Settlement bank behaviour and throughput rules in an RTGS payment system with collateralised intraday credit

Simon Buckle[*]

and

Erin Campbell[**]

Working Paper no. 209

[*] Foreign & Commonwealth Office

[**] Bank of England

This paper represents the views and analysis of the authors and should not be thought to represent those of the Bank of England or Monetary Policy Committee members. We would like to thank a number of people at the Bank and elsewhere for their comments and encouragement at various stages of this work, including Prasanna Gai (Australian National University), Chris Mann (Bank of England), Jamie McAndrews (Federal Reserve Bank of New York), Jim Moser (Federal Reserve Bank of Chicago), Jean-Charles Rochet (University of Toulouse), Hyun Song Shin (London School of Economics), John Trundle (Bank of England), Richard Wellington (Bank of England) and Philip Woo (Monetary Authority of Singapore). We would also like to thank an anonymous referee for useful comments. Any errors remaining are, of course, our own.

Copies of working papers may be obtained from Publications Group, Bank of England, Threadneedle Street, London, EC2R 8AH; telephone 020 7601 4030, fax 020 7601 3298, e-mail mapublications@bankofengland.co.uk

Working papers are also available at www.bankofengland.co.uk/wp/index.html

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

ISSN 1368-5562

Contents

Abstract

We present a simple two-period, two-bank model of an RTGS system with collateralised intraday credit. We show that two types of outcome are possible - inefficient or efficient - depending on whether banks care about payments imbalances between them in the first period. If they do, banks delay payments to each other, increasing their aggregate liquidity requirements. We argue that efficiency is not guaranteed even when banks face repeated interaction in a real payment system, largely because of imperfect information and the competitive dynamics of the payment industry. An efficient outcome can be achieved by the imposition of throughput rules on the value of payments banks must make by a certain deadline. These can both reduce aggregate liquidity requirements and increase the contestability of the payments market, encouraging a higher degree of direct access to payment systems. Throughput rules could therefore also have risk-reduction benefits if they help to reduce the level of tiering in the financial system. We show that the detailed characteristics of these rules are important and address a number of design issues such as how frequently requirements should be set, and whether throughput rules should apply on an aggregate or bilateral basis.

JEL classification: E58, G21, G28, L1.

Key words: Payments, settlement, intraday liquidity, competition, regulation.

Summary

High-value payment systems are critical elements of the economy and typically take one of two forms: deferred net settlement (DNS) and real-time gross settlement (RTGS). In a DNS system, banks make payments to each other during a specified period (usually one day) and then settle the net amounts at the end of that period. Until settlement is completed, banks are effectively extending unsecured and possibly unmonitored loans to each other. The amount of credit risk in such systems was one of the main drivers for the introduction of RTGS systems in Europe and elsewhere.

RTGS systems eliminate the counterparty credit risk present in DNS systems by requiring participants to settle payments on a gross basis in real time. But this credit risk reduction comes at the cost of a requirement for potentially expensive intraday liquidity. Central banks have sought to reduce liquidity costs for settlement banks, for example by providing collateralised intraday liquidity and good system design. Even so, intraday liquidity in RTGS systems is not in general free and unlimited.

An important determinant of the liquidity efficiency of an RTGS payment system is the extent to which the system design gives settlement banks an incentive to manage their payments in a socially efficient way. In an RTGS system, one bank's payments are a source of intraday liquidity for the recipient bank, which it may then subsequently use to make its own payments. If banks recycle liquidity sufficiently quickly, the aggregate requirement for intraday liquidity can be significantly reduced.

This paper provides a simple analytical model with which to study RTGS system design. In the context of this two-bank model, we show that banks will delay payments when they care about payment imbalances between them in the first period, leading to an inefficient degree of liquidity recycling. When banks do not care about first-period payments imbalances, there is no unique equilibrium outcome but one possible symmetric outcome is efficient - when each bank posts the same amount of collateral, equal to half of the value of payments each wants to make, and uses all its available liquidity to make payments in the first period. This results in the maximum possible degree of liquidity recycling and the lowest aggregate collateral requirement.

In practice, banks do care about payment imbalances between them during the day because of competitive and/or liquidity risk concerns. While some degree of liquidity recycling is likely to emerge even in these circumstances, in particular due to the repeated nature of the interaction between settlement banks, we argue that full efficiency is not guaranteed, largely because of imperfect information and the competitive dynamics of the payment industry. Using the model, we show how regulation - in this case a throughput rule - can be used to achieve the efficient outcome even in this situation. Throughput rules, which stipulate the proportion of each settlement bank's usual daily payments that must be made by a certain cut-off time, substantially reduce the overall requirement for intraday liquidity in an RTGS system and may also increase the contestability of the payments market, encouraging a higher degree of direct access to payment systems. Consequently, throughput rules could have risk-reduction benefits if they help to reduce the level of tiering in the financial system.

We also address the question of how to design throughput rules in practice. Our model suggests that increasing the number of throughput rules would enhance efficiency indefinitely, although at an ever diminishing rate. It seems likely, however, that there is some upper limit to the efficient number of throughput rules, for at least two reasons. First, real payments have a finite size and are sometimes very large and urgent. If the value of payments required to be made in a given period was less than the size of a large, urgent payment, banks needing to make such payments would be forced to use more liquidity than they receive back from other banks – the original problem that throughput rules were designed to solve. The second reason why the feasible number of throughput rules will be bounded is that real payments between banks are stochastic and, at least in part, unknown at the start of the day. Assuming the throughput rules are based on the average value of payments, as the number of rules increases, eventually a point will be reached where, on a day when the demand for payments is low, one or more of the banks will just not have sufficient customer payments to meet the final throughput requirement.

A further potentially important design issue highlighted in this paper is that aggregate throughput rules may not be adequate in a payment system with more than two settlement banks, since they could not prevent banks from forming cartels to disadvantage other banks or new potential entrants. While we have no evidence of such behaviour within the UK high-value payments system, there may be merit in considering the feasibility of applying throughput rules on a bilateral basis or putting other equivalent incentive mechanisms in place.

1 Introduction

High-value payment systems are critical elements of the economy. Systems such as Target within the European Union, Fedwire in the United States, and CHAPS in the United Kingdom are typically used to make payments equivalent to annual GDP each week. A large fraction of these payments are related to wholesale activity in the major financial markets. Such high-value payment systems are also crucial for the achievement of several central bank objectives. Efficient and robust payment and settlement systems are needed for the implementation of monetary policy through open market operations. The robustness of major payment systems is also critical for financial stability since payment systems link together the most important domestic and international financial intermediaries, thus providing a direct and almost immediate potential contagion mechanism between banks. Finally, the risk and efficiency characteristics of a payment system may have significant implications for social welfare and bank profitability.

High-value payment systems typically come in two main varieties: deferred net settlement (DNS) and real-time gross settlement (RTGS).[1] In a DNS system, banks make payments to each other during a specified period (usually one day) and then settle the net amounts at the end of that period. Until settlement is completed, banks are effectively extending unsecured and possibly unmonitored loans to each other. At any given point during the period, gross exposures may be large relative to the capital of the receiving institution. Although DNS systems can be made more secure by a variety of methods, concerns about the amount of credit risk in such systems were one of the main drivers for the introduction of RTGS systems in Europe and elsewhere. RTGS payments systems eliminate counterparty credit risk for participants by ensuring that settlement is completed in real-time on a gross basis and they can eliminate all credit risk if settlement is effected using central bank liabilities. But this credit risk reduction comes at the cost of a requirement for potentially expensive intraday liquidity.

There are a number of economic questions that any RTGS payment system designer has to address. A key issue is how costly intraday liquidity should be. Theoretical models (see for example Freeman (1996), Green (1999), Zhou (2000) and Kahn and Roberds (2001)) suggest that costly intraday liquidity in payment systems imposes constraints on system participants that may

[1] The hybrid systems developed in some countries (eg CHIPS in the US and RTGS Plus in Germany) are a third type. These systems aim to combine the liquidity saving features of DNS systems with at least some of the risk-reducing features of RTGS systems. See McAndrews and Trundle (2001) for more details.

give rise to distortions in trading and consumption patterns. From a practical point of view, many private sector bankers argue that the opportunity cost of intraday liquidity is becoming more significant with the advent of systems such as CLS, a settlement facility for foreign exchange deals that aims to eliminate Herstatt risk through direct links between national RTGS systems. Settlement banks increasingly face the challenge of real-time liquidity management on a global basis.[2]

Central banks have tried to reduce the cost of intraday liquidity for settlement banks both directly (eg by extending intraday credit to settlement banks) and indirectly through good system design. Even so, intraday liquidity in RTGS systems is not provided without constraint. If it were, counterparty credit risk would not have been eliminated but simply transferred to the central bank. To limit the risks they face, central banks may cap the amount of credit available intraday through price and/or quantity mechanisms (as in the US system, Fedwire)[3] or only provide credit on a collateralised basis, as in the UK system, CHAPS.

An additional design concern is the extent to which settlement banks have an incentive to manage their payments in a socially efficient way. One bank's payments are a source of intraday liquidity for the recipient bank, which it may then subsequently use to make its own payments. If banks recycle liquidity sufficiently quickly, the aggregate requirement for intraday liquidity can be significantly reduced. When intraday liquidity is costly, however, settlement banks may seek to economise on the amount of collateral they provide to the system by increasing the extent to which they depend on incoming payments to provide the necessary intraday liquidity. While such a bank was accumulating sufficient liquidity, it would tend to delay outgoing payments. The effect of such behaviour would be to reduce the overall level of liquidity recycling within the payment system. This problem has been highlighted in a number of papers, notably Angelini (1998) who framed the problem in terms of a trade-off between the costs of payments delay and the (variable) cost of additional intraday liquidity, and more recently by McAndrews and Rajan (2000). If a sufficiently large number of banks behaved like this, the behaviour would become self-defeating. Settlement banks would either have to provide more liquidity than they would have needed if they had behaved co-operatively or the liquidity shortage in the system could become so severe that payments gridlock would ensue.

[2] CLS went live on 9 September 2002.

[3] Fedwire has recently introduced collateralisation under certain circumstances.

Another issue bearing on the cost of intraday liquidity is the nature of the regulatory regime faced by the settlement banks, as noted by Rochet and Tirole (1996). Financial regulations or payment system rules in a number of countries vary between different types of settlement bank, who may then face different opportunity costs for intraday liquidity. When this is the case, we might expect this to be reflected in settlement bank behaviour within the payment system, a possibility noted in McAndrews (1999).[4]

The aim of this paper is to examine the potential for a particular type of rule, a throughput rule, to improve the efficiency of an RTGS payments system with collateralised intraday credit. Throughput rules stipulate the proportion (by value) of each settlement bank's usual daily payments that must be made by a certain cut-off time. (A more formal definition will be given later on in the paper.) Anticipating our findings, such rules can substantially reduce the overall requirement for intraday liquidity in an RTGS system by removing, or at least reducing, externalities and by encouraging efficient liquidity recycling between system participants. A further possible benefit is that sufficiently demanding throughput rules may increase the contestability of the payments market by reducing the amount of liquidity required to participate and thus open access to higher opportunity cost banks. As well as putting downwards pressure on fees charged to customers, this could have risk-reduction benefits by encouraging a higher degree of direct access to payment systems and consequently reducing the level of tiering.[5]

The rest of the paper is organised as follows. In Section 2, we develop the modelling framework - a simple two-period, two-bank model of an RTGS system with collateralised intraday credit. Banks are assumed to be profit maximisers, gaining revenue from making payments for customers but facing three types of costs: an opportunity cost of collateral placed in the payment system intraday; a cost incurred if any payments are cancelled; and a cost proportional to the extent and sign of any imbalances in the value of payments made by the settlement banks in the first period. In Section 3, we use this framework to examine two different possible outcomes, which we label efficient and inefficient. The inefficient Nash equilibrium results when banks care about first-period payment imbalances between them and both banks delay making payments until the second period. In the efficient outcome, banks recycle liquidity between them to the maximum

[4] It seems possible that settlement banks with a relatively low opportunity cost of intraday liquidity might be able to limit the scope for entry or business expansion by other potential payment providers that have a higher opportunity cost of intraday liquidity.

[5] By tiering, we have in mind a situation in which only a few settlement banks are directly connected to the payment system and these banks provide payment services for other banks as well as for their corporate customers.

possible extent by spreading their payments over both periods, thus reducing the aggregate liquidity requirement. This efficient outcome is only possible when banks do not attach a cost to first-period payment imbalances between them, and even then it is not guaranteed. In Section 4, we consider whether there are sufficient pressures within a real payment system to deliver the efficient outcome in the absence of regulation. We conclude that this is not guaranteed. While some degree of co-operation is likely to emerge spontaneously in real payment systems, its extent may be insufficient to achieve the efficient outcome. Section 5 then examines the potential for one type of regulation, a throughput rule, to deliver the efficient outcome even when banks care about payment imbalances between them. Section 6 discusses a number of issues related to the design and implementation of such throughput rules, while Section 7 concludes.

2 A model of an RTGS system with collateralised credit

This section presents a model of an RTGS payment system with collateralised intraday credit provided by the central bank. There are two settlement banks that provide payment services to their customers through an RTGS system run by a central bank. The central bank does not make payments of its own and its role is confined to acting as a settlement agent and providing intraday liquidity against high-quality collateral.

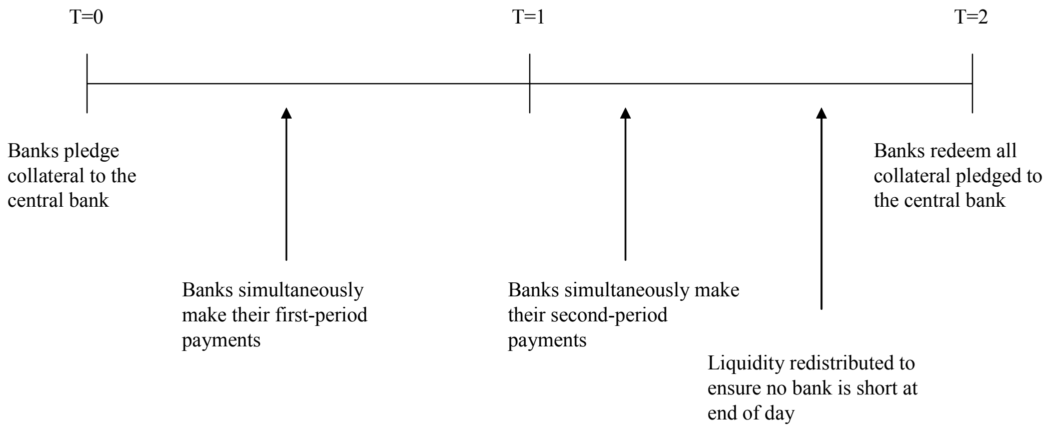
2.1 The timeline

There are three times during the day $(t = 0,\ 1,\ 2)$. There is no wholesale market in intraday liquidity between the settlement banks. In other words, each bank must post its own collateral to generate the required intraday liquidity.[6] At $t = 0$, the start of day, both banks post collateral with the central bank and obtain intraday liquidity of $\mathcal{L}_i^{t=0} = c_i\ for\ banks\ i = 1, 2$ in their accounts at the central bank.[7] We assume that banks cannot increase or decrease collateral during the day. Banks simultaneously make payments to each other twice during the day, once in the first period and once in the second. At the end of the day banks have to redeem all the collateral they have pledged to the central bank. A timeline is presented in Chart 1.

[6] Allowing trading between low and high opportunity cost banks might be more efficient in principle, but if banks are concerned about their market share of the payments business, low opportunity cost banks are unlikely to offer high cost banks a significant discount and thereby threaten their own profitability.

[7] The amount of available liquidity will in general be less than the market value of the assets, reflecting any haircuts that the central bank has applied.

Chart 1: Timeline for the RTGS day


T=0
T=1
T=2
Banks pledge collateral to the central bank
Banks simultaneously make their first-period payments
Banks simultaneously make their second-period payments
Liquidity redistributed to ensure no bank is short at end of day
Banks redeem all collateral pledged to the central bank

2.2 The nature and timing of payments

We assume that each bank's normal planned daily payments to the other are identical and non-stochastic, equal in value to μ. Individual payments are assumed to be sufficiently small that we can ignore any problems associated with the need to deal with payments that are large relative to the size of collateral posted.[8]

Banks can make payments twice during the day - once in the first period and once in the second. Payments are made simultaneously by each bank in each period. The value of payments made in each period is constrained by the amount of intraday liquidity each bank has on its central bank settlement account at the start of that period - the RTGS constraint. If, in the second period, a bank does not have sufficient liquidity to make all the payments it desires, then some of its payments will be cancelled.

In the first period, settlement bank 1 makes payments to bank 2 equal to a fraction α_1 of its available intraday liquidity, c_1:

$$Period\ 1\ payments\ by\ bank\ 1 = \alpha_1 c_1$$

Simultaneously, bank 2 makes payments of $\alpha_2 c_2$. Both the initial level of intraday liquidity, $\mathcal{L}_i^{t=0} = c_i$, and the fraction paid in the first-period, α_i, are choice variables for the banks. After making its first-period payments and receiving those made by bank 2, bank 1's intraday liquidity

[8] This assumption simplifies the analysis considerably but may not hold in practice. We leave analysis of this problem for future research.

is given by:

$$\mathcal{L}_1^{t=1} = (1 - \alpha_1)c_1 + \alpha_2 c_2$$

This then constrains the amount of payments that it can make in the second period:

$$Period\ 2\ payments\ by\ bank\ 1 = Min\left[\mu - \alpha_1 c_1, (1 - \alpha_1)c_1 + \alpha_2 c_2\right]$$

At the end of the second period, bank 1's liquidity, $\mathcal{L}_1^{t=2}$, is made up of any liquidity remaining after it has made all its feasible second-period payments plus the payments it has received from the other bank in the second period:

$$\mathcal{L}_1^{t=2} = Max\left[c_1 + \alpha_2 c_2 - \mu, 0\right] + Min\left[\mu - \alpha_2 c_2, (1 - \alpha_2)c_2 + \alpha_1 c_1\right]$$

The level of cancelled payments depends on not only the collateral provided by bank 1 but also the level of first-period payments made by bank 2:

$$Level\ of\ cancelled\ payments\ (bank\ 1) = Max\left[\left(\mu - c_1 - \alpha_2 c_2\right), 0\right]$$

2.3 The end-of-day liquidity requirement

At the end of the day, $t = 2$, each settlement bank must have sufficient liquidity on its account at the central bank to redeem its collateral. This ensures the separation of intraday and overnight credit and allows monetary policy to be set independently of the liquidity requirements of the payment system. Since the model payment system is closed, there must be sufficient aggregate liquidity in the payment system at all times to allow for redemption. But the liquidity may in principle be unevenly distributed among the banks - at the end of day, one may be short of liquidity and the other long, depending on the nature of their payment flows. In a real payment system, any such imbalances would normally be dealt with through short-term lending in the interbank market.

The extent to which liquidity needs to be reallocated at the end of the day is determined by the requirement that banks must redeem their collateral. In our model, bank 1's end-of-day liquidity requirement (LR_1) is measured by the difference between the collateral placed with the central bank at the start of the day, c_1, and $\mathcal{L}_1^{t=2}$, the amount of intraday liquidity remaining to it after all feasible payments have been made:

$$LR_1 = c_1 - \mathcal{L}_1^{t=2}$$

We assume that a sustainable payment system is one in which banks do not run down their balance sheets to make payments. This implies that the expected liquidity requirement for each bank must be zero, ie $E\left[LR_i\right] = 0$. With non-stochastic payments, the end-of-day liquidity position of each bank must therefore be just sufficient to redeem the collateral it pledged at $t = 0$ without the need for any liquidity reallocation. This simplification does not limit the force of our observations on RTGS system design.

2.4 Behavioural assumptions

In order to determine how settlement banks choose the level of collateral they place with the central bank and the amount of payments they will make in the first period, we need to add some behavioural assumptions. Settlement banks are assumed to be profit maximisers. They gain revenue from making payments for customers but must offset this against the opportunity cost of collateral placed with the central bank. Additionally, settlement banks suffer a financial penalty for cancelled payments. This may be thought of as either an immediate compensation payment to a customer or as a reputational cost leading to loss of future business. Including the cost of cancelling payments makes our set-up similar to that of Kobayakawa (1997). An important difference between his work and ours is that we assume that the costs of cancellation fall only on the sending bank and not on both settlement banks symmetrically.[9]

To these relatively familiar costs, our model adds the possibility that banks incur a further cost proportional to the extent of the payments imbalance between them at the end of the first period, ie $(\alpha_1 c_1 - \alpha_2 c_2)$. The motivation for such a term is that some settlement banks in the UK RTGS system impose bilateral ceilings on the extent to which they will make payments to another settlement bank in the absence of adequate offsetting receipts. In other words, at least some settlement banks appear to care about their position as liquidity providers or receivers during the day.

[9] If one bank cancels a payment, the receiving bank experiences a loss of liquidity associated with the cancelled payment. Kobayakawa argues that both banks inconvenience their customers and thus both banks' reputations suffer equally. We assume the receiving bank's reputation will only suffer to the extent it cancels its own payments because it receives less liquidity. As we shall see this leads to different behavioural predictions. In Kobayakawa's model, the Nash equilibrium is for banks to pay early to avoid the cost of cancelled payments.

2.5 *Payments revenue*

Payments made by settlement banks are assumed to be made on behalf of their customers.[10] We assume that settlement banks gain revenue by charging a fee of F_i per unit value of the payment.[11] This fee may differ between banks. A settlement bank's total revenue from making payments will then be given by the unit fee charged multiplied by the value of payments that it successfully makes. The value of payments made will be the minimum of the desired level of payments, μ, and the sum of its own collateral and its first-period receipts from the other bank. Bank 1's revenue will therefore be given by:

$$TR_1 = F_1 Min\left[\mu, c_1 + \alpha_2 c_2\right] \leq \mu F_1$$

2.6 *The costs of making payments*

Offsetting this revenue are the costs incurred in making payments: the opportunity cost of collateral, the cost of cancelling any payments and the costs of liquidity imbalances between banks in period one. For convenience, we also assume there is no transaction fee charged by the system operator.

The opportunity cost of intraday credit arises because a settlement bank in an RTGS system with collateralised intraday liquidity has to hold the high-quality assets on its balance sheet to gain access to liquidity but might, for profit maximisation reasons, prefer to hold higher yielding assets. Even if this were not the case, there may be costs stemming from having to administer the assets or from having to forgo any alternative profitable use of those assets during the day (eg securities lending). We assume that these opportunity costs are linear and are represented by a bank-specific parameter, γ_i, multiplied by the level of collateral placed with the central bank.[12]

[10] For simplicity, we ignore proprietary payments to other settlement banks.

[11] Since payments are identical in this model, this is equivalent to charging a fixed fee per payment, which seems to be banks' usual practice.

[12] In the UK, the parameter may vary between settlement banks because of differences in regulatory treatment. At present, the UK regulator (the FSA) requires that large UK retail banks (the majority of participants) maintain a minimum level of high-quality, liquid assets against the occurrence of liquidity shocks. This stock of high-quality assets is also eligible for intraday repos with the Bank of England to facilitate payments in CHAPS Sterling. The opportunity cost of using such collateral to facilitate payments is therefore small and perhaps even zero for such banks. Three foreign-owned banks in CHAPS Sterling are regulated in a different way, based on the maturity of their assets and liabilities. They are not explicitly required to hold a stock of specific high-quality liquid assets, although they are advised that this might be prudent. For this group of banks the marginal opportunity cost of collateral is likely to be higher than for the retail banks and might be thought of as the difference between the yield on such safe, liquid assets and some other more risky and less liquid asset that the bank might prefer to hold in its portfolio.

It is worth underlining the fact that we are assuming that the opportunity cost is a fixed cost in this type of RTGS system. Much of the previous related literature dealt with systems such as Fedwire, which - before the introduction of collateral - operated on the basis of a charge for the amount of credit in use at the end of fixed periods during the day. In such systems, the opportunity cost of intraday credit is a variable cost. It might be argued that since collateral can be added or removed from the system during the day, the opportunity cost is also a variable cost in a collateralised RTGS system.[13] But the difficulties and delays in transferring title to securities would seem to preclude them being used in the payment system for part of the day and then being used for other purposes for the remainder of the day. Indeed, the fixed-cost assumption seems to accord with settlement bank practice in the United Kingdom's Sterling RTGS system, where the amount of collateral posted remains largely unchanged through the day.

We model the cost of cancelling payments as a non-linear and increasing function of the level of cancelled payments. If only a small fraction of customer payments are not made in a given day, there may be little or no direct cost or damage to the reputation of the bank. For example, it might be possible to resubmit the payments the next day with little or no inconvenience. If however customers faced a significant probability of their payments being cancelled, then they would be likely to find other ways of making payments, for example by moving their account to another bank, with a consequent loss of revenue to the first bank. So large cancellations seem likely to have a disproportionate impact on a bank's revenue from payments business. For simplicity, we choose to capture this feature by using a quadratic cost term multiplied by a (non-zero) bank-specific parameter, β_i.The larger the value of β_i, the less payments a bank will be willing to cancel.

The third cost in the model reflects settlement banks' aversion to making more outward payments than they receive from other banks in a given period. There are at least two plausible motivations for this observed behaviour and each may hold in part. The first is that if one bank were systematically providing another with liquidity by making more payments than it received for prolonged periods of the day, it would in effect be subsidising the operations of the other bank. This fortunate bank could then post a lower amount of costly collateral and still make its required payments, thus increasing its profitability at the expense of the first bank. By limiting its net provision of liquidity to other banks, a settlement bank thereby constrains the extent to which other banks are able to free-ride on its liquidity.

[13] For example, see the paper on strategic behaviour by banks within RTGS systems by Bech and Garratt (2001).

A second possible reason is that settlement banks may be concerned about the impact any disruption of the payment system could have on their own liquidity position. The greater the imbalance of payments between banks, the more vulnerable the net payer is to operational or other risks that could disrupt a counterparty's ability to make offsetting payments during the remainder of the day. To the extent that liquidity risk is a concern, banks may want to limit the scale of bilateral liquidity imbalances with other settlement banks.

We can capture this aspect of bank behaviour by adding a further cost term proportional to the extent of the payments imbalance between settlement banks at the end of the first period to the model set out above. The total costs for settlement bank 1 then become:

$$TC_1 = \gamma_1 c_1 + \beta_1 \left(Max\left[\mu - c_1 - \alpha_2 c_2, 0\right]\right)^2 + \xi_1(\alpha_1 c_1 - \alpha_2 c_2)$$

3 Possible outcomes

We now examine the possible outcomes of this model under two different assumptions for the parameters ξ_i. We first study the case when $\xi_i > 0$, which leads to what we term the inefficient outcome. We then focus on the possible outcomes when $\xi_i = 0$ and banks do not care about liquidity imbalances between them at the end of period one.

3.1 The inefficient outcome

Each settlement will seek to maximise its profits,

$$\Pi_i = TR_i - TC_i$$

subject to the constraint that $0 \leq \alpha_i \leq 1$. When $\xi_i > 0$, we note that Π_i decreases with α_i. This makes it optimal for bank 1 to set $\alpha_1 = 0$. The same argument applies to bank 2. Any non-zero, positive value of ξ therefore causes each bank to delay all of its payments to the second period, thus raising the aggregate collateral requirement for each bank.

Using this result, we can write the profit function as:

$$\Pi_i = F_i Min\left[\mu, c_i\right] - \gamma_i c_i - \beta_i \left(Max\left[\mu - c_i, 0\right]\right)^2$$

The solution depends on the relative size of the fee, F_i, and the opportunity cost of collateral, γ_i. When the fee charged for making payments is greater than the opportunity cost for both banks, neither will wish to cancel payments and banks will post just sufficient collateral to ensure that this can be achieved. When the opportunity cost of collateral is greater than the fee for both settlement banks, they will always cancel some payments.

Proposition 1 When $F_i > \gamma_i$, bank i will not wish to cancel any payments and the level of collateral it chooses to place in the RTGS system is $c_i^* = \mu$. The payment system is sustainable when both banks choose this level of collateral.

Proof. When $c_1^* < \mu$, bank 1's profits can always be increased by raising the level of collateral posted. But when $c_1^* \geq \mu$, additional collateral does not generate any more revenue but only incurs additional opportunity costs. Bank 1 will therefore post the minimum level of collateral required to ensure it makes all its desired payments, ie:

$$c_1^* = \mu$$

So long as both banks choose this same level of collateral, the expected end-of-day liquidity requirement is zero and the payment system is sustainable.

Proposition 2 When $F_i < \gamma_i$, bank i will always cancel some payments and will optimally choose to post collateral of:

$$c_i^* = \mu - \frac{(\gamma_i - F_i)}{2\beta_i}$$

For a sustainable payment system, the other bank must cancel an equal amount of payments.

Proof. When $F_1 < \gamma_1$, the first-order condition for a maximum is:

$$F_1 - \gamma_1 + 2\beta_1(\mu - c_1) = 0$$

Rearranging yields the desired expression. Since $F_1 < \gamma_1$, the level of collateral must be lower than the level of desired payments, μ, resulting in cancelled payments of $\frac{(\gamma_1 - F_1)}{2\beta_1}$. For a sustainable payment system, we require that $E[LR_i] = 0$. Both banks must therefore cancel the same amount of payments.(14)

This inefficient outcome is the result of potential undercutting. Suppose each bank were to place collateral with the central bank at a lower level than c^*, expecting to make and receive some

(14) This does not mean that banks will set out to cancel a significant amount of payments each day. The amount of cancelled payments tends to zero in the limit as the cost of cancelling payments, β, tends to infinity. Alternatively, with a finite β, the level of cancelled payments will become arbitrarily small as the fee charged approaches the opportunity cost of collateral.

payments in the first period. But any bank making payments in the first period would have no guarantee that the other bank would make offsetting payments in return, even though by doing so both banks could make all their scheduled payments with a lower level of collateral. Indeed, the other bank might see a competitive advantage in not making a reciprocal payment. It could still make all of its payments in the second period while forcing the first bank to cancel some of its payments. The Nash equilibrium in such a situation is for both banks to plan on the basis that neither bank will make a payment in the first period.

3.2 The efficient outcome

The inefficient outcome above is obviously not the best that could be achieved. For example, imagine the case where neither bank desires to cancel payments, ie $F_i > \gamma_i$. If both banks could credibly commit to placing $\frac{\mu}{2}$ of collateral in the system and to using all of this available liquidity to make payments to each other in the first period (ie $\alpha_i = 1$), the aggregate liquidity requirement could be reduced by a factor of two relative to the level of collateral required in the case when $\xi_i > 0$. This efficient outcome is in fact one possible solution of the model when the two banks do not care about payments imbalances between them during the first period. Similar collateral reductions are also possible when $F_i < \gamma_i$.

Proposition 3 When $\xi_i = 0$ and $F_i > 0.5\gamma_i$, there exists a Pareto dominating Nash equilibrium where each bank posts the same amount of collateral, $c^* = \frac{\mu}{2}$, and makes payments of $\frac{\mu}{2}$ in each period.

Proof. Suppose each bank chooses the same level of collateral, c^*, and can credibly commit to making payments equal to c^* in each period. The profit function for bank 1 is then:

$$\Pi_1 = F_1 Min\left[\mu, 2c^*\right] - \gamma_1 c^* - \beta_1 \left(Max\left[\mu - 2c^*, 0\right]\right)^2$$

When $c^* < 0.5\mu$, since $F_1 > 0.5\gamma_1$, bank 1's profits can always be increased by raising the level of collateral posted and reducing the amount of payments cancelled. But when $c^* \geq 0.5\mu$, additional collateral does not generate any more revenue but only incurs additional opportunity costs. Bank 1 will therefore want to post the minimum level of collateral required to ensure it makes all its desired payments, ie:

$$c^* = \frac{\mu}{2}$$

Bank 2 will also want to post the same level of collateral, since $F_2 > 0.5\gamma_2$. No payments are cancelled and therefore the system is sustainable.

Proposition 4 When $\xi_i = 0$ and $F_i < 0.5\gamma_i$, there exists a Pareto dominating Nash equilibrium where each bank posts the same optimal level of collateral given by:

$$c^* = c_i^* = \frac{1}{2}\left(\mu - \frac{(\gamma_i - 2F_i)}{4\beta_i}\right)$$

and makes payments equal to c^* in each period. A necessary condition for the existence of this equilibrium is that each bank cancels the same amount of payments:

$$\frac{(\gamma_1 - 2F_1)}{4\beta_1} = \frac{(\gamma_2 - 2F_2)}{4\beta_2}$$

Proof. Again, suppose each bank chooses the same level of collateral, c^*, and can credibly commit to making payments equal to c^* in the first period. The profit function for bank 1 is then:

$$\Pi_1 = F_1 Min\left[\mu, 2c^*\right] - \gamma_1 c^* - \beta_1 \left(Max\left[\mu - 2c^*, 0\right]\right)^2$$

Differentiating with respect to c^*, the first-order condition is:

$$2F_1 - \gamma_1 + 4\beta_1 (\mu - 2c^*) = 0$$

An optimal solution exists for the parameter range $F_1 < 0.5\gamma_1$. Rearranging yields the desired result. The optimal collateral for bank 2 is derived in exactly the same way and the optimal collateral for each bank must be equal for a solution to exist and also for the payment system to be sustainable.

These symmetric, Pareto dominating Nash equilibria are not, however, guaranteed to occur. Since each bank's level of profit is independent of the level of its first-period payments, there is in fact an infinity of potential solutions, both symmetric and asymmetric, one for each couple (α_1, α_2).[15]

3.3 Comparison of the inefficient and efficient outcomes

The Pareto dominating outcome reduces the collateral required and increases bank profitability relative to the inefficient Nash equilibrium, whether or not the settlement banks cancel payments. This happens because banks now generate the same or more revenue from a much smaller amount

[15] We are grateful to Professor Rochet for drawing this to our attention. Asymmetric outcomes imply that one bank is subsidising the other.

of collateral - the marginal cost of collateral remains fixed but the marginal revenue doubles since both banks make use of each other's collateral in equal measure. Additionally, if banks optimally choose to cancel some payments in the inefficient equilibrium, they either choose not to cancel payments or to cancel fewer payments in the efficient case, depending on the relative size of their opportunity cost of collateral and the customer fee. Whenever banks cancel payments, they make negative profits. We prove these propositions below.

Proposition 5 For a payment system in which both banks satisfy $F_i > \gamma_i$, neither bank will cancel payments in either the inefficient or the efficient outcome. Bank profitability will be greater in the efficient case.

Proof. We have already proved that banks do not cancel payments when $F_i > \gamma_i$. Bank profits in the inefficient case are:

$$\Pi_i^{Ineff} = (F_i - \gamma_i)\mu > 0$$

Profits in the efficient case are:

$$\Pi_i^{Eff} = (F_i - \frac{\gamma_i}{2})\mu > 0$$

which means that

$$\Pi_i^{Eff} - \Pi_i^{Ineff} = \frac{\gamma}{2}\mu_i > 0$$

Proposition 6 For a payment system in which both banks satisfy $0.5\gamma_i < F_i < \gamma_i$, banks will choose not to cancel payments in the efficient case, even though they will choose to cancel payments in the inefficient case. Profitability is greater in the efficient case.

Proof. When $F_i < \gamma_i$, we have already proved that banks will cancel payments in the inefficient case. We have also shown that when $F_i > 0.5\gamma_i$, banks will choose not to cancel payments in the efficient case. It remains to compare profitability in the two cases. Profits in the inefficient equilibrium can be written as:

$$\Pi_i^{Ineff} = \left(\gamma_i - F_i\right)\left\{\frac{\left(\gamma_i - F_i\right)}{4\beta_i} - \mu\right\} < 0$$

which is negative since $\gamma_i > F_i$ and one half of the level of cancelled payments must be smaller than the total level of payments, μ. In the efficient case, profitability is:

$$\Pi_i^{Eff} = (F_i - \frac{\gamma_i}{2})\mu > 0$$

Since by assumption, $F_i > \frac{\gamma_i}{2}$, this must be positive. Hence, the required result:

$$\Pi_i^{Eff} - \Pi_i^{Ineff} = \frac{\gamma_i \mu}{2} - \frac{(\gamma_i - F_i)^2}{4\beta_i} > 0$$

Proposition 7 For a payment system in which $F_i < 0.5\gamma_i$, banks cancel payments in both the inefficient and the efficient cases. However, they will cancel fewer payments in the efficient case and their losses are lower.

Proof. Profitability in the inefficient case is:

$$\Pi_i^{Ineff} = (\gamma_i - F_i)\left\{\frac{(\gamma_i - F_i)}{4\beta_i} - \mu\right\} < 0$$

and in the efficient case:

$$\Pi_i^{Eff} = \frac{(\gamma_i - 2F_i)}{2}\left\{\frac{(\gamma_i - 2F_i)}{8\beta_i} - \mu\right\} < 0$$

Since each bank cancels $\frac{\gamma_i}{4\beta_i}$ fewer payments in the efficient case and less collateral is required, profits increase by:

$$\Pi_i^{Eff} - \Pi_i^{Ineff} = \frac{\gamma_i \mu}{2} - \frac{\gamma_i\left(\frac{3\gamma_i}{4} - F_i\right)}{4\beta_i}$$

So not only is the efficiency of the payment system improved by more rapid recycling of liquidity between the settlement banks, it also makes it possible for a relatively high opportunity cost bank to enter the payments business and compete with a low opportunity cost firm profitably. As well as potentially putting downwards pressure on fees charged to customers, encouraging direct access to the payment system could reduce the degree of tiering and, consequently, reduce the systemic importance of individual players.

4 Can banks achieve efficiency without regulation?

Is there any reason to suppose that settlement banks would themselves eventually arrive at the Pareto dominant co-operative outcome in the absence of regulation? We begin by considering factors that might tend to encourage some degree of liquidity recycling even when banks do care about payments imbalances between them. We then consider whether these factors are sufficient to ensure that real payment systems achieve an adequate level of efficiency. We identify a range of countervailing forces that suggest that this type of RTGS payment systems may not necessarily be able to achieve an adequate level of efficiency in the absence of regulation.

4.1 Factors encouraging greater liquidity recycling

There are at least four factors that suggest that the inefficient outcome of delaying all payments to the second period may be too extreme a prediction for real payment systems. These are: (i) the need for banks to make time urgent payments; (ii) operational risk; (iii) constraints on the levels of collateral even the biggest settlement banks may wish to hold for payments purposes; and (iv) the indefinitely repeated game played by settlement banks in any real payment system. We will now briefly discuss each of these in turn.

Time urgent payments are payments that need to be made by or before a certain time, for example by $t = 1$ in our model. They might include payments made by banks related to their foreign exchange and money market transactions, as well as payments made by firms completing large transactions such as mergers or acquisitions. The cost of not making such high-value payments might outweigh the cost incurred by any subsequent liquidity imbalance and the possible need for additional collateral if the receiving bank did not make offsetting first-period payments in return. If so, banks would want to make these payments in the first period, even if this provided essentially free liquidity to the other settlement bank. Of course, if the amount of time urgent payments that each bank wanted to make was roughly the same and stable over time, this might encourage movement towards, and help sustain, a relatively more efficient outcome.

There is always a risk that some component of the payment system - either in the common infrastructure or within an individual settlement bank - may suffer an operational failure for part or (less likely) all of the payments day. Payments in many RTGS systems typically comprise a small number of really high value payments and a larger number of smaller payments. Banks rely heavily on their sophisticated IT systems to deal with the large volumes of such payments. An operational failure after a bank has sent the bulk of its small-value payments may not be too serious since the remaining high-value payments may be sufficiently small in number that they can be dealt with manually. In these circumstances, banks may wish to make as many as possible of the small-value payments early in the day, even if this means they are providing some free liquidity to others.

The third factor that may encourage some degree of liquidity recycling is simply that the inefficient equilibrium demands a lot of liquidity. In the extreme case described in the model, banks must have one unit of collateral for each unit payment they wish to make. If banks are unable to charge fees greater than the opportunity cost of collateral, they will cancel payments and lose money. Even if the banks could recoup their own opportunity costs through the fees they charged their customers, they could still increase their own profits by co-operating with the other settlement banks. Whatever the relative size of F and γ, banks would therefore seem to have an incentive to co-operate, if only they could trust each other not to free-ride.

This is of course where the repeated nature of the interaction between settlement banks becomes particularly important. A real payment system is more like an indefinitely repeated game than the one-shot game developed above. So provided that free-riders can be punished by the other banks, we might expect some degree of sustainable co-operation to emerge.

4.2 Factors limiting payment system efficiency

There are, however, reasons for believing that some degree of inefficiency might persist, even with repeated interaction between settlement banks. Time urgent payments typically comprise only a small, albeit possibly growing, percentage of total payments. And the smaller-value payments that banks might like to get out of the way early in the day for operational reasons may not amount to more than a small fraction of the total value of payments for a given bank either. In other words, time urgent payments and operational risk together may not ensure that banks recycle liquidity frequently enough during the day to achieve the Pareto dominating outcome.

Competition issues may also be a constraining factor. While liquidity recycling improves the incumbent banks' profitability relative to the inefficient case, it may also facilitate market entry by relatively high cost payment providers. Such actual or threatened entry would tend to drive down the level of fees charged by the incumbents. This is of course good for economic efficiency but bad for the incumbents' profits and there is a risk that they may therefore limit the degree of liquidity recycling in the RTGS payment system. The extent to which this is a concern is likely to depend on the extent of any differences in the opportunity cost facing different potential settlement banks and the degree to which their customers are willing and able to move their payments business to other banks.

A further possible limitation on the degree of co-operation in a real payment system is the fact that it may be very difficult to identify, and therefore punish, free-riders. A real payment system has more (perhaps many more) than two banks and payments between them are stochastic. This means that the signals required to trigger a punishment strategy may be imperfect. Even if free-riders could be accurately identified, differences between banks may subsequently make it difficult to enforce effective punishment strategies, for example a bank facing a low opportunity cost of collateral may be relatively immune from punishment by banks with a relatively high cost of collateral.

These are good reasons for thinking that settlement banks may not necessarily be able, or even want, to achieve the Pareto dominating efficient outcome in practice. Given the scale of the potential welfare loss from inefficient liquidity recycling, the question arises whether there is a regulation that could ensure that settlement banks behaved sufficiently co-operatively. In the next section, we focus on the potential for the use of throughput rules to remove or at least limit the extent of any inefficiencies arising from sub-optimal liquidity recycling between the settlement banks.

5 Throughput rules

A throughput rule is a binding requirement on settlement banks to make a proportion of their normal value of payments by a certain cut-off time during the payments day. The rule may be defined on the basis of the aggregate payments made by each individual bank to all other banks in the system, or alternatively on the basis of each bank’s payments to every other settlement bank on a bilateral basis. A bank’s normal payments can also be defined in a number of ways. For example, as the value of total payments on that day or as the average value over some longer period, which could be retrospective, prospective or overlapping. Rules may also differ in how they are applied; banks’ payments could be assessed against the throughput rule on a daily basis or over a longer period of time.

It might help to give some specific examples of throughput rules. We will confine ourselves here to aggregate throughput rules since in a system with two banks, there is no difference between these and bilateral rules. We discuss some important differences between these types of rules in the next section.

We denote the total value of payments made on the n^{th} day by TP_n and the value of payments made to all other settlement banks by the specified cut-off time on the same day by P_n. Then the following four throughput rules would all require the same proportion, τ of the settlement banks' total daily payments (suitably measured) to be made by the cut-off time:

Definition 1 *The throughput is calculated on a daily basis. Compliance is determined by the size of the ratio of the value of the payments made in period 1 on day n to the average total value of the daily payments over N days, ie:*

$$\frac{P_n}{\frac{1}{N}\sum_{n=1}^{N} TP_n} \geq \tau$$

Definition 2 *The throughput is calculated on a daily basis. Compliance is determined by the size of the ratio of the value of the payments made in period 1 on day n to the total value of the payments on the same day ie:*

$$\frac{P_n}{TP_n} \geq \tau$$

Definition 3 *The throughput rule is calculated over N days. Compliance is determined by the size of the ratio of the average value of the payments made in period 1, measured over N days, to the average value of the total payments measured over the same period, ie:*

$$\frac{\sum_{n=1}^{N} P_n}{\sum_{n=1}^{N} TP_n} \geq \tau$$

Definition 4 *The throughput is calculated over N days. Compliance is determined by the average over a period of N days of the daily ratio of the value of the payments made in period 1 to the total value of the payments made on the same day, ie:*

$$\frac{1}{N}\sum_{n=1}^{N} \frac{P_n}{TP_n} \geq \tau$$

A throughput rule such as Definition 3 or 4 that is calculated over a period can be met in many different ways. For example, a bank could satisfy the rule by behaving according to either Definition 1 or 2 on a daily basis or even by underperforming for some sub-period and then overperforming against the requirement for the remainder of the period to compensate. Daily throughput rules like Definitions 1 and 2 are therefore potentially more demanding than average requirements - a point to bear in mind when considering the implications of our results for a real RTGS system.

In the CHAPS Sterling RTGS system in the United Kingdom, there are two throughput rules that stipulate that banks should make 50% of their payments by value by 12.00 and 75% of their payments by 14.30 on average each month, measured retrospectively. In other words, the CHAPS throughput requirements are of the form set out in Definition 3, with N set to the number of payment days in one calendar month and $\tau_{P1} = 0.5$ for the noon cut-off and $\tau_{P2} = 0.75$ for the 14.30 cut-off. In CHAPS Sterling, there is a reputational cost to not complying with the throughput rules, which increases for repeat offenders.(16)

5.1 *The payments model with throughput rules*

We now wish to incorporate a throughput rule into the RTGS model. We continue to assume that the payments process is non-stochastic and stationary over time. Under these circumstances, all four definitions of a throughput rule given above are equivalent. We assume that the rule stipulates that each bank's first-period payments must be at least equal to $\tau\mu$ or else they incur some penalty cost, proportional to a parameter $\epsilon > 0$. Both the throughput rule and the penalty charge are common between banks. Since the number of payments is bounded, τ satisfies the inequality, $0 \leq \tau \leq 1$.

With these assumptions, the profit function for bank 1 then becomes:

$$\begin{aligned}\Pi_1 &= F_1 Min\,[\mu, c_1 + \alpha_2 c_2] - \gamma_1 c_1 - \beta_1 \left(Max\,[\mu - c_1 - \alpha_2 c_2, 0]\right)^2 \\ &\quad -\xi_1(\alpha_1 c_1 - \alpha_2 c_2) - \epsilon Max[(\tau\mu - \alpha_1 c_1)\,, 0]\end{aligned}$$

The final term reflects the penalty for failing to meet the period 1 throughput rule.

Proposition 8 Provided $\epsilon > \xi_i$, each bank will make first-period payments of exactly $\tau\mu$.

Proof. Provided the penalty, ϵ, exceeds the costs to either bank of withholding payments in the first period (ie $\epsilon > \xi_i$), the profit function is increasing in $\alpha_i c_i$ until first-period payments reach

(16) The loss of reputation is apparent to other members of the payment system, but not the customer requesting the payment.

the level of $\tau\mu$. After this point, ie when $\alpha_i c_i \geq \tau\mu$, the profit function becomes strictly decreasing in α_i. Both banks will therefore make first-period payments exactly equal to $\tau\mu$.

Using this result, the profit function can then be written:

$$\Pi_i = F_i Min\left[\mu, c_i + \tau\mu\right] - \gamma_i c_i - \beta_i\left(Max\left[\mu - c_i - \tau\mu, 0\right]\right)^2$$

Since collateral is a fixed cost in this model, symmetry demands that the maximum profit for settlement banks will involve spreading the payments they want to make evenly over the two periods - as in the Pareto dominating efficient outcome studied in Section 3. In other words, settlement bank profits will be maximised when the throughput requirement, $\tau\mu$, is equal to the collateral they wish to place in the system and when the total payments they desire to make is equal to $2\tau\mu$. If settlement banks placed more collateral than this in the system, that would mean either that the payments flows were unbalanced between periods or that there was unused collateral in the system. In either case, costs could be reduced further.

Of course, we have not specified how many payments banks will want to make. Again there are two cases, depending on the relative sizes of F_i and γ_i.

Proposition 9 When $F_i > 0.5\gamma_i$ and the throughput parameter τ is set at 0.5, each bank will choose to hold collateral at the same Pareto dominating level of $c^* = 0.5\mu$ and no payments will be cancelled.

Proof. We need to express each bank's profits in terms of the throughput requirement and then maximise with respect to τ, ensuring that the solution for each bank is compatible with that for the other. When each bank posts collateral equal to $\tau\mu$, bank 1's profits are:

$$\Pi_1 = F_1 Min\left[\mu, 2\tau\mu\right] - \gamma_1\tau\mu - \beta_1\left(Max\left[\mu - 2\tau\mu, 0\right]\right)^2$$

Differentiating with respect to τ, we see that when $2F_1 > \gamma_1$ bank 1 will not want to cancel payments, since revenue can always be increased by posting more collateral up to the point at which there are no more payments to be made. But it will not increase collateral beyond this point since it then just incurs additional costs but no additional revenues. So, for bank 1, the desired collateral is $c_1^* = \tau\mu = 0.5\mu$. Provided $2F_2 > \gamma_2$, bank 2 will choose the same level of collateral. Since the two banks desire to hold the same level of collateral, this outcome can be realised by the system operator setting $\tau = 0.5$. The end-of-day liquidity requirement will then be zero.

Proposition 10 When $F_i < 0.5\gamma_i$ and the throughput parameter τ is set at 0.5, both banks will desire to hold collateral of:

$$c_i^* = \frac{1}{2}\left(\mu - \frac{(\gamma_i - 2F_i)}{4\beta_i}\right)$$

The Pareto dominating equilibrium can be achieved providing that $c_1^* = c_2^*$. A necessary condition for this equilibrium is that both banks desire to cancel the same amount of payments.

Proof. When $F_1 < 0.5\gamma_1$, there is an optimal level of collateral since bank 1 will wish to cancel some payments. Differentiating the level of profit with respect to τ, the first-order condition is:

$$2F_1 - \gamma_1 + 4\beta_1 (\mu - 2\tau\mu) = 0$$

Rearranging yields the optimal throughput from bank 1's perspective:

$$\tau_1\mu = \frac{1}{2}\left(\mu - \frac{(\gamma_1 - 2F_1)}{4\beta_1}\right)$$

Since $F_2 < 0.5\gamma_2$, the optimal throughput from bank 2's perspective will be:

$$\tau_2\mu = \frac{1}{2}\left(\mu - \frac{(\gamma_2 - 2F_2)}{4\beta_2}\right)$$

A necessary requirement for achieving the Pareto dominating outcome in this case is that these two measures are the same, ie $\tau_1 = \tau_2$. This requires that each bank wants to cancel the same amount of payments:[17]

$$\frac{(\gamma_1 - 2F_1)}{4\beta_1} = \frac{(\gamma_2 - 2F_2)}{4\beta_2}$$

6 Throughput rule design

In this penultimate section, we address two key issues. First, how many throughput rules should be applied within each day? Second, is the aggregate throughput requirement used in CHAPS Sterling, and reflected in this model, adequate when there are more than two banks in the payments system?

6.1 More frequent throughput requirements

It is relatively straightforward to extend the two-period analysis to deal with multiple throughput requirements. Consider the case of $(N$-$1)$ identical throughput requirements, $\tau'\mu$. For simplicity,

[17] If this condition is not satisfied, a sustainable payment system could still be achieved by setting the throughput requirement higher than the optimal level for the bank wishing to make the fewest payments. This would mean that the bank would cancel fewer payments than it would optimally choose to cancel, which would allow the end-of-day liquidity requirement to be eliminated.

we assume that both banks desire to hold the same level of collateral, $c^* = \tau'\mu$, whether or not they wish to cancel any payments. Bank profitability is then given by:

$$\Pi_i = F_i Min\left[\mu, N\tau'\mu\right] - \gamma_i \tau'\mu - \beta_i \left(Max\left[\mu - N\tau'\mu, 0\right]\right)^2$$

Proposition 11 When $NF_i > \gamma_i$, neither bank cancels any payments and $c^* = \frac{\mu}{N}$. Bank profitability is then increasing in N:

$$\Pi_i = (F_i - \frac{\gamma_i}{N})\mu$$

Proof. As for the two-period case with one throughput rule.

Proposition 12 When $NF_i < \gamma_i$, and assuming both banks cancel the same amount of payments, the level of optimal collateral is given by:

$$c^* = \frac{1}{N}\left(\mu - \frac{(\frac{\gamma_i}{N} - F_i)}{2\beta_i}\right)$$

Cancelled payments fall with increased N. Bank profitability is given by:

$$\Pi_i = \frac{\left(\frac{\gamma_i}{N} - F_i\right)^2}{4\beta_i} - \left(\frac{\gamma_i}{N} - F_i\right)\mu$$

and increases with N.

Proof. The relevant first-order condition is:

$$NF_i - \gamma_i + 2N\beta_i(\mu - N\tau'\mu) = 0$$

Solving for the throughput requirement, which is equal to the optimal collateral, yields the required result. We have assumed that the necessary condition for the cancelled payments by each bank to be equal. Cancelled payments are $\frac{(\frac{\gamma_i}{N} - F_i)}{2\beta_i}$, which fall in proportion to N^2. Profits are the sum of revenue minus opportunity costs minus the cost of cancelled payments. Differentiating this, we find that profits increase with N so long as:

$$\mu > \frac{(\frac{\gamma_i}{N} - F_i)}{2\beta_i}$$

which must be satisfied since banks cannot cancel more payments than they have to make.

So how should the system operator choose to set the number of throughput requirements? The simplistic answer is to say that it should be as large as possible given the technology for processing payments. Increasing the frequency of throughput requirements increases the payments velocity (as defined by Schoenmaker (1995)) and will limit the intraday liquidity costs of collateralised RTGS systems. In the limit, all payments could be made with an infinitesimal amount of collateral recycled infinitely frequently during the day. As we have seen, increasing N also facilitates greater competition in the payments business by changing the threshold at which banks choose to cancel payments or not. As N increases, banks with higher and higher opportunity costs of collateral will be able to enter the payments business profitably.(18)

There are two reasons why there will be an upper bound on N. First, actual payments are not infinitely divisible and can often be large relative to the size of collateral posted. If banks were forced to make payments greater than the throughput requirement in any given period because of such a large payment, it could reintroduce some of the externalities that the throughput rules were initially designed to dispel.(19) A second problem arises when we consider that real payment flows are actually stochastic. As the number of throughput requirements increases, eventually a point will be reached where one or more of the banks will not have sufficient payments to meet the final throughput requirement whenever they face a low payments realisation.

If, within these constraints, a large number of throughput rules is better than a smaller number, why does the UK sterling RTGS system have only two throughput requirements? Here it is important to draw the distinction between the model and a real payment system. While there are only two formal throughput requirements in CHAPS Sterling, payments can be made almost continuously through the day whereas in the model payments could be made only once during each period. The throughput rules in CHAPS Sterling do not specify what banks should do at the same level of precision as the rules in our model. In CHAPS Sterling, the throughput rules place a lower bound on the degree of recycling that must take place and help to limit the volatility of payment flows both within and across days. Within these parameters, banks may choose to recycle liquidity more or less quickly, depending on their own and their competitors' characteristics, for example the value of time urgent payments banks need to make and their attitude to operational risk. Liquidity may therefore be recycled much more frequently in CHAPS Sterling than a simple (and incorrect) extrapolation of our model would suggest. Indeed, the value of payments made in CHAPS Sterling is typically around ten times the value of intraday liquidity that banks obtain through intraday repos. We do not know, however, whether or not this represents full efficiency.

(18) There will be diminishing marginal benefits to increasing N, however.

(19) Splitting payments into smaller sizes might help reduce this problem, but may give rise to others, for example ensuring that all parts of a particular payment had been settled.

6.2 Throughput rules with many banks

In the two-bank model, the existence of a throughput requirement works symmetrically and each bank is certain to receive $\tau\,\mu$ in each period that the throughput rule applies. This changes radically in a three-bank model if the throughput rules are enforced on the aggregate level of payments, rather than the level of payments made to individual settlement banks. With an aggregate rule, there is no guarantee that all banks will receive their 'fair' share of payments in periods when the throughput rules apply. For example, in a system with just three banks, two of the three could collude and meet the throughput rules by making payments to each other while delaying payments to the third bank. The third bank would, however, face no choice but to make its payments to meet its own throughput target. Such collusion would increase the third bank's requirement for intraday liquidity and, if such behaviour persisted over time, it might make it uncompetitive for the third bank to remain in the payments business (or even to enter in the first place).

This seems a potentially important weakness in the relatively simple aggregate throughput rules in place in the CHAPS Sterling system, although we have no evidence to suggest that banks are behaving in this way in practice. If it were a problem, one way to overcome it might be to refine the throughput requirements so that they applied on a bilateral basis. This might, however, give rise to some practical difficulties, depending on the pattern and timing of bilateral payments between banks. In particular, small banks with a low level of payments to any one bank might find it difficult to guarantee meeting such bilateral throughput rules.

7 Concluding remarks

We have developed a simple two-bank, two-period model, of an RTGS system with collateralised intraday credit. Using this model, we demonstrated that, when banks care about payments imbalances between them, the aggregate liquidity requirement in the payment system will be inefficiently high due to banks' concerns over free-riding. While a co-operative outcome in which banks recycle liquidity between themselves during the day would both increase bank profitability and payment system performance, this may not be achievable without some form of regulation. One problem is that the structure and transparency of the payment system may be such that co-operation between banks is inherently fragile because of the lack of reliable triggers for the punishment of free-riders or differences in the opportunity cost of collateral between banks.

Another potential impediment is oligopolistic competition within the payments industry. By increasing the degree of liquidity recycling, incumbent banks may make it easier for higher-cost banks to enter the payments market, with the usual negative impact on the incumbents' profitability. In certain circumstances, this may prompt incumbent banks to limit the degree of liquidity recycling in the system.

Since banks do appear to care about payments imbalances between them, for example by operating bilateral limits on net payments to other banks, the case for some degree of regulation appears to be strong. In the United Kingdom this has taken the form of formal throughput requirements specifying the proportion (by value) of payments that must be made by specific times of the day. Throughput requirements force liquidity recycling by all settlement banks in proportion to their level of payments. The total intraday liquidity available to settlement banks is therefore always greater than the amount generated from their own collateral holdings alone. Incorporating such rules into our model allows banks to achieve the Pareto dominating efficient outcome. Additionally, by lowering the threshold at which banks can profitably enter the payments business, such throughput rules are potentially a powerful way of increasing competition in the payments industry, encouraging a higher degree of direct access to payment systems. Consequently, throughput rules could have risk-reduction benefits to the extent that they help to reduce the level of tiering in the financial system.

Our model suggests that increasing the number of throughput requirements would enhance efficiency indefinitely, although at an ever diminishing rate. It seems likely, however, that there is some upper limit to the efficient number of throughput rules, for at least two reasons. First, real payments have a finite size and are sometimes very large and urgent, whereas our model treats them as infinitely divisible and identical in their importance. If the payments throughput demanded in a given period were less than the size of such large payments, banks might be forced into the position where they were once again making more payments to other banks than they were receiving back – the original problem throughput rules were designed to solve. The second reason is that real payments between banks are stochastic and partially unknown at the start of the day. As the number of throughput requirements increases, eventually a point will be reached where one or more of the banks will not have sufficient payments to meet the final throughput requirement whenever they face a low payments realisation.

It is important to recognise, however, that the degree of liquidity recycling in a real RTGS payment system is somewhat more complicated than that in our model. Banks are able to make payments continuously, rather than once in each period in the model. Throughput rules in a real payment system such as CHAPS Sterling seem to place a lower bound on the degree of liquidity recycling in the system and help to limit the volatility of payment flows both within and across days. Within the parameters set by these rules, banks may then choose to recycle liquidity more or less quickly, depending on their own and their competitors' characteristics, for example the value of time-urgent payments that banks need to make, as well as their attitude to operational risk. Liquidity seems to be recycled much more frequently in CHAPS Sterling than a simple (and incorrect) extrapolation of our model would suggest.

A further potentially important point to arise from our analysis is that aggregate throughput rules may not be sufficient to prevent banks from forming cartels to disadvantage other banks or new potential entrants. While we are not aware of any evidence to suggest that this is a problem in CHAPS Sterling, this issue is one that RTGS payment system designers and overseers should be alert to. Bilateral throughput rules would be one way to deal with such anticompetitive behaviour, although these might be difficult to implement in a system with a large number of small banks with unpredictable payments patterns.

References

Angelini, P (1998), 'An analysis of competitive externalities in gross settlement systems', *Journal of Banking and Finance*, Vol. 22, pages 1-18.

Bech, M and Garratt, R (2001), 'The intraday liquidity management game', *USCSB Working Paper*.

Freeman, S (1996), 'The payments system, liquidity, and rediscounting', *American Economic Review,* Vol. 86 (5).

Green, E (1999), 'Money and debt in the structure of payments', *Federal Reserve Bank of Minneapolis Quarterly Review*, Vol. 23, No. 2.

Kahn, C M and Roberds, W (2001), 'Real-time gross settlement and the costs of immediacy', *Journal of Monetary Economics*, Vol. 47, pages 299-319.

Kobayakawa, S (1997), 'The comparative analysis of settlement systems', London, *Centre for Economic Policy Research Discussion Paper*, No. 1667.

McAndrews, J (1999), 'Panel: Thoughts on the future of payments and central banking', *Journal of Money, Credit and Banking*, Vol. 31, pages 671-73.

McAndrews, J and Rajan, S (2000), 'The timing and funding of Fedwire funds transfers', *FRBNY Economic Policy Review*, Vol. 6, No. 2.

McAndrews, J and Trundle, J M (2001), 'New payment system designs: causes and consequences', *Financial Stability Review*, Bank of England, December.

Rochet, J-C and Tirole, J (1996), 'Controlling risk in payment systems', *Journal of Money, Credit and Banking*, Vol. 28(4), pages 833-62.

Schoenmaker, D (1995), ‘A comparison of alternative interbank settlement systems’, London, *LSE Financial Markets Group Discussion Paper Series*, No. 204.

Zhou, R (2000), ‘Understanding intraday credit in large-value payment systems’, Federal Reserve Bank of Chicago, *Economic Perspectives*, Third Quarter.