BANK OF ENGLAND

Working Paper No. 428
Intraday two-part tariff in payment systems

Tomohiro Ota

May 2011

BANK OF ENGLAND

Working Paper No. 428
Intraday two-part tariff in payment systems

Tomohiro Ota[1]

Abstract

This paper studies the optimal intraday pricing in payment systems and its impact on banks' payment behaviour and intraday liquidity management. A model is developed to compare the performance of two different mechanisms to reduce payment delay: a throughput guideline and a tariff that varies over time, and concludes that a linear time-varying tariff achieves a better outcome unless the payment system experiences a system-wide liquidity shock. We show that settlement delay can be socially efficient, contrary to general understanding of the literature, when it reduces the aggregate cost of liquidity. The theoretical model suggests that the tariff eliminates the inefficient settlement delay that does not contribute to lowering the cost, while leaving the socially efficient delay.

Key words: Payment, RTGS, two-part tariff, throughput guideline.

JEL classification: E42, E58, G21.

[1] Bank of England. Email: tomohiro.ota@bankofengland.co.uk

The views expressed in this paper are those of the author, and not necessarily those of the Bank of England. The author wishes to thank Matthew Brewis, James Chapman, Rodney Garratt, John Jackson, Marius Jurgilas, Thomas Nellen, Ben Norman, Jochen Schanz, Anne Wetherilt, seminar participants at the Bank of England and Swiss National Bank, and an anonymous referee for the helpful comments and suggestions. Any remaining errors are of course the responsibility of the author.
This paper was finalised on 31 January 2011.

The Bank of England's working paper series is externally refereed.

Information on the Bank's working paper series can be found at
www.bankofengland.co.uk/publications/workingpapers/index.htm

Publications Group, Bank of England, Threadneedle Street, London, EC2R 8AH
Telephone +44 (0)20 7601 4030 Fax +44 (0)20 7601 3298 email mapublications@bankofengland.co.uk

ISSN 1749-9135 (on-line)

Contents

Summary

Timely and liquidity-efficient settlement of payments is an important policy objective for central banks. Settlement delay is, however, recognised as a potential problem in major payment systems. This paper studies two possible solutions to the problem of settlement delay, throughput guidelines and a time-varying tariff, compares their performances, and discusses the design of a time-varying tariff.

The economics of payment literature generally assumes that early payments are always good. Banks have an incentive to delay their payments to minimise the cost of liquidity. By delaying payments until other banks make payments to them, they can free-ride the cash inflow to make their own payments. Since every bank delays aiming at the free-riding, no bank can successfully recycle payment inflow from others. The 'competition of delay' is socially inefficient. This paper also confirms the inefficiency of the 'competition of delay', but finds that delaying payments is not always inefficient. It is socially optimal for a bank with a higher cost of liquidity to delay its payments and for a bank with a lower cost to make early payments. By doing so, the payment system can establish an efficient role-sharing to minimise the aggregate cost of intraday liquidity. That is, the low-cost bank prepares more intraday liquidity than a high-cost bank, and the high-cost bank can recycle the payment inflows (cash) from the low-cost bank for its payments for free. The delay need not be long — just until the bank with the higher cost of liquidity has received funds in.

The typical solution to the delay, the throughput guidelines adopted by the United Kingdom and others, is to penalise a bank if it fails to make a certain fraction of payments by predetermined deadlines. The model in this paper shows that these guidelines have potential drawbacks. First, they do not penalise payment delay until the deadline. As a result, they may create a bunching of payments just before the deadline, as the guidelines provide greater incentives for banks to make last-minute payments. Second, they impose the same deadline on all banks in the payment system even if they have different liquidity costs. This inhibits heterogeneous banks from the efficient role-sharing.

The second solution, the time-varying tariff adopted by Switzerland and others, penalises late payments in a different way. A payment system with such a tariff charges member banks a fee

(tariff), which is increasing over time, on each payment. This paper shows that a linear time-varying tariff can overcome the potential drawbacks of throughput guidelines. The tariff allows each member bank to determine its optimal payment schedule, according to its cost of liquidity. The efficient delays are retained, while the inefficient 'competition of delay' is eliminated. The tariff itself is independent of the cost — ie a system operator does not need to monitor each bank's cost of liquidity, which would be costly or infeasible, to design the optimal tariff.

We also show that the tariff fails to encourage early payments in the specific situation where banks simultaneously experience a large rise in liquidity cost, as in a liquidity crisis. Otherwise, the tariff improves the efficiencies of the payment system by minimising the aggregate cost of liquidity and discouraging inefficient settlement delay, compared with the throughput guidelines.

1 Introduction

Timely and liquidity-efficient settlement of payments is an important policy objective for central banks. Settlement delay is, however, recognised as a potential problem in real-time gross settlement (RTGS) systems, especially.[1] Bech and Garratt's (2003) seminal paper describes the mechanism of delay. In RTGS, banks are required to prepare cash to make payments. Banks can obtain cash by borrowing from the central bank, or by receiving a payment inflow from another bank. The former option allows the banks to make a timely payment, but they have to bear the cost of liquidity outflow.[2] The latter option gives free cash to the banks, but they have to postpone their payments until they receive payment inflows. A delay may also be costly for banks, because the customers who request the payments possibly prefer timely settlement (Bech and Garratt denote this as the cost of delay). And if all the banks wait for payment inflows from other banks, no payment would be made until the last minute, resulting in significant intraday settlement delay.

Bech (2008) mentions three possible solutions to the problem. First, central banks can lower the cost of liquidity. If the cost is negligibly small, their banks have fewer incentives to await payment inflows for free cash.

The second one is a throughput guideline. In the United Kingdom, member banks of CHAPS, the country's large-value payment system, are required on average, over the course of a month, to settle 50% of the daily value of their payments by noon and 75% by 2.30 pm. There is no defined penalty for violating the guideline in practice, but the violation incurs a small reputation loss, which is believed to be sufficient to discipline the banks. CHIPS in the United States has a similar arrangement.

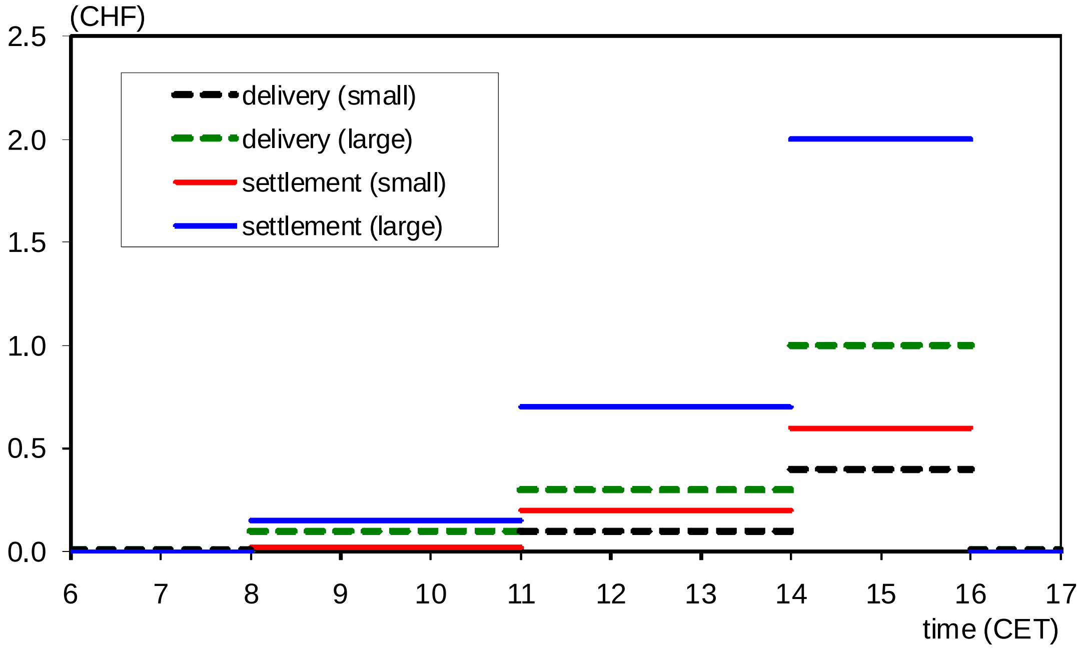
The third solution is a time-varying tariff, which SIC in Switzerland adopts. Many payment system operators charge member banks a small fee (tariff) to process each payment.[3] Most operators charge a fixed tariff on each payment, irrespective of the payment timing. The tariff of

[1] See, for instance, Committee on Payment and Settlement Systems (2005), McAndrews and Rajan (2000), and Armantier, Arnold and McAndrews (2008).

[2] The banks have to pledge collateral to the central bank to obtain cash, even though the borrowing rate is normally zero across countries. The cost of liquidity is one measure of the opportunity cost of the collateral.

[3] See Manning, Nier and Schanz (2009) for detail of the tariff structure of payment systems (page 172, for example).

Chart 1: The processing fees (tariffs) of SIC



*A delivery fee is charged when a bank submits a payment order, and the settlement fee is charged when the order is settled. 'Small' means the fee for payments smaller than CHF 100,000, while 'large' means the fee for large-value payments (over CHF 100,000).

SIC, on the other hand, is increasing over time through the day and by the size of the payment (see Chart 1). The time-varying tariff penalises later payers by charging a higher fee on later (and larger) payments.

The payments literature has mainly focused on the first option, eg in the context of liquidity-saving mechanisms (see eg Martin and McAndrews (2008); or Willison (2005)). In this paper, instead, we limit our attention to the latter two: throughput guidelines and time-varying tariffs. The aim of the paper is to examine whether the time-varying tariff would be more efficient or not comparing with throughput guidelines, and what that tariff could look like. The following is a summary of the paper's results.

The two options, throughput guidelines and time-varying tariffs, are characterised by their different natures to penalise late payments. The throughput guidelines are an example of trigger strategies, which impose a lump-sum penalty on an agent if the agent fails to satisfy a preset criterion. Time-varying tariffs do not have such thresholds: the tariffs penalise banks continuously as they delay payments. In reality SIC's time-varying tariffs does not continuously penalise late payments as Chart 1 shows, but this paper focus on the case of continuous penalty to clarify the conceptual difference.

Trigger strategy (throughput guidelines) has several potential drawbacks. First, it does not penalise payment delay until the deadline, and there is no additional penalty even if a bank keeps delaying payments after failing to meet the deadline.[4] As a result, it may create a bunching of payments just before the deadline, as the strategy incentivises banks to make last-minute payments. Second, a settlement agent who operates the system has to choose an appropriate deadline that maximises the efficiency of the system. This may not be an easy task, because the optimal payment timing of each bank is not observable for the agent. Third, the throughput guideline enforces an identical (or, at least, similar) payment schedule to all the member banks. This may be another source of inefficiency, because each member bank's optimal payment timing is unlikely to be the same. Theoretically, the central bank may be able to tailor the deadline to each one of the member banks' cost structures. But this is an unrealistic option because it requires the central bank to observe each bank's costs of liquidity and delay.

The model in this paper will show that a linearly increasing tariff over the course of the day can overcome the potential drawbacks of the throughput guideline. A significant benefit of the tariff is allowing potentially optimal payment schedules to be achieved by each member bank. If a bank finds its cost of liquidity low, then the bank can make payments earlier to save on the cost of delay and the tariff. On the other hand, if a bank's cost of liquidity is high, it may find it optimal to delay the payments: the bank has to bear a higher cost of delay and the tariff, but has more chance to receive free cash inflows from other banks. This paper focuses on the case where member banks are heterogeneous only in the costs of liquidity and assumes that the banks are identical in the cost of delay, for the sake of simplicity.

Allowing banks greater flexibility over their payment schedules through a time-varying tariff has two benefits. First, the central bank does not have to know the costs of each member bank. The central bank still needs to know the average cost of liquidity to determine the slope of the time-varying tariff, but it can apply the same tariff function to all member banks. Banks modify their payment timings based on their own cost structure, which is private information. In other

[4] In the United Kingdom, there are two deadlines at noon and 2:30pm - but it does not change the argument. No further penalty is applied.

words, the tariff achieves the constrained social optimum, where the central bank cannot observe all available information.

Second, the flexibility of the tariff not only provides a benefit to each member bank, but also improves social welfare. The welfare improvement comes from the incentive to improve management of intraday liquidity created by the tariff. If there are two banks, which are heterogeneous in their costs of liquidity, then the aggregate cost of liquidity is minimised when the bank with low liquidity cost borrows more cash, and the bank with high cost recycles the cash received from the low-cost bank to make its own payments.

From a policy perspective, settlement systems are normally required to achieve cost recovery. They should not make a loss, but should not make an excess profit either. The throughput guidelines do not need to consider cost recovery, since these impose a non-pecuniary penalty. This paper shows that, under a time-varying tariff, there is an easy way to achieve cost recovery, even though it entails a pecuniary penalty.

Theoretically, the most relevant paper is Walsh (1995), who also studies a linearly increasing penalty that enables central bankers to commit to the target inflation rate. While Walsh (1995) studies how the penalty works in a principal (government) - agent (central banks) relationship, we initially look at two agents' (banks') competition to show when and how settlement delay occurs in payment systems. This paper then considers a feasible measure, a linearly increasing tariff, by which the principal (the central bank) eliminates inefficient settlement delay.

Other relevant literatures in payment economics are the studies of throughput guidelines and of two-part tariffs in payment systems. The study of Buckle and Campbell (2003) is an example of the former. Holthausen and Rochet (2005, 2006) study optimal pricing in a large-value payment system when the central bank maximises welfare. Holthausen and Rochet (2006) show that a volume-discounting price policy, ie the per-transaction fee is lowered where participants make a large number of payments, is optimal, when the central bank cannot observe each participant's degree of willingness to make payments and the payment system is not allowed to make any profit or loss (full cost recovery). Holthausen and Rochet (2005) study a case when the central bank provides a public good or service, and show that a subsidy to a public payment system can be optimal when it is competing with a private system, to ensure the appropriate provision of the public service. This paper also studies a two-part tariff as we will see below, but is different from the literature in many aspects: eg instead of small payers being penalised by the volume discount, late payers are penalised in this model. To my knowledge, this is the first paper studying intraday pricing of large-value payment systems and its impact on payments behaviour and intraday liquidity management, Holthausen and Rochet, however, focus on studying daily pricing across systems.

Another contribution of this paper is to develop a way of modelling payment behaviour. To date, payment behaviour has been mainly studied with a simple discrete time scale: mostly two periods, morning and evening. But in this model, the payment timing is chosen from a continuous set and a Cournot duopoly game (with two-sided incomplete information) is played. There are several interesting implications arising from this set-up for payments behaviour in general (for example, it may suggest the reason why banks have been recently paying earlier in Fedwire); and significantly it shows that settlement delay is not always inefficient.

The remainder of the paper is organised as follows. Section 2 describes the model. Section 3 describes the equilibrium of the game when the banks choose their actions at their discretion. Section 4 provides the socially optimal outcome as a benchmark; and two possible measures to improve efficiency are discussed in Section 5. Section 5 also shows that the time-varying tariff works better than throughput guidelines. Section 6 provides a holistic discussion of the tariff and concludes.

2 Payment behaviour: basic model and delay

Suppose that there are two banks 1 and 2 in an RTGS payment system. The banks have many payment obligations to be settled by the end of the day. It is assumed that their payment behaviour is summarised by one variable, the delay of these payments (d_1 and d_2), because it is infeasible for the banks to manage each one of the payments. The delay of bank $i \in \{1, 2\}$ is denoted as $d_i \in [0, T]$, ie their delays are chosen from the continuous bounded segment: 0 represents a point in time when the bank settles the payment obligations as soon as possible, and T represents the maximum possible delay. Bank i chooses d_i before the payment system opens, and cannot change this during the day.

A bank determines the delay based on two costs. First, if it delays settling payments, it has to bear an intraday delay cost that is increasing over time. This partly represents a reputation loss for the bank when it fails to settle their customers' payments in a timely fashion.

Second, banks can save on their cost of liquidity by delaying their payments. In RTGS, a bank who pays first in the system has to obtain cash at the cost of intraday liquidity. The recipient of the payment does not have to do so — it can recycle the received cash for its own payment for free. The liquidity cost is, therefore, a decreasing function of delay, given that the other bank's delay is fixed: the later a bank pays compared to its the counterparty, the lower its liquidity cost.

2.1 *Definition of the game*

Bank i chooses d_i to minimise the following loss function at the beginning of the day.

$$\min_{d_i} V_i(d_i) = l(d_i; d_j, c + \varepsilon_i) + k(d_i)$$

The loss function V is additively separable. $l(d_i; c, d_j, \varepsilon_i)$ is the liquidity cost function, which is assumed to be C^2 class. The liquidity cost function has three parameters: c, market-wide liquidity cost; d_j, the counterparty j's delay; and ε_i, bank i's idiosyncratic liquidity cost. ε_i for $\forall i$ is private information throughout the game, but the distribution is publicly known; ε_i follows a uniform distribution on a closed set X, and is identically and independently distributed across banks. Lastly ε_i follows a uniform distribution and $E[\varepsilon_i] = 0$ for $\forall i$, ie $\varepsilon_i \in [-\bar{\varepsilon}, \bar{\varepsilon}]$. The banks are identical except for ε: so $V_i(\cdot) = V_j(\cdot)$ if $\varepsilon_i = \varepsilon_j$.

$k(d_i)$ is the function of intraday delay, which is also C^2 class. The delay cost is independent of the counterparty's delay d_j.

For the sake of simplicity and analytical tractability, we approximate the functions by Taylor expansion up to the second order. By expanding the liquidity cost function at $d_i = d_j = 0$ and $c + \varepsilon_1 = c$, we have:

$$
\begin{aligned}
l_1(d_1, d_2, c+\varepsilon_1) &\simeq \hat{l}_1(d_1, d_2, c+\varepsilon_1) \\
&= l_{1,d_1}(0,0,c)\cdot d_1 + l_{1,d_1d_2}(0,0,c)\cdot d_1\cdot d_2 + l_{1,d_1\varepsilon_1}(0,0,c)\cdot d_1\cdot\varepsilon_1 \\
&\quad + l_{1,d_2}(0,0,c)\cdot d_2 + l_{1,d_2\varepsilon_1}(0,0,c)\cdot d_2\cdot\varepsilon_1 \\
&= \alpha d_1 + \beta d_1 d_2 + \gamma d_1\varepsilon_1 + \delta d_2 + \eta d_2\varepsilon_1 \qquad \textbf{(1)}
\end{aligned}
$$

where $\alpha = l_{1,d_1}(0,0,c)$ is a first-order derivative coefficient; $\partial l_1(d_1 = 0, d_2 = 0, c+\varepsilon_1 = c)/\partial d_1$. The other parameters are defined in the same manner: $\beta = l_{1,d_1d_2}(0,0,c) = \partial^2 l_1(d_1 = 0, d_2 = 0, c+\varepsilon_1 = c)/\partial d_1\partial d_2$, $\gamma = l_{1,d_1\varepsilon_1}(0,0,c)$, $\delta = l_{1,d_2}(0,0,c)$ and $\eta = l_{1,d_2\varepsilon_1}(0,0,c)$. All the other coefficients, eg $l_{1,d_2d_2}(0,0,c)$ and $l_1(0,0,c)$, are assumed to be zero for the sake of simplicity. We would have a similar result even if we did not simplify the parameter set. Since the banks are assumed to be identical *ex ante*, bank 2 has the same parameters and δ: ie $l_{1,d_1}(0,0,c)$=$l_{2,d_2}(0,0,c)$ $l_{1,d_1d_2}(0,0,c) = l_{2,d_2d_1}(0,0,c)$ and so on.

The delay cost function k is approximated in the same manner. Bank 2 has the same parameters θ and μ as it is symmetric.

$$
\begin{aligned}
k(d_1) &\simeq \hat{k}_1(d_1) \\
&= k_{1,d_1}(d_1 = 0)\cdot d_1 + \frac{1}{2}\cdot k_{1,d_1d_1}(d_1 = 0)\cdot d_1^2 \\
&= \theta d_1 + \frac{1}{2}\mu d_1^2
\end{aligned}
$$

The coefficients of the approximated liquidity cost function, eg $l_{i,d_i} = l_{i,d_i}(0; 0, l_c)$, $l_{i,d_id_j} = l_{i,d_id_j}(0; 0, l_c)$ are also defined in the same way.

We impose several restrictions on the derivatives to reflect the intuition provided above:

A1: $\alpha < 0$

A2: $\delta > 0$

A3: $\alpha + \delta = 0$

A4: $\beta < 0$

A5: $\gamma < 0$

A6: $\gamma + \eta = 0$

A7: $\mu > 0$ and $\theta = 0$

Delaying payments allows a paying bank to lower its liquidity cost, by utilising payment inflows from the counterparty for the paying bank's own payment. This is assumption A1. But this is costly for the counterparty since its net payment position temporarily increases due to the delay in payment inflow — the counterparty has to prepare additional cash to make its payments. A2, $\delta > 0$, describes this negative externality.

A3 assumes that relative delay matters. If two banks delay their payments to the same extent, the liquidity-saving effect α and the negative externality δ offset each other. This is not an essential assumption: see footnotes 5, 6 and 8 for the cases when we do not assume A3.

A4 assumes that the more bank 2 delays its payments, the more bank 1 can save on the cost of liquidity by delaying payments. If we assume $\beta = 0$ instead, the banks choose their optimal degree of delay d_i irrespective of their counterparty's delay.

A5 means that banks with higher cost of liquidity (ie larger ε_i) find payment delay more attractive. This is an intuitive assumption: if the liquidity cost is high, payment inflows (a source of free cash) become more valuable and the paying bank finds it optimal to delay its payments. This assumption plays an important role when we consider the welfare effect of various tariffs.

A6 is parallel to assumption A3. $\eta > 0$ implies that the negative externality of a counterparty's delay becomes larger if the bank's liquidity cost is high. This negative externality and the liquidity-saving effect γ (A5) also offset each other: see footnotes 5 and 6 for the cases when we ease this assumption.

A7 assumes that the intraday delay cost is increasing and convex. As is observed in many payment system, participants of payment systems aim to settle all payment obligations by the end of the day. As Committee on Payment and Settlement Systems (1997) notes, banks try to avoid delays in time-critical transfers. But at the end of the day, most (if not all) payment obligations become time-critical. This implies an increasing and convex delay cost function k. Furthermore, if the loss function V is concave, we always have corner solutions at $d_i = 0$ or $d_i = T$: this is not an interesting case as we will see below. For the sake of simplicity I further assume $\theta = 0$. See footnotes 5,6 and 8 for the cases when we do not assume $\theta = 0$.

3 Discretionary payment timing

The objective functions of banks are defined as follows:

$$\min_{d_1 \in [0,T]} \hat{l}_1(d_1, d_2, c + \varepsilon_1) + \hat{k}_1(d_1) - \lambda_1 d_1 - \lambda_3(T - d_1)$$
$$\min_{d_2 \in [0,T]} \hat{l}_2(d_2, d_1, c + \varepsilon_2) + \hat{k}_2(d_2) - \lambda_2 d_2 - \lambda_4(T - d_2)$$

$\hat{l}_i$ and $\hat{k}_i$ are the approximated loss functions and λ_j are Lagrangian multipliers that ensure $d_i \in [0, T]$. The Kuhn-Tucker first-order conditions of bank 1 are:

$$\alpha + \beta d_2 + \delta\varepsilon_1 + \mu d_1 - \lambda_1 + \lambda_3 = 0$$
$$\lambda_{1,3} \geq 0 \quad \text{and} \quad \lambda_1 d_1 = \lambda_3(T - d_1) = 0$$

The best response function d_1^{BR} is then derived:

$$d_1^{BR}(E[d_2], \varepsilon_1) = (-1)\frac{\alpha}{\mu} - \frac{\beta}{\mu} \cdot E\left[d_2(d_1, \varepsilon_2)\right] - \frac{\gamma}{\mu} \cdot \varepsilon_1 + \frac{1}{\mu} \cdot \lambda_1 - \frac{1}{\mu} \cdot \lambda_3 \qquad \textbf{(2)}$$

Since bank 1 cannot observe ε_2, the best response function becomes a function of expected $d_2(d_1, \varepsilon_2)$. The best response function of bank 2 is calculated in the same way.

Here we limit our attention to the internal solutions: $\lambda_h = 0$ for $h = \{1, 2, 3, 4\}$. (The cases of corner solutions are mentioned later.) At any equilibrium, $E\left[d_i(d_j, \varepsilon_i)\right] = E\left[d_i^{BR}(E[d_j], \varepsilon_i)\right]$ for any $i \neq j$. The expected best response functions are

$$\begin{aligned} E\left[d_1^{BR}(d_2, \varepsilon_1)\right] &= (-1)\frac{\alpha}{\mu} - \frac{\beta}{\mu} \cdot E\left[d_2^{BR}(d_1, \varepsilon_2)\right] \qquad \textbf{(3)} \\ E\left[d_2^{BR}(d_1, \varepsilon_2)\right] &= (-1)\frac{\alpha}{\mu} - \frac{\beta}{\mu} \cdot E\left[d_1^{BR}(d_2, \varepsilon_1)\right] \qquad \textbf{(4)} \end{aligned}$$

Solving the simultaneous equations, we have

$$E\left[d_1^{BR}(d_2, \varepsilon_1)\right] = E\left[d_2^{BR}(d_1, \varepsilon_2)\right] = (-1) \cdot \frac{\alpha}{\mu + \beta}$$

Substituting this into the best response functions, we have

$$\begin{aligned} d_1^*(E[d_2], \varepsilon_1) &= (-1) \cdot \frac{\alpha}{\mu + \beta} - \frac{\gamma}{\mu} \cdot \varepsilon_1 \\ d_2^*(E[d_1], \varepsilon_2) &= (-1) \cdot \frac{\alpha}{\mu + \beta} - \frac{\gamma}{\mu} \cdot \varepsilon_2 \end{aligned}$$

Since we assume $\lambda_h = 0$ for $h = \{1, 2, 3, 4\}$, $d_i^*(E[d_j], \varepsilon_i) \in (0, T)$ for any $i \neq j$. The condition to have this is, from the complementary slackness condition,

$$\begin{aligned}(-1) \cdot \frac{\alpha}{\mu + \beta} &> \frac{\gamma}{\mu} \cdot \varepsilon_1 \\ (-1) \cdot \frac{\alpha}{\mu + \beta} &< T + \frac{\gamma}{\mu} \cdot \varepsilon_1\end{aligned}$$

Rearranging this, for any i,

$$\frac{\mu}{\gamma} \cdot \frac{(-1) \cdot \alpha}{\mu + \beta} < \varepsilon_i < \frac{\mu}{\gamma} \cdot \frac{(-1) \cdot \alpha}{\mu + \beta} - \frac{\mu}{\gamma} \cdot T \qquad \textbf{(5)}$$

We denote the open interval as Ξ. Since $E[\varepsilon_i] = 0$ we need the lower bound (the left-hand side of the equation **(5)**) to be strictly negative and the upper bound (the right-hand side) to be strictly positive. These conditions are satisfied when $\alpha < 0$ and $\mu + \beta > 0$ (these are the conditions for the simultaneous equations **(3)** and **(4)** to have a unique equilibrium) and when T is sufficiently large. Now we have proved the following proposition.[5]

Proposition 1 If the assumptions A1-7 hold and if $\mu + \beta > 0$, then there exists a unique pure-strategy Bayesian Nash Equilibrium $d_i^*(E[d_j], \varepsilon_i) = (-1) \cdot \frac{\alpha}{\mu+\beta} - \frac{\gamma}{\mu} \cdot \varepsilon_i$ $(i \neq j)$, where $\varepsilon_i \in X \subset \Xi$.

See the appendix for the proof.

An important implication is that there is a constant term of delay $\frac{(-1)\cdot\alpha}{\mu+\beta} > 0$ irrespective of the idiosyncratic liquidity cost ε_i. This comes from the 'competition of delay' between banks: ie each bank tries to submit payments later than the counterparty. The 'competition', theoretically a variation of Cournot competition, can be clearly seen in the best response function (equation **(2)**), in which d_1^{BR} is a positive function of $E[d_2]$. The positive coefficient of $E[d_2]$ comes from the

[5] The assumptions A3 and A6 are unnecessary for the proposition. The assumption A7, $\theta = 0$, can be eased to $\theta \geq 0$, but we need to have $\alpha + \theta < 0$ to ensure the existence of internal solutions.

assumption A4 which provides an additional incentive for a bank to delay when its counterparty delays.

Banks postpone payments by $\frac{(-1)\cdot\alpha}{\mu+\beta}$ owing to a fear that they are preceded by their counterparties (delaying payments), and choose an additional delay if they themselves find it costly to obtain funds ($\frac{\gamma}{\mu} \cdot \varepsilon_i$). Banks shorten the delay if the delay cost is large (large μ), or if the liquidity cost is low (small α or ε_i). We will see that the 'competition of delay' is socially inefficient in the following section.

The constant term 'competition of delay' does not diverge to positive infinity, because at some stage the marginal benefit banks can obtain by paying later than their counterparties (β) is outweighed by the marginal cost of delay, μ. This is what the assumption $\mu + \beta > 0$ ensures. If $\mu + \beta \leq 0$, instead, banks always find it optimal to respond to their counterparties' delay by delaying even more, and the competition of delay continues until d_i reaches its upper bound T. As we will discuss later in Section 6, the corner solution is trivial since ε_i is irrelevant in equilibria.

4 Socially optimal payment timing

In the 'competition of delay' game, banks delay payments to minimise their loss functions, but in equilibrium, the delay becomes costly for both. This is because delay by a bank raises the other bank's cost of liquidity and thus incentivises the other bank's delay as a countermeasure. If both banks delay, the delay creates a deadweight loss.

In this section, we calculate a socially optimal outcome, minimising the aggregate loss of the banks, as a benchmark to clarify the deadweight loss. We will first derive the first-best solution, where there is no information asymmetry. We then derive another solution, under a more realistic assumption that a social planner cannot observe ε_i.

4.1 First-best solution

A social planner of the economy, which is possibly a central bank or a settlement agent, is assumed to minimise the aggregate loss of two banks 1 and 2. The social planner can observe ε_1 and ε_2. The social planner's loss function is thus defined as follows:

$$\min_{d_1,d_2} \hat{l}_1(d_1, d_2, c+\varepsilon_1) + \hat{k}_1(d_1) + \hat{l}_2(d_2, d_1, c+\varepsilon_2) + \hat{k}_2(d_2) \\ -\lambda_1 d_1 - \lambda_2 d_2 - \lambda_3(T-d_1) - \lambda_4(T-d_2) \tag{6}$$

By solving this, we have the following proposition.[6]

Proposition 2 The first-best solution d_i^{1st} is determined as follows:

$$d_i^{1st} = \max\left[\frac{\gamma}{\mu - 2\beta}\left(\varepsilon_j - \varepsilon_i\right), 0\right]$$

See appendix for the proof. The proposition gives rise to several important implications. First, the socially optimal delay is a function of the difference between the two banks' idiosyncratic costs of liquidity. Even if ε_1 is large, bank 1 makes the payments earlier if $\varepsilon_2 > \varepsilon_1$ (note that $\frac{\gamma}{\mu-2\beta} < 0$). The intuition is straightforward. If bank 1's liquidity cost is cheaper than that of bank 2, it is optimal for bank 1 to obtain more liquidity and to make its payments earlier than bank 2, in order to allow bank 2 to free-ride on the payment inflow from bank 1. Bank 1, as a result, has

[6] If A3 does not hold and if $\alpha + \delta < 0$, then $E[d_i^{1st}] > 0$ when $\mu + 2\beta > 0$ (a slightly stronger condition than the condition of Proposition 1, since the social planner counts the externality effects of two banks - the intuition of the condition is the same as the Proposition 1). This is because $\alpha + \delta < 0$ implies that the private benefit of own delay (α) outweighes its negative externality (δ). Ie delay is socially optimal until the (net) benefit is dominated by the increase of the delay cost (μ). We do not consider this case further since the assumption $\alpha + \delta < 0$ is counterintuitive.

If $\alpha + \delta > 0$ then d_i^{1st} has a negative constant term. It does change the overall argument below. $\theta > 0$ also provides a negative constant term.

Even if A6 does not hold and $\gamma + \eta \neq 0$, the main arguments below are unchanged as long as $\eta > 0$. See appendix for the detail.

to prepare a larger amount of intraday liquidity, but the aggregate cost of liquidity of the two banks is minimised since bank 1's liquidity cost is lower than bank 2.

Second, the socially efficient expected delay $E[d_i^{1st}]$ is equal to zero. In other words, on average no delay is socially optimal. This proves that the discretionary payment d_i^* is not socially desirable since it has a positive constant term. The suboptimality of the constant term is intuitive. For a given payment timing of its counterparty, a bank finds it optimal to delay and free-ride on the counterparty's liquidity knowing that its own delay increases the liquidity cost of its counterparty. The counterparty then also finds it optimal to delay further. The 'competition of delay' game continues until the marginal decline in the liquidity cost is outweighed by the increase of the delay cost. Since the two banks delay to the same extent in the (discretionary) equilibrium (Proposition 1), the banks cannot save any liquidity cost because free-riding is not possible. This is obviously socially inefficient: the 'competition of delay' ends up with a higher delay cost alone.

These implications tell us that, from an individual bank's point of view and from the social planner's point of view, settlement delay can be socially efficient when banks are heterogeneous ($\varepsilon_1 \neq \varepsilon_2$). The role-sharing arrangement that a bank with a lower funding cost prepares more liquidity and pays earlier works when banks are heterogeneous. And the socially desirable delay is a function of the difference of the liquidity cost ($\varepsilon_i - \varepsilon_j$). On the other hand, the propositions above also specify an inefficient component of settlement delay; the constant 'competition of delay' part.

4.2 When the social planner has imperfect information

The assumption that the social planner can observe individual bank i's idiosyncratic cost ε_i is, however, infeasible in most cases. Central banks may be able to observe the average cost of liquidity among banks, which is the market rate, but it is at least difficult, if not impossible, to monitor the liquidity cost of each one of the banks.

If the idiosyncratic cost ε_i is unobservable for the social planner, it can at most minimise the expected cost of the aggregated loss function:

$$\min_{d_1,d_2} E\left[\hat{l}_1(d_1,d_2,c+\varepsilon_1)+\hat{k}_1(d_1)+\hat{l}_2(d_2,d_1,c+\varepsilon_2)+\hat{k}_2(d_2)\right]$$
$$-\lambda_1 d_1-\lambda_2 d_2-\lambda_3(T-d_1)-\lambda_4(T-d_2)$$

The solution of the problem is almost identical to the first-best case, and we have the following:

Proposition 3 $d_i^{feasible}=0$ for any $i\in\{1,2\}$.

Theoretically, the social planner may be able to design a mechanism that provides an incentive for banks to reveal their private information (ε_i). This may be a better solution than imposing $d_i^{feasible}=0$, but we do not consider this possibility. This is because the social planner would have to design a complicated system whereby payment timing and the tariff are adjusted based on banks' announcement of their private information (possibly every day). Considering the current arrangements in payment systems, such a complicated arrangement is infeasible.

$d_i^{feasible}$ provides a better outcome than the discrete d_i^* in the sense that $d_i^{feasible}$ eliminates the inefficient 'competition of delay', but is less efficient than d_i^{1st} and d_i^* in the sense that it ignores the idiosyncratic factors $\{\varepsilon_i\}_{i=1,2}$.

5 Incentive mechanisms for early payments

In this section, we seek means to eliminate the inefficiency identified in the previous section. Two options are discussed here. The first one is throughput guidelines, which force banks to submit a fraction of their payments by a certain intraday deadline. CHAPS in the United Kingdom, CHIPS in the United States and some other payment systems adopt this option. The second one is a time-varying tariff, which charges a higher processing fee for payments that are submitted for settlement later on the day. SIC in Switzerland and STR in Brazil have introduced this option. We will compare the options and also find the optimal design of the time-varying tariff.

There are three important assumptions underlying the discussions in this section. First, the social planner cannot observe $\varepsilon_{1,2}$. Second, the social planner cannot enforce the payment action d_i directly. Third, the social planner cannot design a mechanism that is a function of ε_i.

5.1 Throughput guidelines

Throughput guidelines require banks to make some fraction of a given day's payments by certain intraday deadlines.[7] Normally there is no pecuniary penalty for breach of the guidelines. Reputation loss is considered to be sufficient to enforce banks' early payments.

By imposing sufficiently strong penalties for violation, the social planner (a settlement agent or a central bank) can enforce $d_1^{feasible}$. In other words, throughput guidelines can be interpreted as a system by which the social planner, who is not able to observe ε_i, aims to maximise the welfare of the system. The guidelines have some drawbacks, however.

The first reason is equivalent to the drawback of $d_1^{feasible}$ discussed in the previous section. Since the enforced payment timing is independent of bank specific factors (ε_i), there is no flexibility for the banks to adjust their payment behaviours based on ε_i. We have seen in the previous section that it is socially optimal for a bank with a lower liquidity cost to pay early and for a bank with a higher liquidity cost to pay later, as this minimises the aggregate cost of liquidity. The throughput guidelines do not allow for such flexibility.

Second, there is no incentive for banks to make payments well ahead of the deadline. Since there is no penalty for 'last-minute' payments made shortly before the deadline, it is optimal for the banks to make payments an instant before the deadline: this creates bunching of payments. We do not discuss this inefficiency here in order to focus on the flexibility issue: the deadline is assumed to be $d_i = 0$ as shown above, so no 'delay before deadlines' can be made.

[7] To be precise, in the United Kingdom, the member banks are required to satisfy the guideline on a monthly average, not on a daily basis. But the same argument can be applied to the UK guidelines as well, by assuming that ε is unchanged throughout the month (alternatively, we can interpret the model as studying banks' behaviour on the last day of the month). See Section 6 for further discussion on this issue.

5.2 *Time-varying tariff*

An alternative mechanism to incentivise early payments is a time-varying tariff: if banks pay late, they have to pay a higher processing fee to the system operator. We denote the tariff as a function $t(d_i)$, which is increasing in d_i. The objective function is now defined as follows (we do not restrict d_i from above, because it is obviously bounded above):

$$\min_{d_1} \hat{l}_1(d_1, d_2, c + \varepsilon_1) + \hat{k}_1(d_1) + t(d_1) - \lambda_1 d_1$$
$$\min_{d_2} \hat{l}_2(d_2, d_1, c + \varepsilon_2) + \hat{k}_2(d_2) + t(d_2) - \lambda_2 d_2$$

Solving these, and determining $t(d_i)$ to eliminate the constant strategic delay term in Proposition 1, we have the following:[8]

Proposition 4 An optimal tariff function is $t(d_i) = t_0 - \alpha d_i$. The equilibrium delay is $d_i^{**} = \max\left[-\frac{\gamma}{\mu}\varepsilon_i, 0\right]$.

See the appendix for the proof. t_0 is an integral constant.

Since we assume $\alpha < 0$, $t(d_i)$ is increasing in d_i. α is a decreasing function of c, thus the tariff should be designed to be steeper (ie the tariff should increase quickly over time) if the average liquidity cost c is high. And since α is a constant parameter, the tariff should increase linearly

[8] If $\theta > 0$ (easing assumption A7) or $\alpha + \delta \neq 0$ (easing A3), the optimal tariff function takes a slightly different form, because the expected first-best solution has a non-zero constant term, as we have seen in footnote 6.

Focusing on the case $\alpha + \delta \neq 0$, d_i^{1st} has a constant term $\frac{(-1)(\alpha+\delta)}{\mu+2\beta}$. The optimal tariff function has to incorporate this, so we have the following:
$t(d_i) = t_0 - \left\{\alpha - .\frac{(\alpha+\delta)(\mu+\beta)}{\mu+2\beta}\right\} \cdot d_i$
This is still a linear function against d_i, and the characteristics of the function are unchanged.

Assumption A6 is irrelevant to the proposition.

over time. This suggest, however, that an exponentially increasing tariff, such as the SNB adopts, may excessively penalise later payers.

A linearly increasing tariff has several characteristics. First, the tariff eliminates only the delay coming from the inefficient 'competition of delay' (see Section 2) and allows banks to adjust their payment timing in accordance with the bank's own idiosyncratic factors ε_i. Second, the settlement agent does not have to tailor the tariff for each bank: the same tariff enforces the different optimal payment timing according to banks' own ε_i.

Third, there is a straightforward mechanism to help ensuring the cost recovery of the system. The settlement agent is usually required not to make any profit, and also not to make any loss. Cost recovery can be achieved by adjusting the constant term t_0. t_0 is an integral constant so any t_0 does not change the incentive structure we have discussed. If the system makes an excess profit, the agent can lower t_0 (or, return the equivalent value to the member banks).

In addition, if t_0 is appropriately chosen to achieve zero profit, the time-varying tariff acts as an income transfer mechanism, from late payers to early payers. This is the essence of the tariff. The fact that cash can be obtained for free by receiving payment inflows incentivises settlement delay in RTGS. The tariff forces the recipient (ie late payer) to pay some liquidity cost of the payment inflow to the payer (ie early payer). The recipient who can recycle the inflow for free should be happy to pay the fee, as long as its own liquidity cost is higher than the fee: and thus this transfer from late payers to early payers improves efficiency.

5.3 Performance of the two options

The tariff, however, cannot achieve the first best (Proposition 2). This is because each bank i knows its own private information ε_i alone, and thus cannot determine its payment timing d_i based on counterparty j's ε_j. Furthermore, the social planner needs to design the tariff function independently of the unobservable variables ε_1 and ε_2. Ie the tariff function cannot change the coefficient of ε_i in Proposition 1; nevertheless the coefficient of ε_i is different from the first-best case (see Proposition 2).

In this section, we compare the aggregated loss function associated with the tariff and with the throughput guidelines, to confirm the efficiency of the tariff. In other words, we will check the following inequality, meaning that the expected aggregate loss associated with the tariff is smaller than the loss with the throughput guideline:

$$E\left[L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)\right] < E\left[L(0, 0, c, \varepsilon_1, \varepsilon_2)\right] = 0$$

where $L(d_1, d_2, c, \varepsilon_1, \varepsilon_2) = \hat{l}_1(d_1, d_2, c + \varepsilon_1) + \hat{k}_1(d_1) + \hat{l}_2(d_2, d_1, c + \varepsilon_2) + \hat{k}_2(d_2)$ is the aggregate loss of the system given payment timings and liquidity costs. The left-hand side is the loss if the tariff is implemented, and the right-hand side is the loss with the throughput guideline and the deadline $d_i = 0$. Since d_i^{**} is not differentiable at $d_i = 0$ (ie at $\varepsilon_i = 0$) we need to consider four different cases to calculate $E\left[L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)\right]$ as follows:

$$\begin{aligned} E\left[L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)\right] = & \int_0^{\bar{\varepsilon}} \int_0^{\bar{\varepsilon}} L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) \psi(\varepsilon_1) \psi(\varepsilon_2) d\varepsilon_1 d\varepsilon_2 \\ & + \int_0^{\bar{\varepsilon}} \int_{-\bar{\varepsilon}}^{0} L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) \psi(\varepsilon_1) \psi(\varepsilon_2) d\varepsilon_1 d\varepsilon_2 \\ & + \int_{-\bar{\varepsilon}}^{0} \int_0^{\bar{\varepsilon}} L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) \psi(\varepsilon_1) \psi(\varepsilon_2) d\varepsilon_1 d\varepsilon_2 \\ & + \int_{-\bar{\varepsilon}}^{0} \int_{-\bar{\varepsilon}}^{0} L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) \psi(\varepsilon_1) \psi(\varepsilon_2) d\varepsilon_1 d\varepsilon_2 \end{aligned} \quad \textbf{(7)}$$

where ψ is the probability density function of ε_i. By substituting $d_i^{**} = \max\left[-\frac{\gamma}{\mu}\varepsilon_i, 0\right]$, we have the following proposition.

Proposition 5 The expected aggregate loss under the tariff $E\left[L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)\right]$ is strictly smaller than the one under the throughput guidelines $E\left[L(0, 0, c, \varepsilon_1, \varepsilon_2)\right]$.

See the appendix for the proof. The proof has some interesting implications for the mechanism of the tariff. Equation **(A-1)** in the appendix shows that the tariff can underperform the throughput guidelines if ε_1 and ε_2 are both positive and take a similar value. Suppose that

$\varepsilon_1 = \varepsilon_2 > 0$. Two banks delay their payments expecting that their counterparties pay earlier, but they cannot save the liquidity cost because both delay to the same extent. In this case, the tariff can be worse than throughput guidelines which simply force the banks to make payments at $d_i = 0$. Even if ε_1 and ε_2 are both strictly positive, the tariff can improve efficiency if the distance between ε_1 and ε_2 is large. This is because the main benefit of the tariff lies in allowing a flexible role-sharing arrangement when banks are heterogeneous.

A potential inefficiency comes from the assumption that banks cannot observe their counterparty's idiosyncratic cost of liquidity ε_i. In the first-best solution, delay is a function of the difference of the idiosyncratic costs and thus the possibility of inefficient delay is eliminated. If the signs of ε_1 and ε_2 are different, the tariff works better than the throughput guidelines for sure. This is because the different signs ensure the heterogeneity of the banks.

The tariff underperforms most when both ε_1 and ε_2 take a similar large positive value. This could happen when the banks are experiencing a large-scale liquidity shortage, in other words, a liquidity crisis. In these circumstances, the tariff fails to encourage early payments. Potential policy options during crises are discussed in the following section.

6 Discussion

6.1 How the tariff works

The policy implications of this paper are summarised as follows.

The model shows that the time-varying tariff works better than throughput guidelines. This is because the tariff allows for efficient delay, ie the role-sharing arrangement between heterogeneous banks, while the guidelines do not. Both however eliminate the inefficient 'competition of delay' component. In other words, the tariff leads to better intraday liquidity management. By making banks choose their payment timings, the social planner (eg the central bank) can allow efficient delay conditional on each bank's liquidity cost, which is unobservable for the planner. It is, however, not the first-best solution with symmetric information, because each bank can observe only its own cost and thus assumes that the counterparty's hidden idiosyncratic cost is zero (the expected value of the cost). Especially when two banks are nearly identical, the optimal tariff creates inefficient delay in some cases. This is because banks delay to a similar extent and fail to save the liquidity cost, while their delay costs increase. The optimal tariff, however, works better on average than throughput guidelines since the welfare improvement in the cases when the banks are heterogeneous dominates the welfare loss in the cases of nearly homogeneous banks.

The optimal tariff has simple features that make it easier to implement. First, it should increase linearly over time, as long as the loss function is convex. (If the loss function is concave, a wide variety of incentive mechanisms can be optimal and the curvature of the tariff does not matter.) A convex tariff may excessively penalise late payers. Second, the constant term of the tariff function, t_0, can be freely adjusted to help ensure cost recovery (zero profit) for the settlement agent, which is normally required (as discussed in Section 5.2). Third, the same tariff function can be applied to all (possibly heterogeneous) banks. The same function can support different optimal payment timings, in accordance with their hidden idiosyncratic liquidity costs. The social planner needs to consider the market-wide (average) liquidity cost c alone.

There is a situation, however, where the tariff underperforms throughput guidelines. As shown in the previous section, if both of the banks experience a large hike of their idiosyncratic liquidity costs, the tariff cannot work properly to eliminate the inefficient delay. In the worst cases where the costs are extremely high, the banks' optimal payment timing d_i reaches to the closing time T of the payment system, which can be labelled as 'gridlock'.[9]

There are two conceivable options to the gridlock. The first option is a steeper tariff that incentivises some banks to obtain intraday liquidity to resolve the gridlock, although changing tariffs during a day may not be feasible. Another option is to provide more intraday liquidity. Since the slope of the tariff is a function of the market-wide liquidity cost c, the necessity of a steeper tariff means that the market-wide intraday liquidity cost is increased during the gridlock period. The social planner can lower this cost by providing intraday liquidity.

[9] See Bank for International Settlements (1993), for example, for the detail of gridlock. This paper does not specifically study gridlock, a corner solution, since the assumption made at the Proposition 1 ensures internal solution.

The tariff is the second-best solution in this model. It is inferior to the social planner's first-best d_i^{1st}, because the tariff only utilises a bank i's own idiosyncratic cost of liquidity, ε_i and assumes the counterparty j's cost is zero, which is the expected value of ε_j. The first-best solution is a function of the difference of the costs. But as Proposition 5 shows, the tariff is superior to the throughput guidelines because the guidelines ignore the idiosyncratic costs. But all of these options are better than the discretional payment behaviour in the Proposition 1, since the 'competition of delay' increases the cost of delay without reducing the cost of liquidity, which is inefficient.

An important issue in the implementation of the tariff is how to determine the optimal slope. Since Proposition 4 tells the optimal slope of the tariff is a function of the cost of intraday liquidity, the optimal slope can be approximated from the interest rate of intraday liquidity. There are several empirical studies on this. Furfine (2001) estimates that increasing the duration of the US Federal funds loans (uncollateralised) by one hour raises the interest rate by 0.9 basis point. Baglioni and Monticini (2008), using an Italian tick-by-tick payments data, estimate that the hourly price (annualised) of money is 0.44 basis point. Kraenzlin and Nellen (2010) obtain a similar number (from 0.4 to 0.5 basis point), using Switzerland's data. On a related issue, Fedwire charges 36 basis point (annualised, 24 hours basis) for a daylight overdraft. This hourly fee (annualised) is obtained by $36 \cdot \frac{1}{24} = 1.5$ basis point.[10] As Furfine (2001) argues, the hourly cost of intraday liquidity is unlikely to be higher. This is because the obtained hourly fee ignores exemption (see footnote 13), and the fee structure is asymmetric (charged only for positive overdraft balance, zero for the negative balance).

Using these estimates and Fedwire data, we can provide an illustrative example of the liquidity-saving effect resulting from delaying a payment. Assume that a bank postpones a $4 million payment (the average size of payments in Fedwire in the last decade) from 9 am till noon (3 hours) and assume further that the bank can free-ride on a payment inflow from a counterparty for the same amount ($4 million). If the hourly cost of intraday liquidity is 0.4, 0.9 and 1.5 basis point, the liquidity-saving effect of the three hours delay is $1.3, $3.0 and $4.9 respectively, which approximate the marginal tariff it should increase from 9 am till noon.

[10] To be precise, the actual fee paid is the gross fee minus a lump-sum exemption, so the effective hourly fee is smaller than 1.5 basis point.

Since the Fedwire's per-item fee for each domestic payment is \$0.3,[11] the constant term of the optimal tariff (t_0) may be a negative number for cost recovery if the tariff increases by \$1.3, \$3.0 or \$4.9 in 3 hours as illustrated above. Since the liquidity-saving effect of delay becomes larger as the size of payments increases, the optimal tariff would be negative in the morning and positive in the evening for some payment systems settling significantly large-value payments. The early payers, as a result, look like they are making money by joining the payment system — but it is an appropriate compensation of early payments (in other words, early payers are losing money without such kinds of tariffs). If a payment system mainly settles small-value payments such as SIC, the optimal tariff would take a small positive value in the morning and slowly increasing during a day. Another implication is that the tariff could be charged for the value of payments, not for the number of payment items, because the cost of liquidity is a function of value, not volume.

Another important outcome of the time-varying tariff is the prioritisation of smaller payments, as Rochet and Tirole (1996) argue. Since payment systems normally charge fees for each payment, independent of the size of these payments, banks find it optimal to make small payments first and large payments later under time-varying tariffs, as is observed in SIC. Rochet and Tirole (1996) regard this 'pre-sorting' as a beneficial liquidity-saving arrangement. But the model discussed here cannot capture the 'pre-sorting' mechanism, since the set of payments is abstract in the model.

6.2 Implications of the model

The model shows that settlement delay is not always inefficient, contrary to general understanding of the literature. If banks in a payment system are heterogeneous in terms of their liquidity costs, it can be socially efficient for banks with a higher liquidity cost to delay payments. Banks with a lower cost of liquidity can obtain more intraday liquidity and pay earlier, while the high-cost banks can free-ride on the cash received from low-cost banks. The delay by the high-cost banks therefore improves social welfare. In other words, the settlement delay by high-cost banks allows for a socially efficient role-sharing between high-cost banks and low-cost banks.

[11] For the first 14,000 transfers per month.

On the other hand, as discussed in Section 4, the 'competition of delay' (the constant term of Proposition 1) is socially inefficient, because banks cannot save any liquidity cost when they delay while the delay costs increase. Eliminating this component improves social welfare.

In addition, the equilibrium payment timing is not a corner solution. Previous two-period models of payment systems implicitly focus on corner solutions (the beginning of the morning and the end of the evening). Since the equilibrium in this paper is an internal solution, the optimal payment timing is flexibly adjusted as the costs of liquidity and delay change. This feature is convenient when we study, for example, how payment timing can be adjusted by a fall of liquidity cost.

If we have a corner solution, eg because $\mu + \beta \leq 0$, changes of ε no longer matter — the optimal d is 0 or T, irrespective of the level of ε_i. In this case, we do not need to consider an optimal mechanism to incentivise early payments. We just need to force $d = 0$ by all means. We do not discuss the case further since it is non-interesting.[12]

6.3 Directions for future study

There are several possible extensions. The paper assumes that banks know the amount of their payment obligations *ex ante*, and thus can determine the delay index d for sure. But if there are some unexpected payments requests during a day, banks cannot perfectly control d: d may have a random error in this case. For instance, the delay d would be determined by the bank's choice $\hat{d}$ plus a white noise term s. With noise, a trigger strategy (by which the violation of a predetermined deadline triggers a fixed penalty), such as throughput guidelines, may work well, as Walsh (2002) shows for instance. This is left for future study.

Another potential extension would be to assume that breaches of the throughput guidelines are penalised only when a bank repeatedly violates the guidelines in a month, as currently CHAPS does in the United Kingdom. This feature also may give some flexibility to banks in the system because they can pay later one day and pay earlier another day according to their situations. This

[12] The corner solution can be considered to be similar to Bech and Garratt (2003) in continuous time scale.

extension is more difficult to model from a pure-theoretical point of view (see, eg, Matsushima (2004)).

From a policy perspective, it is an interesting question how the optimal tariff changes when a central bank introduces a liquidity-saving mechanism or a new scheme for intraday liquidity provision is introduced as the Fed is currently working on. The model discussed here provides several implications on these. Liquidity-saving mechanisms are likely to reduce the cost of intraday liquidity, the optimal slope of time-varying tariff would be flatter. If a new policy for intraday liquidity provision raises the cost of liquidity, the optimal slope will be steeper. A detailed study is needed, however, to discuss these issues further.

Appendix: Proofs

Proof of Proposition 2

$$\hat{l}_1 = \alpha d_1 + \beta d_1 d_2 + \gamma d_1 \varepsilon_1 + \delta d_2 + \eta d_2 \varepsilon_1$$

$$\hat{l}_2 = \alpha d_2 + \beta d_2 d_1 + \gamma d_2 \varepsilon_2 + \delta d_1 + \eta d_1 \varepsilon_2$$

The first-order conditions of the function **(6)** are:

$$\alpha + \beta d_2 + \gamma \varepsilon_1 + \mu d_1 + \delta + \beta d_2 + \eta \varepsilon_2 - \lambda_1 + \lambda_3 = 0$$

$$\alpha + \beta d_1 + \gamma \varepsilon_2 + \mu d_2 + \delta + \beta d_1 + \eta \varepsilon_1 - \lambda_2 + \lambda_4 = 0$$

Solving these:

$$d_1 = \frac{(-1)\alpha}{\mu} - \frac{\delta}{\mu} - \frac{\gamma}{\mu}\varepsilon_1 - \frac{\eta}{\mu}\varepsilon_2 - \frac{2\beta}{\mu}d_2 + \frac{\lambda_1 - \lambda_3}{\mu}$$

d_2 is defined in the same way. Substituting these and assume $\lambda_j = 0$ for all $j \in \{1, 2, 3, 4\}$, we have:

$$d_1 = \frac{(-1)(\alpha + \delta)}{\mu} - \frac{\gamma}{\mu}\varepsilon_1 - \frac{\eta}{\mu}\varepsilon_2 - \frac{2\beta}{\mu}\left\{\frac{(-1)\alpha}{\mu} - \frac{\delta}{\mu} - \frac{\eta}{\mu}\varepsilon_1 - \frac{\gamma}{\mu}\varepsilon_2 - \frac{2\beta}{\mu}d_1\right\}$$

Solving for d_1,

$$d_1 = \frac{(-1)(\alpha + \delta)}{\mu + 2\beta} + \frac{2\beta\gamma - \mu\eta}{\mu^2 - 4\beta^2}\varepsilon_2 + \frac{2\beta\eta - \mu\gamma}{\mu^2 - 4\beta^2}\varepsilon_1$$

From A3 and A6, we have:

$$\begin{aligned} d_1 &= \frac{2\beta\gamma + \mu\gamma}{\mu^2 - 4\beta^2}\varepsilon_2 + \frac{-2\beta\gamma - \mu\gamma}{\mu^2 - 4\beta^2}\varepsilon_1 \\ &= \frac{\gamma}{\mu - 2\beta}(\varepsilon_2 - \varepsilon_1) \end{aligned}$$

The first-best solution is, therefore,

$$d_1^{1st} = \max\left[\frac{\gamma}{\mu - 2\beta}(\varepsilon_2 - \varepsilon_1), 0\right]$$

$\frac{\gamma}{\mu-2\beta} < 0$ thus d_1^{1st} is a positive function of ε_1 and a negative function of ε_2. λ_1 and λ_2 can be non-zero if $d_1^{1st} = 0$: the derivation of these is skipped. Since we have assumed a sufficiently large T, $\lambda_3 = \lambda_4 = 0$ in any case.

Proof of Proposition 4

We keep the assumptions made in Proposition 1, which ensures internal solutions, but we still need to formulate the Kuhn-Tucker conditions because we have the tariff function.

$$\alpha + \mu d_1 + \beta d_2 + \gamma \varepsilon_1 + \frac{\partial t}{\partial d_1} - \lambda_1 = 0$$

$$\lambda_1 \geq 0 \quad \text{and} \quad \lambda_1 d_1 = 0$$

The best response function d_1^{BR} is then derived.

$$d_1^{BR}(E[d_2], \varepsilon_1) = (-1)\frac{\alpha}{\mu} - \frac{\beta}{\mu} \cdot E\left[d_2(d_1, \varepsilon_2)\right] - \frac{\gamma}{\mu}\varepsilon_1 - \frac{1}{\mu}\left(\frac{\partial t}{\partial d_1} - \lambda_1\right)$$

Substitute this into the equation $E[d_2^{BR}(E[d_1], \varepsilon_2)]$, which is defined in a similar way, and we have:

$$E[d_2^{BR}(E[d_1], \varepsilon_2)] = \frac{(-1)\alpha}{\mu + \beta} + \frac{\beta}{\mu^2 - \beta^2}\left(\frac{\partial t}{\partial d_1} - \lambda_1\right) - \frac{\mu}{\mu^2 - \beta^2}\left(\frac{\partial t}{\partial d_2} - \lambda_2\right)$$

Likewise, we have:

$$E[d_1^{BR}(E[d_2], \varepsilon_1)] = \frac{(-1)\alpha}{\mu + \beta} + \frac{\beta}{\mu^2 - \beta^2}\left(\frac{\partial t}{\partial d_2} - \lambda_2\right) - \frac{\mu}{\mu^2 - \beta^2}\left(\frac{\partial t}{\partial d_1} - \lambda_1\right)$$

We will design the function t to eliminate the source of deadweight loss $\frac{(-1)\alpha}{\mu+\beta}$, so we have the following simultaneous equations:

$$\frac{\alpha}{\mu + \beta} = \frac{\beta}{\mu^2 - \beta^2}\left(\frac{\partial t}{\partial d_2} - \lambda_2\right) - \frac{\mu}{\mu^2 - \beta^2}\left(\frac{\partial t}{\partial d_1} - \lambda_1\right)$$

$$\frac{\alpha}{\mu + \beta} = \frac{\beta}{\mu^2 - \beta^2}\left(\frac{\partial t}{\partial d_1} - \lambda_1\right) - \frac{\mu}{\mu^2 - \beta^2}\left(\frac{\partial t}{\partial d_2} - \lambda_2\right)$$

By solving these equations, we have

$$\frac{\partial t}{\partial d_1} - \lambda_1 = \frac{\partial t}{\partial d_2} - \lambda_2 = (-1)\alpha$$

Integrating these, and $\lambda_i \cdot d_i = 0$, we have:

$$t(d_i) = t_0 - \alpha d_i$$

Proof of Proposition 5

The welfare of the system under the tariff $t(d_i)$ for given ε_1 and ε_2, $L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)$, is defined as follows:

$$
\begin{aligned}
& L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) \\
= \; & \alpha \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_1, 0\right] + \beta \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_1, 0\right] \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_2, 0\right] + \gamma \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_1, 0\right] \cdot \varepsilon_1 \\
& +\delta \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_2, 0\right] + \eta \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_2, 0\right] \cdot \varepsilon_1 + \frac{1}{2}\mu \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_1, 0\right]^2 \\
& +\alpha \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_2, 0\right] + \beta \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_2, 0\right] \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_1, 0\right] + \gamma \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_2, 0\right] \cdot \varepsilon_2 \\
& +\delta \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_1, 0\right] + \eta \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_1, 0\right] \cdot \varepsilon_2 + \frac{1}{2}\mu \cdot \max\left[-\frac{\gamma}{\mu}\varepsilon_2, 0\right]^2
\end{aligned}
$$

If $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, we have:

$$
\begin{aligned}
L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) &= \alpha\frac{-\gamma}{\mu}\varepsilon_1 + \beta\frac{-\gamma}{\mu}\varepsilon_1\frac{-\gamma}{\mu}\varepsilon_2 + \gamma\,\frac{-\gamma}{\mu}\varepsilon_1\varepsilon_1 \\
&\quad +\delta\frac{-\gamma}{\mu}\varepsilon_2 + \eta\frac{-\gamma}{\mu}\varepsilon_2\varepsilon_1 + \frac{1}{2}\mu\left(\frac{-\gamma}{\mu}\varepsilon_1\right)^2 \\
&\quad +\alpha\frac{-\gamma}{\mu}\varepsilon_2 + \beta\frac{-\gamma}{\mu}\varepsilon_2\frac{-\gamma}{\mu}\varepsilon_1 + \gamma\,\frac{-\gamma}{\mu}\varepsilon_2\varepsilon_2 \\
&\quad +\delta\frac{-\gamma}{\mu}\varepsilon_1 + \eta\frac{-\gamma}{\mu}\varepsilon_1\varepsilon_2 + \frac{1}{2}\mu\left(\frac{-\gamma}{\mu}\varepsilon_2\right)^2 \\
&= 2\frac{\beta\gamma^2}{\mu^2}\varepsilon_1\varepsilon_2 - 2\frac{\eta\gamma}{\mu}\varepsilon_1\varepsilon_2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_1^2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_2^2 \\
&\quad - (\alpha+\delta)\frac{\gamma}{\mu}\varepsilon_1 - (\alpha+\delta)\frac{\gamma}{\mu}\varepsilon_2
\end{aligned}
$$

From A3 ($\alpha + \delta = 0$) and A6 ($\gamma + \eta = 0$), we have:

$$
\begin{aligned}
L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) &= 2\gamma^2\frac{\beta+\mu}{\mu^2}\varepsilon_1\varepsilon_2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_1^2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_2^2 \\
&= \frac{\gamma^2}{\mu}\left\{2\frac{\beta+\mu}{\mu}\varepsilon_1\varepsilon_2 - \frac{1}{2}\varepsilon_1^2 - \frac{1}{2}\varepsilon_2^2\right\} \qquad \textbf{(A-1)}
\end{aligned}
$$

Notice that $L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)$ can be positive when $\frac{\beta+\mu}{\mu} > \frac{1}{2}$ and $\varepsilon_1 \simeq \varepsilon_2 > 0$. If $\varepsilon_1 > 0$ and $\varepsilon_2 \leq 0$, $L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)$ is:

$$\begin{aligned}
L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) &= \alpha \frac{-\gamma}{\mu}\varepsilon_1 + \gamma \frac{-\gamma}{\mu}\varepsilon_1\varepsilon_1 + \frac{1}{2}\mu\left(\frac{-\gamma}{\mu}\varepsilon_1\right)^2 \\
&\quad +\delta\frac{-\gamma}{\mu}\varepsilon_1 + \eta\frac{-\gamma}{\mu}\varepsilon_1\varepsilon_2 \\
&= \frac{-\gamma}{\mu}(\alpha+\delta)\varepsilon_1 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_1^2 + \eta\frac{-\gamma}{\mu}\varepsilon_1\varepsilon_2
\end{aligned}$$

From A3 ($\alpha + \delta = 0$) and A6 ($\gamma + \eta = 0$), we have:

$$\begin{aligned}
L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) &= \frac{\gamma^2}{\mu}\varepsilon_1\varepsilon_2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_1^2 \\
&\leq 0
\end{aligned}$$

Since the banks are symmetric, we have a similar welfare for the case $\varepsilon_1 \leq 0$ and $\varepsilon_2 > 0$.

$$\begin{aligned}
L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2) &= \frac{\gamma^2}{\mu}\varepsilon_1\varepsilon_2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_2^2 \\
&\leq 0
\end{aligned}$$

If $\varepsilon_1 \leq 0$ and $\varepsilon_2 \leq 0$, $L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)$ is trivially zero.

Substituting these into the equation **(7)**:

$$\begin{aligned}
E\left[L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)\right] &= \int_0^{\bar{\varepsilon}}\int_0^{\bar{\varepsilon}} \left\{2\gamma^2\frac{\beta+\mu}{\mu^2}\varepsilon_1\varepsilon_2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_1^2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_2^2\right\}\psi(\varepsilon_1)\psi(\varepsilon_2)d\varepsilon_1 d\varepsilon_2 \\
&\quad + \int_0^{\bar{\varepsilon}}\int_{-\bar{\varepsilon}}^{0} \left\{\frac{\gamma^2}{\mu}\varepsilon_1\varepsilon_2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_1^2\right\}\psi(\varepsilon_1)\psi(\varepsilon_2)d\varepsilon_1 d\varepsilon_2 \\
&\quad + \int_{-\bar{\varepsilon}}^{0}\int_0^{\bar{\varepsilon}} \left\{\frac{\gamma^2}{\mu}\varepsilon_1\varepsilon_2 - \frac{1}{2}\frac{\gamma^2}{\mu}\varepsilon_1^2\right\}\psi(\varepsilon_1)\psi(\varepsilon_2)d\varepsilon_1 d\varepsilon_2
\end{aligned}$$

The first term can be written as follows:

$$
\begin{aligned}
& \int_0^{\bar{\varepsilon}} \int_0^{\bar{\varepsilon}} \left\{ 2\gamma^2 \frac{\beta+\mu}{\mu^2} \varepsilon_1 \varepsilon_2 - \frac{1}{2} \frac{\gamma^2}{\mu} \varepsilon_1^2 - \frac{1}{2} \frac{\gamma^2}{\mu} \varepsilon_2^2 \right\} \frac{1}{4\bar{\varepsilon}^2} d\varepsilon_1 d\varepsilon_2 \\
= & \int_0^{\bar{\varepsilon}} \left[\gamma^2 \frac{\beta+\mu}{\mu^2} \varepsilon_1 \varepsilon_2^2 - \frac{1}{2} \frac{\gamma^2}{\mu} \varepsilon_1^2 \varepsilon_2 - \frac{1}{6} \frac{\gamma^2}{\mu} \varepsilon_2^3 \right]_0^{\bar{\varepsilon}} \frac{1}{4\bar{\varepsilon}^2} d\varepsilon_1 \\
= & \frac{1}{4\bar{\varepsilon}^2} \left[\frac{1}{2} \gamma^2 \frac{\beta+\mu}{\mu^2} \varepsilon_1^2 \bar{\varepsilon}^2 - \frac{1}{6} \frac{\gamma^2}{\mu} \varepsilon_1^3 \bar{\varepsilon} - \frac{1}{6} \frac{\gamma^2}{\mu} \varepsilon_1 \bar{\varepsilon}^3 \right]_0^{\bar{\varepsilon}} \\
= & \frac{1}{24} \frac{\gamma^2}{\mu} \left\{ 3 \frac{\beta+\mu}{\mu} - 2 \right\} \bar{\varepsilon}^2
\end{aligned}
$$

The second term is:

$$
\begin{aligned}
& \int_0^{\bar{\varepsilon}} \left[\frac{1}{2} \frac{\gamma^2}{\mu} \varepsilon_1 \varepsilon_2^2 - \frac{1}{2} \frac{\gamma^2}{\mu} \varepsilon_1^2 \varepsilon_2 \right]_{\bar{\varepsilon}}^0 \frac{1}{4\bar{\varepsilon}^2} d\varepsilon_1 \\
= & \frac{1}{4\bar{\varepsilon}^2} \left[\frac{1}{6} \frac{\gamma^2}{\mu} \varepsilon_1^3 \bar{\varepsilon} - \frac{1}{4} \frac{\gamma^2}{\mu} \varepsilon_1^2 \bar{\varepsilon}^2 \right]_0^{\bar{\varepsilon}} \\
= & \frac{1}{48} \frac{\gamma^2}{\mu} (-1) \bar{\varepsilon}^2
\end{aligned}
$$

The third term is equivalent to the second one. Now we have:

$$
\begin{aligned}
E\left[L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)\right] &= \frac{1}{24} \frac{\gamma^2}{\mu} \left\{ 3 \frac{\beta+\mu}{\mu} - 2 \right\} \bar{\varepsilon}^2 + 2 \cdot \frac{1}{48} \frac{\gamma^2}{\mu} (-1) \bar{\varepsilon}^2 \\
&= \frac{1}{24} \frac{\gamma^2}{\mu} \bar{\varepsilon}^2 \left(3 \frac{\beta+\mu}{\mu} - 2 - 1 \right)
\end{aligned}
$$

Since $\beta < 0$, $\frac{\beta+\mu}{\mu} < 1$ and it is sufficient to show the following:

$$E\left[L(d_1^{**}, d_2^{**}, c, \varepsilon_1, \varepsilon_2)\right] < 0$$

Since the aggregate loss under the throughput guidelines, $E\left[L(0, 0, c, \varepsilon_1, \varepsilon_2)\right]$ is obviously zero ($d_1 = d_2 = 0$ for any cases), now we complete proving that the aggregate loss under the tariff is smaller than the loss under the throughput guidelines.

References

Armantier, O, Arnold, J and McAndrews, J (2008), 'Changes in the timing distribution of Fedwire Funds transfers', *Federal Reserve Bank of New York Economic Policy Review,* Vol. 14, No. 2.

Baglioni, A and Monticini, A (2008), 'The intraday price of money: evidence from the e-MID interbank market', *Journal of Money, Credit, and Banking,* Vol. 40, No. 7, pages 1,533-40.

Bank for International Settlements (1993), 'Payment Systems in the Group of Ten countries', prepared by the Committee on Payment and Settlement Systems of the central banks of the Group of Ten countries, Basle.

Bech, M (2008), 'Intraday liquidity management: a tale of games banks play', *Federal Reserve Bank of New York Economic Policy Review,* Vol. 14, No. 2.

Bech, M and Garratt, R (2003), 'The intraday liquidity management game', *Journal of Economic Theory,* Vol. 109, No. 2, pages 198-219.

Buckle, S and Campbell, E (2003), 'Settlement bank behaviour and throughput rules in an RTGS payment system with collateralised intraday credit', *Bank of England Working Paper No. 209.*

Committee on Payment and Settlement Systems (1997), 'Real-time gross settlement systems', *CPSS Publications No. 22.*

Committee on Payment and Settlement Systems (2005), 'New developments in large-value payment systems', *CPSS Publications No. 67.*

Furfine, C H (2001), 'Banks as monitors of other banks: evidence from the overnight Federal Funds market', *Journal of Business,* Vol. 74, No. 1, pages 33-57.

Holthausen, C and Rochet, J-C (2005), 'Efficient pricing of large value interbank payment systems', *Journal of Money, Credit, and Banking,* Vol. 38, No. 7, pages 1797-818.

Holthausen, C and Rochet, J-C (2006), 'Incorporating a 'public good factor' into the pricing of large-value payment systems', *European Central Bank Working Paper No. 507.*

Kraenzlin, S and Nellen, T (2010), 'Daytime is money', *Swiss National Bank Working Paper No. 2010-06.*

Manning, M, Nier, E and Schanz, J (2009), *The economics of large-value payments and settlement: theory and policy issues for central banks*, Oxford University Press.

Martin, A and McAndrews, J (2008), 'Liquidity-saving mechanisms', *Journal of Monetary Economics,* Vol. 55, No. 3, pages 554-67.

Matsushima, H (2004), 'Repeated games with private monitoring: two players', *Econometrica,* Vol. 72, No. 3, pages 823-52.

McAndrews, J and Rajan, S (2000), 'The timing and funding of Fedwire Funds transfers', *Federal Reserve Bank of New York Economic Policy Review,* Vol. 6, No. 2.

Rochet, J-C and Tirole, J (1996), 'Controlling risk in payment systems', *Journal of Money, Credit, and Banking,* Vol. 28, No. 4, pages 832-62

Walsh, C E (1995), 'Optimal contracts for central bankers', *American Economic Review,* Vol. 85, No. 1, pages 150-67.

Walsh, C E (2002), 'When should central bankers be fired?', *Economics of Governance,* Vol. 3, No. 1, pages 1-21.

Willison, M (2005), 'Real-Time Gross Settlement and hybrid payment systems: a comparison', *Bank of England Working Paper No. 252*.