{e^{\rho T}}{\rho} - \frac{1}{\rho e^{-\rho\Delta}} = \frac{1}{\rho}(e^{\rho T} - 1)$.

given by $\frac{\gamma}{\rho}K\left(e^{\rho T}-1\right)$. The expected total balance adjustments to consumers are similarly determined and given by $\frac{\gamma}{\rho}L\left(e^{\rho T}-1\right)$. Finally, expected balance adjustments over all agents who have received no trading opportunities over this time interval can be determined as follows. For each interval of length Δ, a measure $2\gamma\Delta$ of agents are engaged in transactions ($\gamma\Delta$ of them as consumers and $\gamma\Delta$ as producers). Therefore, the measure of agents who are not involved in any transactions over an interval of length Δ is $(1-2\gamma)\,\Delta$. As a result, the aggregate balance adjustments for non-trading activities over the interval of length T are given by

$$\begin{aligned}&(1-2\gamma)\,\Delta B_{\Delta}+\cdots+(1-2\gamma)\,\Delta B_{(m-1)\Delta}+(1-2\gamma)\,\Delta B_{m\Delta}\\&=\frac{(1-2\gamma)}{\rho}B\left(e^{\rho T}-1\right).\end{aligned} \tag{56}$$

Market clearing during the settlement stage is then given by the following equation

$$\frac{1}{\rho}\left(e^{\rho T}-1\right)\left[\gamma pK+\gamma pL+(1-2\gamma)pB\right]=-\delta. \tag{57}$$

Using the above balance adjustments, one obtains

$$pB=-\delta\rho\frac{1}{e^{\rho T}-1}-\gamma c(q). \tag{58}$$

The worst possible balance adjustment is assigned to agents that either never traded or never produced in the interval $[0,T]$. Following the above discussion, this adjustment is given by $\frac{1}{\rho}\left(1-e^{-\rho T}\right)pB$. This implies that the only PC that is potentially binding is given by

$$\frac{1}{\rho}\left(e^{\rho T}-1\right)pB+\frac{\gamma}{\rho}(u(q)-c(q))-\delta\frac{e^{-\rho T}}{1-e^{-\rho T}}\geq 0. \tag{59}$$

This constraint is identical to the one in discrete-time, simply adjusting for the continuous time discount factor. Given these adjustments, an optimal PS chooses q and T in order to solve the following maximization problem:

$$\max_{q,T} \frac{\gamma}{\rho}(u(q)-c(q)) - \delta \frac{e^{-\rho T}}{1-e^{-\rho T}} \tag{60}$$

$$\text{subject to}$$

$$\frac{1}{\rho}\left(e^{\rho T}-1\right) pB + \frac{\gamma}{\rho}(u(q)-c(q)) - \delta\frac{e^{-\rho T}}{1-e^{-\rho T}} \geq 0,$$

$$pB = -\delta\rho\frac{1}{e^{\rho T}-1} - \gamma c(q).$$

The objective function expresses the discounted lifetime utility of a representative participant. The second constraint summarizes the PC that is potentially binding, while the third constraint summarizes the IC and the market clearing conditions that must be satisfied in any incentive feasible PS. The equality in the last equation follows from the fact that the PS works for the largest set of parameters if it makes all IC exactly bind. The constraint set can be rewritten to obtain

$$\frac{\gamma}{\rho}\left[u(q) - e^{\rho T}c(q)\right] \geq \delta\frac{1}{1-e^{-\rho T}}, \tag{61}$$

or

$$\left(1-e^{-\rho T}\right)u(q) - \left(e^{\rho T}-1\right)c(q) \geq \frac{\rho}{\gamma}\delta. \tag{62}$$

The objective function is strictly concave in (q, T). In order to guarantee that the constraint set is convex, we need an additional assumption. Given any T (q), the function on the left-hand side of the above inequality is concave in q (T). However, the left-hand side is not necessarily jointly concave in (q, T) due to the second term, which is a product of two convex functions. We have the following sufficient condition for the constraint set to be convex.[26]

Lemma 8. *Suppose that $c(q)$ is log-convex. Then $e^{\rho T}c(q)$ is a strictly convex function in (q, T), and the constraint set is convex.*

Proof. A function is log-convex if its natural logarithm is convex. Since $c(q)$ is log-convex, we have that

$$\frac{\partial^2 \ln c(q)}{\partial q^2} = \frac{c(q)c''(q) - (c'(q))^2}{(c(q))^2} > 0. \tag{63}$$

[26] A weaker condition is given by $-\frac{1}{e^{\rho T}}u''(q)c(q) \geq c'^2 - c''(q)c(q)$.

The first term of the LHS in equation (61) is strictly concave in q, while the RHS is strictly convex in T. The remaining term has a Hessian given by

$$H(q,T) = \begin{pmatrix} \rho^2 e^{\rho T} c(q) & \rho e^{\rho T} c'(q) \\ \rho e^{\rho T} c'(q) & e^{\rho T} c''(q) \end{pmatrix}. \tag{64}$$

The first principal minor is positive, while the second principal minor is positive if and only if

$$c(q)c''(q) - (c'(q))^2 > 0. \tag{65}$$

Hence, as $c(q)$ is log-convex, $e^{\rho T} c(q)$ is convex. The result follows since the sum of two concave functions is concave. □

Taking first-order conditions with respect to q and T, we obtain the following characterization of the solution:

$$\frac{u'(q) - c'(q)}{c'(q)} = \frac{\lambda}{1+\lambda}\left(e^{\rho T} - 1\right), \text{ and} \tag{66}$$

$$\frac{\delta}{c(q)} \frac{\rho}{\gamma} \left(\frac{1}{e^{\rho T} - 1}\right) = \frac{\lambda}{1+\lambda}\left(e^{\rho T} - 1\right), \tag{67}$$

where λ is the multiplier on the single constraint. This leads us to the following.

Lemma 9. *Let $c(q)$ be log-convex. For any optimal PS with settlement, we have $\hat{q} < q^*$.*

Proof. Since δ, γ, and ρ are positive, and the optimal settlement length is finite ($\hat{T} \in (0, \infty)$), we must have that $\lambda > 0$. Hence, equation (66) implies that $u'(\hat{q}) - c'(\hat{q}) > 0$. Since c is increasing and strictly convex, and u is increasing and strictly concave, this implies that $\hat{q} < q^*$. □

Eliminating λ from the first-order conditions (66) and (67) we obtain a single first-order condition

$$\frac{\gamma}{\rho} \frac{u'(q) - c'(q)}{c'(q)} \left(e^{\rho T} - 1\right) = \frac{\delta}{c(q)}. \tag{68}$$

This condition, together with the constraint (61), characterize the solution $(\hat{q}, \hat{T})$. Solving these equations yields the optimal length of the transactions stage, $\hat{T}$, as a function of ρ and

$\hat{q}$; i.e.,

$$\frac{u(q)}{u'(q)}\frac{c'(q)}{c(q)} = e^{\rho T}. \tag{69}$$

The optimal transaction size, $\hat{q}$, is given by

$$u(q)\left(1 - \frac{c'(q)}{u'(q)}\right) + c(q)\left(1 - \frac{u'(q)}{c'(q)}\right) = \delta\frac{\rho}{\gamma}. \tag{70}$$

A solution to the last equation exists by the Intermediate Value Theorem. Furthermore, any solution must lay in an interval $[\underline{q}, q^*]$, where $\underline{q} > 0$. The problem is that the left-hand side of equation (70) is non-monotonic. Hence, there will, in general, be more than one solution to this equation. The optimal solution, however, corresponds to the one closest to (and below) q^*. The next Proposition relies solely on the fact that at this solution, $\hat{q}$, the left-hand side of equation (70) is *locally* strictly decreasing.

Lemma 10. *Assume that $c(q)$ is log-convex. As the settlement cost, δ, increases, the optimal transaction size, $\hat{q}$, as well as the optimal length of the transactions stage, $\hat{T}$, decrease.*

Proof. We establish first that $\hat{q}$ and $\hat{T}$ move in the same direction; i.e., that $\frac{d\hat{T}}{d\hat{q}} > 0$. Differentiating the left-hand side of equation (69) with respect to q, we obtain

$$\frac{1}{(u'(q)c(q))^2}\left[u(q)u'(q)\left(c(q)c''(q) - (c'(q))^2\right) + c(q)c'(q)\left((u'(q))^2 - u(q)u''(q)\right)\right], \tag{71}$$

which is strictly positive, as u is strictly increasing and strictly concave, while c is log-convex.

Next, we show that $\hat{q}$ is decreasing in δ. Denote the left-hand side of equation (70) by $\Gamma(q)$. Differentiating $\Gamma(q)$ with respect to q and collecting terms we obtain

$$\frac{\partial\Gamma}{\partial q} = c''(q)\left[\frac{c(q)}{c'(q)}\frac{u'(q)}{c'(q)} - \frac{u(q)}{u'(q)}\right] + u''(q)\left[\frac{c'(q)}{u'(q)}\frac{u(q)}{u'(q)} - \frac{c(q)}{c'(q)}\right]. \tag{72}$$

We can rewrite equation (70) as

$$\frac{\gamma}{\rho}\frac{u'(q) - c'(q)}{c'(q)}\left(\frac{u(q)}{u'(q)}\frac{c'(q)}{c(q)} - 1\right) = \frac{\delta}{c(q)}. \tag{73}$$

Since $u'(q) > c'(q)$, for $q < q^*$, we obtain that $\frac{u(q)}{u'(q)} > \frac{c(q)}{c'(q)}$. Letting $q \to q^*$, this implies that

$$\frac{c(q)}{c'(q)}\frac{u'(q)}{c'(q)} - \frac{u(q)}{u'(q)} < 0, \tag{74}$$

and

$$\frac{c'(q)}{u'(q)}\frac{u(q)}{u'(q)} - \frac{c(q)}{c'(q)} > 0. \tag{75}$$

Hence, $\frac{\partial \Gamma}{\partial q} < 0$, or, equivalently, the left-hand side of equation (70) is strictly decreasing for q sufficiently close to q^*. Furthermore, $\Gamma(q)$ converges to 0 as $q \to q^*$. Hence, $\Gamma(q) > 0$ for q sufficiently close to q^* and, by the continuity of $\Gamma(q)$, there must exist a solution to equation (70) for small enough $\delta > 0$. Finally, since $\Gamma(q) \downarrow 0$ as $q \to q^*$, we must have that $\Gamma'(\hat{q}) \leq 0$ (with Γ having possibly a local maximum at $\hat{q}$). This completes the proof. □

References

[1] Fujiki H., Green E.J., and A. Yamazaki (1999) “Sharing the Risk of Settlement Failure,” Federal Reserve Bank of Minneapolis Working Paper 594D

[2] 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

[3] Green, E.J., and R.H. Porter (1984) “Non-Cooperative Collusion Under Imperfect Price Information,” *Econometrica* 52 (1), p. 87-100

[4] Kahn C.M. (2006): “Why Pay?,” Plenary talk, *Conference on the Economics of Payments II*, Federal Reserve Bank of New York.

[5] Kahn C.M., and W. Roberds (1998) “Payment System Settlement and Bank Incentives,” *The Review of Financial Studies* 11(4), p. 845-870

[6] Kahn C.M., and W. Roberds (2001) “Real-Time Gross Settlement and the Costs of Immediacy,” *Journal of Monetary Economics*

[7] Kiyotaki N., and R. Wright (1989) “On Money as a Medium of Exchange,” *Journal of Political Economy* 97, p. 927-954

[8] Kiyotaki N., and R. Wright (1993) “A Search-Theoretic Approach to Monetary Economics,” *American Economic Review* 83, p. 63-77

[9] Kocherlakota N. (2005) “Zero Expected Wealth Taxes: A Mirrless Approach to Dynamic Optimal Taxation”, *Econometrica* 73(5), p. 1587-1621

[10] Koeppl T., Monnet C., and T. Temzelides (2006) “A Dynamic Model of Settlement,” *Journal of Economic Theory*, forthcoming

[11] Lagos R. and R. Wright (2006) “A Unified Framework for Monetary Theory and Policy Analysis,” *Journal of Political Economy*

[12] Mirrlees J. (1971) “An Exploration in the Theory of Optimal Income Taxation,” *Review of Economic Studies* 38, p. 175-208

[13] Rocheteau G. and R. Wright (2005) “Money in Search Equilibrium, in Competitive Equilibrium, and in Competitive Search Equilibrium,” *Econometrica* 73, p. 175-202

[14] Spear S., and S. Srivastava (1987) “On Repeated Moral Hazard with Discounting,” *Review of Economic Studies* 54(4)