Christopher Becher, Marco Galbiati, and Merxe Tudela

The Timing and Funding of CHAPS Sterling Payments

- Participants in CHAPS Sterling often use incoming funds to make payments, a process known as liquidity recycling.
- Liquidity recycling can be problematic if participants delay their outgoing payments in anticipation of incoming funds.
- An analysis of CHAPS payment activity shows that the level of liquidity recycling, though high, is stable throughout the day—a condition attributable to three features of the system.
- First, the settlement of time-critical payments in CHAPS supplies liquidity early in the day—liquidity that can be recycled to fund less urgent payments.
- Second, CHAPS throughput guidelines provide a centralised coordination mechanism that essentially limits any tendency toward payment delay.
- Third, the relatively small direct membership of CHAPS facilitates coordination, enabling members to maintain a constant flux of payments during the day.

1. Introduction

The use of real-time gross settlement (RTGS) systems for the settlement of large-value payments offers considerable advantages, the principal one being the elimination of the credit risk that can arise between participants in deferred net settlement systems. However, in comparison with deferred net settlement systems, RTGS systems require relatively large amounts of liquidity to support payment activity. This liquidity can be sourced from the settlement agent (usually a central bank in the case of large-value payments systems)—in the form of intraday overdrafts—or from incoming payments from other participants.

Obtaining intraday liquidity from a central bank is typically costly. In order to minimise this cost and to take advantage of incoming payments as a funding source, participants may choose to delay outgoing payments. However, payment delay may itself prove costly. Participants face a trade-off, therefore, between the cost of borrowing from the central bank and the expected cost of delaying payments. McAndrews and Rajan (2000) explore this trade-off in a study of payment behaviour in Fedwire. They describe how the use of incoming funds to offset outgoing payments allows participants to avoid incurring costly overdrafts from the central bank and hence reduces the liquidity cost of making payments. Such offsetting can be achieved to a greater extent during activity peaks, so banks are induced to coordinate their payments around, and thereby to reinforce, these peaks.

Christopher Becher, a policy officer at the European Commission, was an analyst at the Bank of England at the time this article was written; Marco Galbiati is an economist and Merxe Tudela a senior economist at the Bank of England.

Correspondence: <marco.galbiati@bankofengland.co.uk>

The authors thank Mark Manning, James McAndrews, participants at workshops at the Bank of England, and an anonymous referee for invaluable input. The views expressed are those of the authors and do not necessarily reflect the position of the Bank of England, the European Commission, the Federal Reserve Bank of New York, or the Federal Reserve System.

This article investigates the factors influencing the timing and funding of payments in the CHAPS Sterling system, drawing where appropriate on comparisons with payment activity in Fedwire. In the next section, we discuss theoretical approaches to the study of payment behaviour and their application to CHAPS Sterling. The empirical analysis of the timing and funding of CHAPS Sterling payments follows in Sections 3 and 4, respectively. Section 5 concludes.

2. Theoretical Studies of Payment Behaviour

Several theoretical studies have addressed the incentives facing participants in RTGS systems. Many focus on the afore-mentioned trade-off between the cost of liquidity and the expected cost of delaying payments.

2.1 Definition of Terms

The measurement of the *cost of liquidity* varies according to the regime employed by the settlement agent (in the cases described in this article, that agent is the central bank). When credit is supplied unsecured, the cost typically takes the form of an explicit overdraft fee. Credit may also be provided against eligible collateral, in which case the cost to the participant is the opportunity cost of posting eligible securities with the central bank and hence forgoing alternative uses for those assets.

The *cost of delay* may take several forms. Financial penalties may be incurred for failure to make time-critical payments by specified deadlines, such as for settlement payments in ancillary systems or repayments of interbank loans. In addition, failure to make customer payments on time, or indeed at all, on the intended settlement date may result in reputational costs and a loss of future business. Also, as we discuss later, the reputation of a participant within a payments system may suffer if it is perceived to be delaying payments in order to "free-ride" on liquidity provided by others.

2.2 Theoretical Approaches

Bech and Garratt (2003) model the trade-off using a game-theoretical approach, analysing the behaviour of two banks, both of which receive random payment requests from customers at the beginning of a morning and an afternoon period. Both banks face a fixed cost of delaying payments and of posting collateral for a morning or afternoon period. The analysis is repeated under priced and collateralised intraday liquidity regimes, as employed in Fedwire and CHAPS, respectively.[1]

Under a collateralised regime, Bech and Garratt find that both early and delayed payments are possible equilibria, depending on the relative costs of liquidity and delay. The efficient equilibrium is for both banks to pay early. However, for certain levels of delay and liquidity costs, the participants are found to be in a prisoner's dilemma, in which the dominant strategy for both is to delay payments until the afternoon, even though both would benefit if payments were made in the morning. This incentive to delay arises because it is possible to avoid the cost of posting collateral in the morning and instead to incur the (cheaper) delay cost. In these cases, there would be a welfare improvement if the participants could be induced to coordinate and to pay earlier.

This article investigates the factors influencing the timing and funding of payments in the CHAPS Sterling system, drawing where appropriate on comparisons with payment activity in Fedwire.

Under a regime of priced credit, Bech and Garratt again find that multiple equilibria are possible. However, in this case, participants stand to benefit from synchronising payments with each other, since no cost is incurred by either participant if payments are "offset" within the time period over which overdraft fees are calculated. The equilibrium outcome will thus depend not only on the relative costs of liquidity and delay but also on the likelihood that the other bank will receive a payment request. In the specific case where the expected cost of delay is lower than the credit fee, and payment flows are skewed toward the afternoon, Bech and Garratt find that the efficient equilibrium involves delay until the afternoon.

In a similar study, Kobayakawa (1997) models the choice of whether to delay payments in RTGS systems under varying intraday credit arrangements. Again, the relative costs of liquidity and delay drive equilibrium selection. Under a system of priced credit, Kobayakawa (like Bech and Garratt) finds that delayed settlement is an equilibrium, since each participant seeks to avoid incurring an overdraft by delaying payments and thereby "free-riding" on the liquidity provided by the other participant. Under a collateralised regime, Kobayakawa finds a unique equilibrium in which both participants pay early.

[1] Bech and Garratt also examine the case of free intraday credit; however, those results are not discussed here.

However, in this case, the result is obtained by assuming that the opportunity cost of collateral is a sunk cost, and so liquidity is in effect free when the game is played. This is unsatisfactory, since it takes no account of the participants' incentives to reduce the cost of liquidity by economising on the value of collateral posted. Consequently, we focus in the following discussion on the Bech and Garratt model.

Mills and Nesmith (2008) adapt the Bech and Garratt model to look at the effect of settlement risk on timing decisions in payments and securities settlement systems, concentrating on the differential impact of overdraft costs on the two types of systems. A main contribution of this paper is to describe a rationale for delays overlooked in the literature: namely, that banks may withhold payments until they receive information on the others' ability to send funds, in order to obtain a better forecast of the costs of funding their own payments. More precisely, in the model, banks choose between paying "early" and paying "late." There are no delay costs, so, in the absence of settlement risk, "early" is a *weakly* dominated strategy,[2] which may still appear in equilibrium (although only in risk-dominated ones). Introduction of settlement risk definitely tilts the balance against the "early" strategy, because

The overarching conclusion of these [theoretical] works is that institutional features and, in particular, intraday credit regimes have a powerful effect on banks' incentives; as a consequence, they largely determine a payments system's performance.

in this case "early" becomes a *strictly* dominated strategy ("late" outperforms it against *any* action by the opponent, as settlement risk imposes some overnight overdraft in probability terms). Thus, settlement risk introduces further reasons to delay, eliminating the "early payment" equilibria.

Building on previous work, Martin and McAndrews (2008) construct a model with a continuum of banks, each making a unit payment to one other bank and each having to decide whether to pay "early" or "late." Banks are assumed to face random delay and liquidity costs, determined in turn by bank-specific shocks that drain (or increase) the available liquidity. The paper shows that, depending on the cost parameters (costs of delay and of overdrafts), on the time-criticality of payments, and on the probability and size of liquidity shocks, the resulting equilibria feature different degrees of delay. More specifically, some or all banks, depending on the shocks received, decide to delay their payments. Martin and McAndrews also explore the effect of a liquidity-saving mechanism on the banks' incentives. This is shown to mitigate the strategic complementarity of banks' strategies by allowing banks to release payments conditional on the receipt of payments. The paper shows that, in this case, the extent of delay in equilibrium also depends on the pattern of payments (whether payments can be offset in pairs or multilaterally), which has implications for the system's efficiency.

The overarching conclusion of these works is that institutional features and, in particular, intraday credit regimes have a powerful effect on banks' incentives; as a consequence, they largely determine a payments system's performance. We draw on—and, where necessary, modify—these theoretical predictions to study payment behaviour in Fedwire and in CHAPS Sterling.

2.3 Implications for Payment Behaviour in Fedwire

The Federal Reserve System supplies intraday liquidity to Fedwire members in the form of uncollateralised daylight overdrafts. Subject to net debit caps, participants can incur overdrafts at any time, which incur a charge calculated as the average per-minute overdraft during the day, multiplied by an effective daily rate, less a deductible.[3]

As Bech and Garratt (2003) describe, a corollary of this charging structure is that participants can avoid overdraft charges by synchronising payments. As long as incoming payments of at least equivalent value to outgoing payments are received within a minute, no overdraft will be required and hence no charge incurred. This sets the scene for a pure coordination game, in which participants attempt to synchronise payments in order to minimise the average overdraft position over the course of the day. The theory would also predict that if the overdraft fee is deemed to be high relative to the expected cost of delay, the efficient equilibrium will involve the delay of payments until later in the day.

This theoretical finding is consistent with the empirical results obtained by McAndrews and Rajan (2000) on the timing and funding of payments in Fedwire. Faced with costly intraday liquidity, participants appear to delay payments until an end-of-day activity peak, during which the probability of

[2] "Early" performs no better than "late," and it performs just as well as "late" if the other also pays early, as offsetting payments incur no charge.

[3] The effective rate is currently equivalent to an annual rate of 36 basis points.

receiving funds from other participants is greater. It is argued that this synchronised delay reinforces the activity peak. The "focal points" for this coordination appear to be provided by ancillary system settlement deadlines (in particular, in CHIPS and DTC). As McAndrews and Rajan note, though, the outcome of this apparent coordination may not be socially efficient, since all participants might stand to benefit from reduced liquidity costs if coordination could be improved so as to take full account of liquidity externalities.

Armantier, Arnold, and McAndrews (2008) extend this analysis in their study of recent changes in the timing of Fedwire funds transfers. Among other things, they explore alternative hypotheses for the timing of late-afternoon payment peaks, with particular reference to the change in the timing of late-afternoon Fedwire transfers following a move to a later CHIPS settlement time. The tendency for Fedwire transfers to be made after ancillary system positions are settled may reflect the "focal point" hypothesis described above. However, it may also be that the liquidity released by ancillary system settlement may trigger a "cascade" of payments, to the extent that participants are liquidity-constrained.[4] Along similar lines, the settlement may also release credit lines, thereby permitting more payments to be made. Additionally, the authors suggest that uncertainty surrounding the size of ancillary system payouts may lead to payments being deferred until after the settlement deadline—that is, once uncertainty has been resolved. The data do not allow for a clear distinction to be made between the competing hypotheses; however, there is sufficient evidence to suggest that the coordination described in the earlier paper is only part of the story.

2.4 Implications for Payment Behaviour in CHAPS Sterling

The Bank of England provides intraday liquidity to members of CHAPS Sterling in the form of interest-free overdrafts secured against eligible collateral. The maximum value of liquidity granted is equal to the value of collateral securities posted, less a "haircut" to take account of movements in the value of the collateral securities. In contrast with Fedwire, where the total cost of liquidity is driven by the *average* overdraft incurred, the cost of liquidity in CHAPS Sterling is driven by the *maximum* overdraft position incurred during the day, since the value of collateral posted must be at least equal to this position.

The cost of posting collateral derives from the fact that the securities posted (or the funds used to obtain the required securities) could be used for alternative purposes; participants therefore face an opportunity cost. As described in Box 1, the upper bound to this cost has been estimated to be of the order of 7 basis points per annum, although for domestic banks subject to the Stock Liquidity Regime the opportunity cost may be significantly lower and may even approach zero.

The Bech and Garratt (2003) model predicts that this regime will result in multiple equilibria, with the selection of an equilibrium dependent on the relative magnitudes of the cost of delayed payment and the opportunity cost of posting collateral. The low opportunity cost of posting collateral for many CHAPS Sterling members may thus be expected to favour an early rather than a delayed payment equilibrium. That said, it is difficult to quantify the cost of delay associated with all but a small number of time-critical payments. Anecdotal evidence suggests that costs of delay are low for the majority of payments.

In contrast with Fedwire, where the total cost of liquidity is driven by the average overdraft incurred, the cost of liquidity in CHAPS Sterling is driven by the maximum overdraft position incurred during the day, since the value of collateral posted must be at least equal to this position.

Certain qualifications are required in applying this model to CHAPS Sterling. In particular, in the Bech and Garratt model, the benefit from delaying payments in a collateralised regime derives from the assumption that it is less costly to post collateral for the afternoon than for the whole day (and hence that there is an incentive to avoid posting collateral in the morning). This in turn rests on the assumption that it is possible to invest surplus liquidity for a fraction of the day—or, in other words, that there exists an intraday market for liquidity. It is not obvious that this incentive applies in CHAPS Sterling since, in the absence of an intraday market, it is probably no cheaper to post collateral for a morning or afternoon than for a full day. Once collateral is committed to the payments system, the cost for the full day is incurred.[5]

It is possible to modify the Bech and Garratt model to incorporate an incentive to delay that does not rely on the existence of an intraday market for liquidity. By delaying

[4] See also Beyeler et al. (2006).

[5] It is nonetheless possible for banks to withdraw liquidity intraday. As we argue here, while it may not be possible to lend in an intraday interbank market, the collateral could in principle be committed to another payments system. In this case, the ability to commit collateral for only part of a day could be considered valuable.

payment and taking advantage of incoming funds, a participant may be able to reduce the maximum overdraft position and hence reduce the aggregate collateral requirement for the day (or avoid posting collateral altogether). It can be shown that a similar prisoner's dilemma outcome emerges from this model, with both participants defecting despite the mutual benefit from paying early, unless they can somehow be induced to coordinate earlier in the day.[6]

However, the finding that some participants may seek to reduce their aggregate collateral requirements by delaying payments must be seen in the context of the empirical observations that participants in CHAPS typically post collateral at the beginning of the day (that is, they do not generally wait to determine whether collateral posting is required) and many post collateral to a value well in excess of liquidity usage (for the system as a whole, maximum liquidity used is only around one-third of the maximum collateral posted).[7] Two factors appear to be particularly influential in explaining this behaviour.[8]

First, as discussed in Box 1, the low opportunity cost of posting collateral (or of maintaining positive reserve account balances) means that, for many banks there appears to be little incentive to delay posting collateral until later in the day, since the potential savings to be made from reducing the aggregate value of collateral posted are small.

Second, the distribution of "time critical" payments, for which the expected cost of delay is high, appears to be skewed toward the morning in CHAPS Sterling. For example, pay-ins to CLS Bank must be made by 11:00 a.m. in order to avoid significant financial penalties, and market convention dictates that overnight interbank loans should be repaid the following morning. Even for those banks for which collateral posting is relatively costly, the expected cost of delay for time-critical payments may be so high as to warrant posting sufficient collateral at the beginning of the day to ensure that liquidity is available to make these payments without the need for recourse to incoming funds. The existence of throughput guidelines may serve to reinforce the incentive to post liquidity "up front" (as we discuss in more detail in Section 2.5).

[6] We are indebted to Peter Gibbard for these insights.

[7] It should be noted that CHAPS Sterling payments are not the only claim on the available liquidity. Participants in the United Kingdom's securities settlement system, CREST, are able to transfer liquidity from CHAPS Sterling settlement accounts to separate accounts designated for the settlement of the cash legs of securities transactions. Liquidity is also available for the settlement of positions in other ancillary systems, such as CLS Bank, retail payments systems (BACS and C&CC), and LCH.Clearnet. Unlike CLS Bank pay-ins, the latter transfers do not take place in CHAPS Sterling and therefore are not recorded in the payment data in this article.

[8] In addition to the factors described, there may be frictions associated with obtaining eligible securities during the day that tend to encourage early posting. Under normal circumstances, intraday repos with the Bank of England are unwound at the end of the day and the securities are held in custody by the Bank of England overnight, to be reposted the following morning.

Box 1

The Cost of Liquidity in CHAPS Sterling

James and Willison (2004) estimate the opportunity cost of posting eligible collateral as the difference between the unsecured interbank rate and the secured-lending repo rate. By posting collateral, the bank forgoes the opportunity to repo the securities and to lend the funds obtained at a higher rate in the interbank market.

As James and Willison acknowledge, this is only part of the story. U.K. banks are subject to the Stock Liquidity Regime (SLR), under which they are required to hold liquid assets sufficient to cover net sight deposit and five-day wholesale cash outflows. These assets cannot be repurchased overnight to generate cash in the interbank market, but they can be posted with the Bank of England to generate liquidity in CHAPS Sterling (since the SLR requirements are measured only at the end of the day). For those banks subject to the SLR, the opportunity cost of posting collateral—and hence the cost of liquidity—may be even lower than the 7 basis point estimate.

Three foreign settlement banks in CHAPS Sterling—accounting for around 14 percent of transactions by value or 11 percent of transactions by volume—are not subject to the SLR. These banks are instead subject to the Maturity Mismatch Regime, which does not require banks to hold eligible liquid assets if committed outflows equal expected inflows. For these banks, the opportunity cost of posting collateral may be higher. However, the use by these banks of cross-border collateral arrangements (including the ability to back sterling payments with euro cash collateral) implies that estimation of the cost of liquidity for foreign banks would require analysis of the cost of generating liquidity in other jurisdictions.

Following reform of the Bank of England's money market operations in May 2006, participants are now able to hold remunerated reserve account balances with the Bank of England. These balances can be used to fund payments, so members can choose to provide liquidity in this way rather than by posting eligible securities. The relative opportunity cost of holding reserve account balances may vary from member to member, although for some foreign banks, particularly those that do not routinely hold sterling collateral, reserve balances may be a relatively attractive source of liquidity.

The value of collateral posted to support these time-critical payments depends on the expectation of the size of time-critical payment flows. Consider a single period in which delay costs are zero up to a certain time (say, the deadline for a time-critical payment) and very high thereafter. If all payment instructions are known at the beginning of the period in which time-critical payments must be made—and therefore all banks are aware of whether they will be net payers or net receivers at

the end of the period—it can be shown that each net payer must raise liquidity at least equal to the value of its net payments and that this amount will be necessary and sufficient to settle all payments in the system (Box 2). Net receivers will be able to meet their payment obligations using incoming funds and hence will not need to raise additional liquidity.

Box 2

Liquidity Requirements for Time-Critical Payments

For simplicity, assume that banks receive all payment instructions exogenously from their customers. These instructions are denoted by x_t^{ij}—that is, at time t, bank i is requested by its customer(s) to pay the amount x to bank j. A payment from bank i to bank j at time t is denoted p_t^{ij}. Banks choose whether to settle payment instructions immediately or to queue them internally. So, p_t^{ij} and x_t^{ij} need not be the same.

Consider the case in which delay costs are zero up to a certain time T and subsequently so high that all payments must be settled within T. It follows that:

$$\sum_{t=0..T} x_t^{ij} = \sum_{t=0..T} p_t^{ij} \quad \forall i,j\,.$$

The payment balance of bank i against bank j at time t is defined as:

$$b_t^{ij} = \sum_{s=0..t} p_s^{ji} - \sum_{s=0..t} p_s^{ij}\,.$$

Bank i is a net payer for the period if its total payment balance at T is negative—that is, if $b_T^i = \sum_{j \neq i} b_T^{ij} < 0$. Since customer orders are exogenous, banks cannot affect whether they will be net payers or net receivers. We assume, however, that banks know with certainty at the beginning of the period which type they will be.

We define I as the set of net payers. Each net payer i has to raise liquidity to a value at least equal to $L^i = -\sum_{j \neq i} b_T^{ij}$ in order to execute its payment instructions. Hence, $\sum_{i \in I} L^i$ is the minimum liquidity required to settle all payment instructions. This amount will also be sufficient to settle all payments by time T if, first, every net payer raises L_i at time zero and pays it out immediately and, second, at any t, every bank i uses all of its liquidity to make payments up to that value (or less, up to the exhaustion of queued orders).[a] Because delay costs are zero up to T, this pattern is optimal for all banks. We can therefore conclude that, when delay costs are zero up to a time-critical threshold and very high thereafter, the banks' interests are aligned and compatible with the efficient use of liquidity. All payments are settled using only the minimum liquidity L.

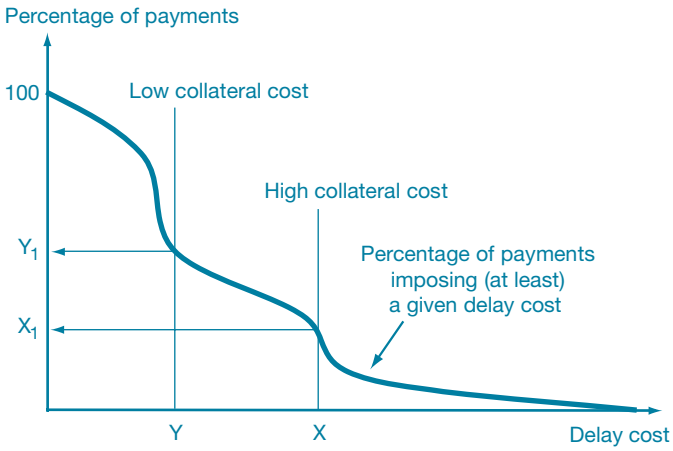
[a] We are abstracting here from 1) the indivisibility of payments—that is, additional liquidity may be required if payments cannot be split and settled in tranches, and 2) the possibility that no bank is a net payer—meaning that all payments net out exactly. In this case, a bargaining process would be required to define who is to provide liquidity, given that liquidity is required, yet no one needs to post collateral if somebody else does.

Banks' incentives are therefore aligned and consistent with the efficient use of liquidity.

However, this result requires that all participants know at the beginning of the period whether they will be net payers or net receivers at the end. In reality, participants will not typically possess full information about payment flows at the time when collateral posting decisions are made, and hence they will face uncertainty about liquidity requirements. Faced with high threshold delay costs, participants will wish to insure themselves against the risk of being net payers and—if the cost of failing to make a time-critical payment is sufficiently high—they may choose to post liquidity at the beginning of the period to a value at least equal to the maximum anticipated gross value of the time-critical payments. For payments with low delay costs, by contrast, participants may be willing to rely on incoming funds rather than post additional collateral. We illustrate this scenario using a simple stylised framework (see exhibit).

Here, the choice of the value of collateral posted at the beginning of the day is determined by the intersection of the expected cost of delay (which varies across payments; in the exhibit, X_1 percent of payments incurs a delay cost of at least X) and the cost of collateral (which, as described above, is fixed once collateral is posted). The value of collateral posted must be sufficient to ensure that time-critical payments—for which the cost of delay is greater than the cost of liquidity—can be made without the need for recourse to incoming funds. By contrast, for those payments for which the cost of delay is lower than the opportunity cost of posting collateral (the proportion of payments 100 percent *minus* X_1 for participant X), participants may be willing to rely on the recycling of incoming funds instead of posting additional liquidity. Such payments—particularly those of high value—will typically be delayed until after time-critical payments are settled, especially when there is uncertainty over the liquidity demands of time-

Delay and Liquidity Costs in CHAPS Sterling System



critical payments.[9] In addition, to the extent that there is reliance on incoming funds, the liquidity released by the settlement of time-critical payments may result in a "liquidity cascade," as queued payments are released. In this way, the early settlement of time-critical payments may serve to catalyse liquidity recycling later in the day.

The proportion of payments to which this applies will vary according to the cost of liquidity. As the exhibit illustrates, banks for which the opportunity cost of posting collateral is relatively low (for example, cost of collateral Y) will post a larger stock of collateral and hence may tend to fund a greater volume of payments (Y_1) from posted liquidity than from incoming funds. However, banks for which liquidity is relatively costly may post less collateral and remain more reliant on the recycling of incoming payments to fund outgoing payments.

Of course, as in the single-period example above, we must also take account of the limited information available to participants when decisions are made. While a proportion of payment instructions may be known at the start of the day (which we would expect to be relatively large when payment activity is driven by proprietary rather than customer business), there may remain considerable uncertainty about the size and distribution of incoming and outgoing payments, including payments with high delay costs. Faced with such uncertainty about aggregate liquidity demands, participants would be expected to hold a buffer of liquidity in excess of the quantity predicted in this framework in order to withstand unforeseen liquidity demands. The tendency to maintain such liquidity cushions—which are indeed observed in practice—will also contribute to the entire system's resilience to liquidity shocks, such as the operational failure of one or more banks to make payments.

2.5 Liquidity Recycling in CHAPS Sterling

We have discussed how the apparently low opportunity cost of collateral for many participants and the high expected delay costs associated with a subset of payments will tend to favour posting collateral at the beginning of the day. This may serve to reduce the incentive for payment delay and hence avoid the prisoner's dilemma outcomes described in the theoretical models. However, we have also seen that for payments for which the cost of delay is relatively low, participants may rely to a greater extent on incoming funds as a funding source in order to avoid squeezing the precautionary buffer of spare liquidity—or indeed to avoid posting additional collateral.

The efficiency with which incoming funds are recycled will depend on the extent to which participants collectively maintain the flow of liquidity around the system, perhaps via proactive payment coordination. McAndrews and Rajan (2000) describe how such coordination is achieved in Fedwire through the delay of payments until an end-of-day activity peak, but suggest that the observed level of coordination may be inefficiently low—and hence liquidity costs inefficiently high—due to collective-action problems, exemplifying the uncooperative outcome in the prisoner's dilemma game described by Bech and Garratt (2003). In principle, CHAPS Sterling members could suffer from a similarly uncooperative outcome in which some members defect and withhold payments, thereby curtailing the ability of others to take advantage of incoming funds.

The efficiency with which incoming funds are recycled will depend on the extent to which participants collectively maintain the flow of liquidity around the system, perhaps via proactive payment coordination.

In practice, the early settlement of time-critical payments will contribute to the recycling of liquidity by ensuring that payments begin to flow early in the day. In addition, certain features of CHAPS Sterling may be particularly conducive to achieving a cooperative outcome and hence to ensuring that liquidity is recycled efficiently. For example, the CHAPS Clearing Company imposes a set of *throughput guidelines*, whereby participants are expected to make 50 percent of payments by value by 12:00 p.m. and 75 percent by 2:30 p.m., as an average over a calendar month, with the explicit intent of improving the efficiency of liquidity usage in the system. While enforcement of the guidelines relies on peer pressure rather than legal compulsion (Box 3), the guidelines are largely observed in practice. As Buckle and Campbell (2003) demonstrate, the existence of such guidelines acts to countervail any tendency toward payment delay and hence serves to promote liquidity recycling earlier in the settlement day, thereby enhancing the efficiency of the payments system.

[9] Empirically, banks tend to settle a large volume of low-value payments early in the day. This might appear to contradict the predictions of this model, since the inherent delay cost associated with any one of these payments is likely to be low. However, expected delay costs for these payments *collectively* may be high, since processing of large volumes later in the day may prove difficult and costly in the event of an operational incident.

Box 3
Enforcement of CHAPS Throughput Guidelines

If a CHAPS Sterling member breaches the throughput guidelines in three consecutive months, that member is required to provide reasons to the CHAPS Clearing Company and to outline the steps taken to ensure that deadlines are met going forward. The participant will be given the opportunity to provide evidence that, over the period in question, failure to meet the guidelines resulted from a lack of payment instructions rather than a shortage of available liquidity.

If the member breaches the guidelines in six consecutive months, or in three consecutive months on two occasions, and has been unable to provide evidence as set out above, it will be obliged to attend a "Star Chamber" hearing. At the Star Chamber, the member's CHAPS board director will be required to explain the steps being taken to resolve the issues and to return performance to acceptable service levels and guidelines.

There is no defined penalty for the breach: As a rule, peer pressure is felt to be sufficient. However, the CHAPS Rules give the company manager the power to suspend or exclude a member "in material breach" of the provisions of the procedural rules, or where, in the opinion of the CHAPS Clearing Company, circumstances have arisen that could be "prejudicial" to the system or represent a threat to its "security, integrity, or reputation."

In addition, the concentrated structure of CHAPS Sterling appears more conducive to coordinated behaviour than does the structure of Fedwire, which has a broader membership. CHAPS has fifteen direct members (including the Bank of England), and the majority of payments are made by a core of four participants. The Fedwire network is more extensive, as around 9,500 participants access the clearings directly.[10] This results in very different network topologies, which in turn has implications for the flow of liquidity around the systems.[11] In particular, the concentration of payment flows among a small group of banks in CHAPS Sterling leads naturally to a higher level of recycling throughout the day than would occur in a more dispersed system, since each unit of liquidity paid out is more likely to be returned quickly if payments are flowing between fewer banks. Furthermore, within a "small club" of participants, the behaviour of each participant is highly visible to others. If one participant defects and fails to provide liquidity to the system, other participants may adopt a punishment strategy, such as delaying their own payments to that member. This cost associated with being perceived as "free-riding" on liquidity provided by peers in a repeated game may thus induce a cooperative outcome.

One specific mechanism possibly enforcing such discipline is the use of *bilateral net sender limits*. This is the simple liquidity management rule whereby bank A ceases to make payments to bank B if the *net* flux of payments from A to B reaches a certain (positive) *limit*; in other words, B is "punished" if it is seen "not to reciprocate." Anecdotal evidence suggests that this mechanism is indeed applied by some CHAPS members. The appendix formalizes the argument, but the logic behind bilateral net sender limits is that they create "interperiod spillovers," increasing the cost of delaying payments by depriving the recalcitrant bank of liquidity in subsequent periods. As a result, banks are encouraged to make payments promptly, to the benefit of the system.[12] As shown in the exhibit, the effect of such limits is to shift the delay curve to the right. In this framework, the result would be to increase the value of collateral posted at the beginning of the day. The effect on liquidity usage would depend on the effect on liquidity recycling over the course of the day.

But even though centralised throughput guidelines and decentralised mechanisms may serve as coordination devices, the use of bilateral net sender limits (in the form described above) would not enforce coordination on a particular time of the payment day and hence would not necessarily overcome a tendency to delay payments until the end of the day. Acting in tandem, however, throughput guidelines and bilateral coordination mechanisms can be expected to both enhance the efficiency of liquidity recycling and to smooth the intraday distribution of payments. We seek evidence of these effects in the empirical analysis that follows.

3. The Timing of CHAPS Sterling Payments

We now turn to an empirical analysis of the timing and funding of payments in CHAPS Sterling. Based on the discussion above, we would expect the low opportunity cost of posting collateral and the high expected delay costs associated with a subset of payments to limit the degree to which members delay payments in order to take advantage of incoming funds. We

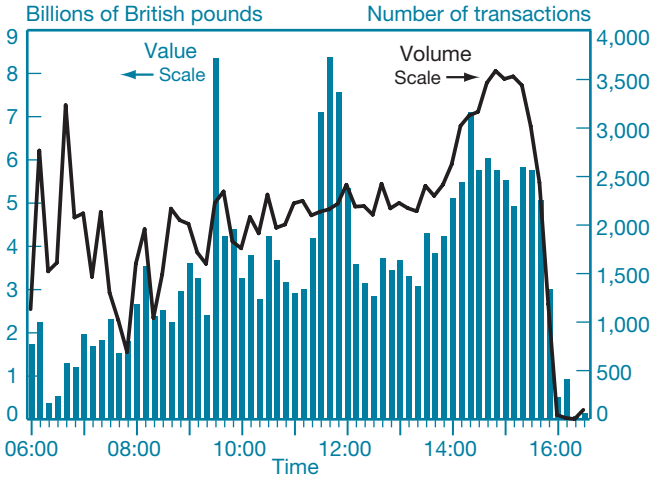
[10] Only a small proportion of these banks use Fedwire heavily, however.

[11] The network of payments between settlement banks in CHAPS, however, is underlain by a more extensive network of payments between the originators of payments and the end recipients. The characteristics of this network are similar to those of Fedwire. See Soramäki et al. (2006) and Becher, Millard, and Soramäki (2007).

[12] Bilateral limits also have the important function of reducing the impact of "liquidity sinks," created when a bank is able to *receive* payments but is unable to *release* funds—for example, as a consequence of a technical outage. By restricting flows to the "sink" bank, bilateral limits reduce the amount of liquidity that is syphoned out of the system.

Chart 1

Value and Volume of Payments of All Banks in CHAPS Sterling System

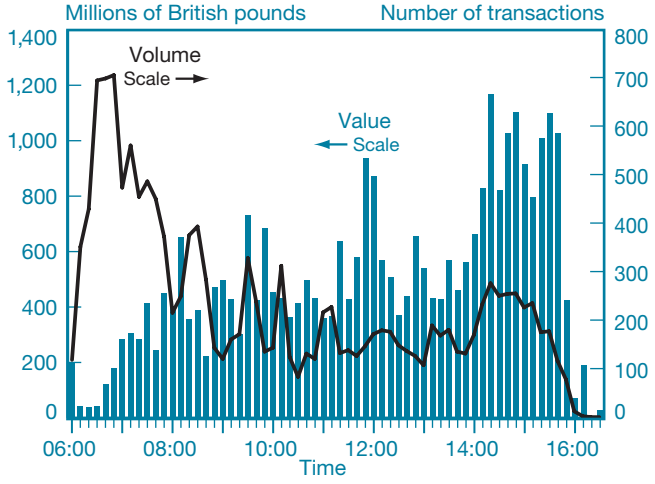


Sources: CHAPS payment database; Bank of England calculations.

Note: We calculate the figures as daily average values and volumes in ten-minute intervals, using data from October 2006.

Chart 2

Value and Volume of Payments of Foreign Banks in CHAPS Sterling System



Sources: CHAPS payment database; Bank of England calculations.

Note: We calculate the figures as daily average values and volumes in ten-minute intervals, using data from October 2006.

would expect to see this reflected in the intraday payment distribution. We also hypothesise that structural features of the CHAPS Sterling system, in particular the imposition of throughput guidelines and the "small club" membership, will promote the recycling of liquidity and smooth the distribution of payments throughout the day.

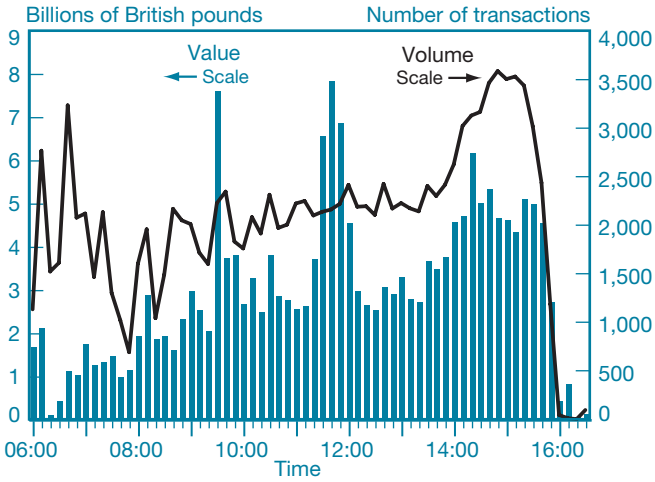
We first consider the intraday pattern of payments in CHAPS Sterling. The system opens for normal service at 6:00 a.m. and closes at 4:20 p.m. CHAPS settlement banks can initiate transfers on behalf of themselves and their clients normally until 4:00 p.m., although settlement members may make transfers on their own behalf, or on behalf of other credit institutions and certain money market participants, for the purpose of settling their end-of-day positions after this time.

The intraday profiles of payments in CHAPS Sterling are shown in Charts 1-3, alongside the profile of Fedwire payments in Chart 4.[13] The profile of payments by all banks (Chart 1) displays three distinct value peaks: the first around 9:30 a.m.; the second between roughly 11:30 a.m. and 12:00 p.m.; and a third, sustained peak between 2:00 p.m. and 4:00 p.m. The value profiles are similar for foreign and domestic banks (Charts 2 and 3, respectively: on average, 55 percent of foreign bank payments are made by noon, compared with 58 percent

[13] The analysis in this section is based on data for CHAPS Sterling payment flows only; other transfers, such as settlement payments for BACS and C&CC, are not included. Our results apply to one month only (October 2006); however, our analysis has been repeated for data from June 2005 and January 2006, with similar results.

Chart 3

Value and Volume of Payments of Domestic Banks in CHAPS Sterling System



Sources: CHAPS payment database; Bank of England calculations.

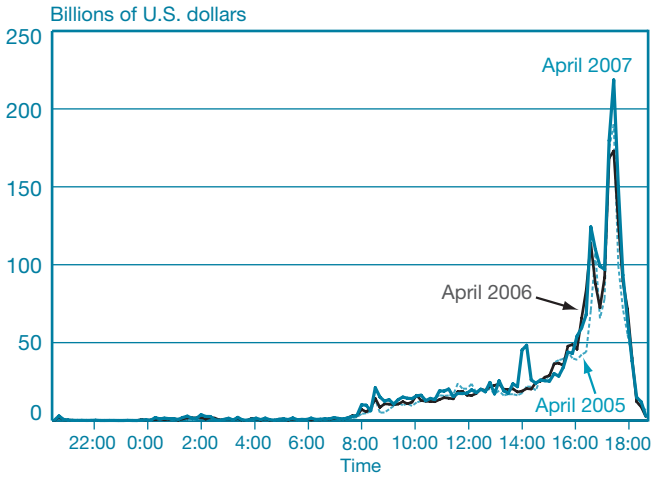
Note: We calculate the figures as daily average values and volumes in ten-minute intervals, using data from October 2006.

for domestic banks. Both domestic and foreign banks make around 25 to 30 percent of payments by value during the end-of-day value peak. The profile contrasts with that of payments in Fedwire, which exhibits strong concentration of payment value at the end of the day (Chart 4).

Chart 4

Value of Payments Made by Time of Day in Fedwire

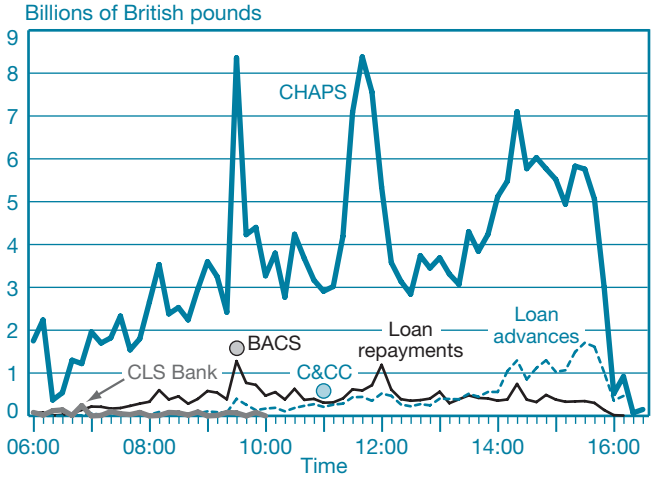
Daily Average by Month



Source: Federal Reserve Bank of New York.

Chart 5

Effect of Payment Events on the Intraday Distribution of CHAPS Sterling System Payments



Sources: CHAPS payment database; Bank of England calculations.

Note: We calculate the figures as daily average values and volumes in ten-minute intervals, using data from October 2006.

Setting aside for the moment the effect of strategic behaviour, we note that the observed *value* peaks correspond well with scheduled payment events, particularly those associated with time-critical payments and throughput deadlines. This result is illustrated in Chart 5.

The first peak (9:30 a.m.) temporally corresponds both with the timetable for pay-ins to CLS Bank, which can be made during a payment window between 7:00 a.m. and 11:00 a.m., and with the settlement of multilateral net positions in the BACS retail payments system.[14] In both cases, the value of the settlement payments involved is small relative to the total value of payments made at these particular times. However, the observed peak may reflect the tendency to delay payments until after the time-critical payments have been made, when any uncertainty around the value of these settlements has been resolved. The settlement payments may also release liquidity for the settlement of subsequent payments. Sharp value peaks also occur ahead of the throughput deadlines, at noon and 2:30 p.m., suggesting that the guidelines do impact significantly on the intraday distribution of payments. In fact, there is prima facie evidence that payments are delayed until the period immediately before the deadlines, which may reflect strategic behaviour.[15]

Finally, the value peaks may also reflect the routine patterns of activity in the overnight interbank market: Late-afternoon value peaks are likely to be reinforced by the creation of overnight loans for the purposes of position-squaring, which must typically be repaid the following morning. Going forward, the effect of overnight markets on payment profiles and liquidity usage is fertile ground for future research.

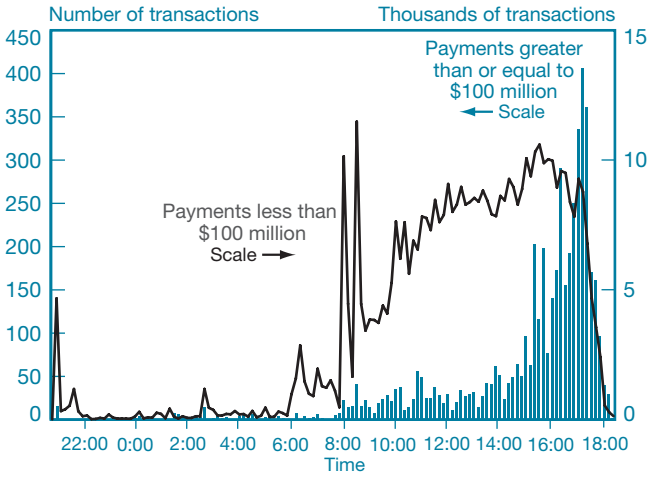
The *volume* profile of CHAPS Sterling payments is relatively smooth throughout the day for the system as a whole. Volume peaks occur shortly after opening and again late in the day. The volume profile is notably different for the set of foreign banks: Volumes are highly concentrated during the first two hours and fall away sharply thereafter. Approximately 40 percent of foreign banks' payments by volume are made by 8:00 a.m., compared with only around 15 percent for domestic banks. This may in part reflect the settlement of payments queued between the opening of continental European markets and the opening of CHAPS Sterling.

The concentration of payment value in Fedwire appears to be driven by a distinct skew in large-value payments toward the end of the day (Chart 6). This distribution may reflect institutionally imposed timings for certain types of payments, such as for CHIPS and DTC settlement; the creation of overnight loans; and settlement payments in financial markets. This is consistent with McAndrews and Rajan's (2000) observation that this peak existed prior to the imposition of overdraft fees. However, as discussed above, the peak may additionally serve as a focal point for, and be reinforced by, proactive payment coordination.

[14] The plots shown do not include these payments.

[15] The noon peak also follows the settlement of positions in the C&CC. This may be influential, although the very low value of settlement payments in this system suggests that it is unlikely to trigger a liquidity cascade or to generate material uncertainty for participants.

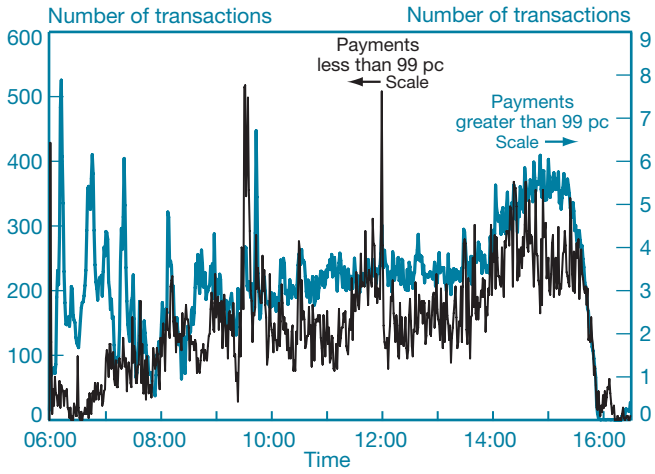
Chart 6
Distribution of Fedwire Payments by Size of Payment



Source: Federal Reserve Bank of New York.

Note: $100 million was the 99th percentile for payment size on March 19, 2007.

Chart 7
Distribution of CHAPS Sterling System Payments by Size of Payment
One-Minute Intervals, Daily Average, October 2006



Sources: CHAPS payment database; Bank of England calculations.

Large-value payments in CHAPS Sterling, defined as those that fall within the 99th percentile of the distribution of payment sizes, account for around 75 percent of daily payment value. As Chart 7 illustrates, the volume of payments per minute in this category is very small, and the intraday distribution is smoother than in Fedwire (Chart 6). Peaks in the incidence of large-value payments occur at 9:30 a.m., 12:00 p.m. (perhaps related to the first throughput guideline), and at the end of the day (consistent with the timing of large position-squaring payments in the interbank market). This suggests that the distribution of large-value payments is driven by the purposes of the payments in question and less by a generalised tendency to delay until the end of the day.

The early concentration of time-critical payments in CHAPS Sterling (in conjunction with other ancillary system settlement payments and the possible need for additional liquidity transfers to CREST) may also help explain why the average value of payments made during the first two hours of opening is low. Participants may be reluctant to commit liquidity to other large-value payments until these time-critical payments have been made. This applies less forcefully to low-value payments; indeed, it is apparent from Chart 7 that low-value payments are released into the system early, perhaps reflecting their low consumption of liquidity.[16]

The contrast between the payment profiles in CHAPS Sterling and in Fedwire—in particular, the observation that payments are much less concentrated at the end of the day in CHAPS Sterling—provides some initial evidence that the incentives for payment delay are weaker in CHAPS Sterling than in Fedwire.[17] The profile of CHAPS Sterling payments is clearly influenced by the existence of time-critical payments and throughput guidelines. To consider whether these patterns are also influenced by the strategic behaviour of members, we now attempt to disaggregate the sources of funding. In particular, we assess whether there is evidence that the use of incoming funds varies by time of day and, following McAndrews and Rajan (2000), whether peaks in liquidity recycling coincide with peaks in payment activity.

4. The Funding of CHAPS Sterling Payments

4.1 Methodology

To decompose the sources of funding of CHAPS Sterling payments, we distinguish between two sources of funding: 1) payments received from other CHAPS Sterling participants within a specified time interval and 2) account balances held at the Bank of England, funded both by collateral posting and the maintenance of positive reserve account balances.

[16] The tendency to settle a high volume of low-value payments early in the day may also reflect the relative complexity of settling high volumes of low-value payments in the event of an operational disruption later in the day. See also footnote 9.

[17] Of course, it is not possible to observe delay directly from the intraday payment profile. This would require knowledge of the timing of payment instructions, as well as of settlement.

Box 4
Measurement of "Offsetting" Payments in CHAPS Sterling

We define the value of payments made from member i to member j within minute t as p_{ij}^{t}.

The total value of payments made within that minute is therefore $\sum_{i,j} p_{ij}^{t}$ and the value of net payments for each member, i, is $\sum_{j}(p_{ij}^{t} - p_{ji}^{t}) = N$.

The value of payments *not* offset within a minute is equal to the sum of net payments for the set of members for which N is positive, or, equivalently, $\frac{1}{2}\sum_{i}|N|$.

The *value of offsetting payments* is then calculated as the value of gross payments made, less the value of payments not offset: $\sum_{i,j} p_{ij}^{t} - \frac{1}{2}\sum_{i}|N|$.

In their study of payment activity in Fedwire, McAndrews and Rajan (2000) consider incoming payments as a funding source for outgoing payments only if those incoming payments offset outgoing payments *within the same one-minute interval*. This definition follows naturally from the overdraft charging structure, as fees are based on the outstanding overdraft at the end of each minute. Provided that all payments are offset by incoming payments within that same minute, irrespective of the ordering of the payments, no charge is incurred.

We choose to adopt the same methodology for the measurement of incoming payments as a funding source in CHAPS Sterling (Box 4). While recognising that payments cannot be "offset" within a minute in the same way as in Fedwire (if outgoing payments are made first, intraday liquidity will be required even if it is subsequently replenished by incoming funds), this measure is useful both as a point of comparison with Fedwire and as an indicator of liquidity recycling in CHAPS Sterling.[18] It should be noted, however, that this particular measure does not capture the recycling of funds hoarded from previous time periods, which is a significant omission. We return to this point in a subsequent discussion.

[18] It is also questionable whether active offsetting of payments within a minute is a realistic representation of members' liquidity management processes. However, as discussed in Section 2.5, the use of bilateral limits may result in such behaviour being observed, since outgoing payments may be released immediately when incoming payments create headroom under a bilateral limit.

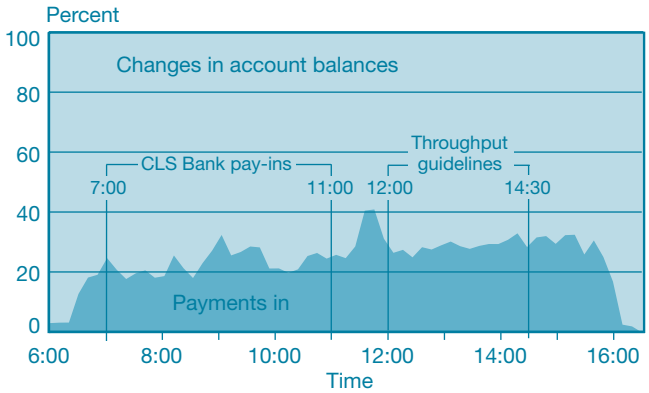
4.2 Results

We now decompose gross payments into the constituent funding sources (Chart 8). While the *absolute* value of payments "offset" by incoming payments increases when the value of gross payments increases, the *proportion* of payments funded by incoming payments remains comparatively stable throughout the day. On average, during the day, around 23 percent of payments made are funded using incoming payments; the use of incoming payments peaks at around 42 percent shortly before noon.

Compared with the results for Fedwire (Chart 9), CHAPS Sterling does not exhibit a pronounced peak in the share of incoming payments as a funding source during the end-of-day value peak, although the proportion of payments funded by incoming funds is above the daily average at this time. It is, however, notable that a distinct peak occurs shortly before the first throughput deadline (at 12:00 p.m.), suggesting that incoming payments are a particularly important funding source at this time. This may ease the liquidity demands of meeting the throughput deadline; indeed, this concentration may be a product of deliberate payment coordination on the focal point(s) provided by throughput guidelines.

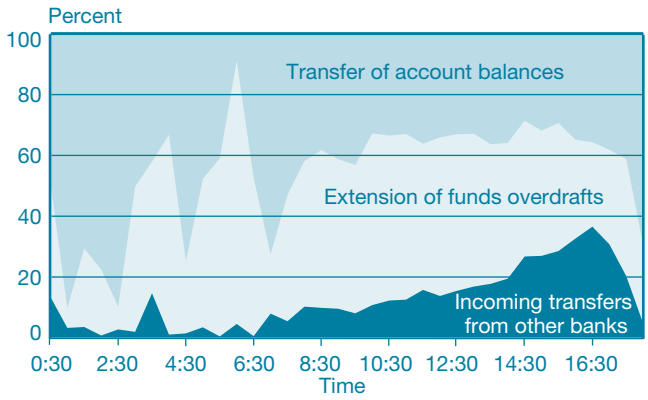
Perhaps the most striking observation from Charts 8 and 9, however, is that on this measure, the *average* level of liquidity recycling within each minute is considerably higher in CHAPS Sterling than in Fedwire. In other words, the *level* of liquidity recycling, whether as a result of active coordination or otherwise, appears to be greater throughout the day in CHAPS Sterling. This result is consistent with our hypothesis that the

Chart 8
Shares of Funding Sources of CHAPS Sterling System Payments
October 2006



Sources: CHAPS payment database; Bank of England calculations.

Chart 9
Shares of Funding Sources of Fedwire Funds Transfers
Average of Four Days over Half-Hour Intervals



Source: Federal Reserve Bank of New York.

Note: Because few payments are made between 12:30 a.m. and 8:30 a.m., the variation in the shares of funding sources during that period of the day is driven by a small number of payments.

more concentrated structure of CHAPS Sterling is likely to be more conducive to liquidity recycling throughout the day, both as a natural consequence of there being fewer participants and as a result of bilateral coordination resulting from "small club" behaviour. Such behaviour, in combination with the distribution of time-critical payments and the effect of throughput guidelines, may ensure that liquidity continues to flow smoothly through the system.

4.3 Liquidity Recycling and Liquidity Constraints

We noted earlier that the opportunity cost of posting collateral (or holding positive reserve account balances) may not be uniform across all CHAPS Sterling participants. In particular, the cost of collateral, and hence of liquidity, should be higher for foreign banks, since they are not subject to the Stock Liquidity Regime. If this is the case, foreign banks would have a stronger incentive to fund payments using incoming funds, and they would be seen to attain higher recycling ratios. Is this indeed the case?

To answer that question, we consider the relationship between the maximum proportion of liquidity drawn down and the overall level of liquidity recycling achieved by each member during the day, measured as the ratio of the value of total daily gross payments to the maximum value of liquidity

Recycling and Liquidity Usage
Daily Average, October 2006

Recycling Ratio (r)	Number of Banks	Liquidity Used (Minimum-Maximum Range, in Percent)
$0 < r \leq 5$	2	68.0 - 85.1
$0 < r \leq 10$	5	37.0 - 99.7
$r > 10$	5	12.5 - 58.5

Sources: CHAPS payment database; Bank of England calculations.

Notes: We exclude the Bank of England and CLS Bank. The Royal Bank of Scotland and Natwest are treated as a single entity (the Royal Bank of Scotland Group), although they retain separate settlement accounts.

used (*the recycling ratio*). This measure is not subject to the critique of the previous section, since it does capture the benefits of liquidity hoarded over multiple periods.

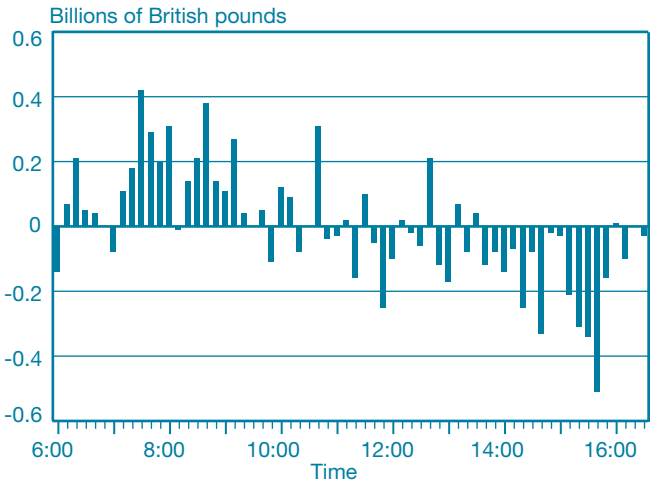
There is a wide variation in the extent of liquidity recycling achieved by CHAPS Sterling members (see table). Only five banks—all domestic—achieve recycling ratios greater than 10, and in each of these cases the proportion of liquidity drawn down is relatively low.[19] The foreign banks achieve lower recycling ratios and draw down a correspondingly large proportion of available liquidity. This implies that foreign banks may indeed face greater liquidity constraints than domestic ones, as suggested earlier in our discussion of the cost of collateral. As noted, we would expect the incentive to delay payments in order to take advantage of liquidity recycling to be correspondingly high, but this expectation is not supported by the data. How might this be explained?

From Section 3, we know that the intraday value profiles of *outgoing* payments made by domestic and foreign banks are similar. However, the ability to recycle liquidity and thereby lower aggregate liquidity requirements during the day also depends on the distribution of *incoming* payment flows. Charts 10 and 11 clearly illustrate that the pattern of net payments is very different on aggregate for domestic and foreign banks, even though all members must comply with the throughput guidelines.

Domestic banks are, on aggregate, net recipients of funds in the morning and net suppliers of funds in the afternoon. Foreign banks exhibit the opposite trend: Net payments are negative until late morning and become positive thereafter. This implies that domestic banks tend to accumulate funds during the morning and then pay these funds out in the afternoon, thereby reducing intraday liquidity usage and increasing the recycling ratio. When the flows are reversed,

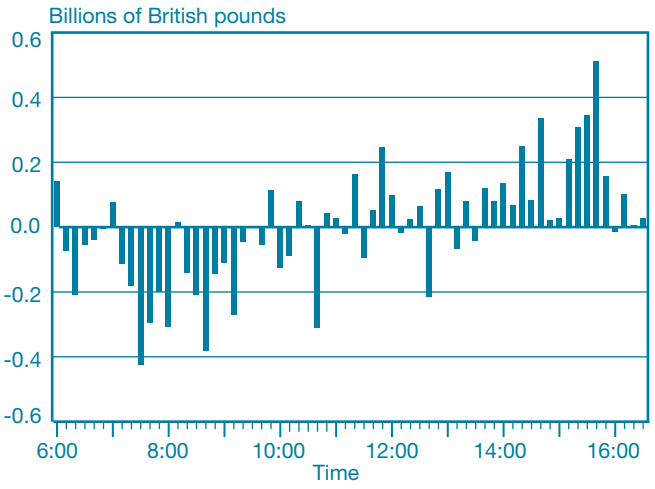
[19] Note that foreign and domestic banks are not differentiated in the table.

Chart 10
Net Payments for Domestic Banks in CHAPS Sterling System
Daily Average, October 2006



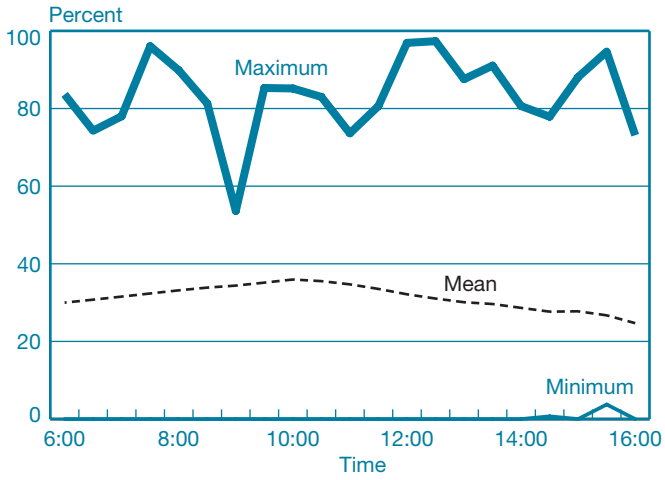
Sources: CHAPS payment database; Bank of England calculations.

Chart 11
Net Payments for Foreign Banks in CHAPS Sterling System
Daily Average, October 2006



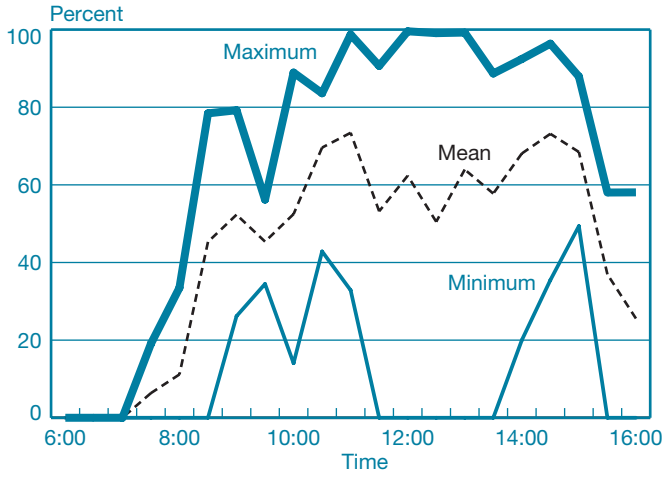
Sources: CHAPS payment database; Bank of England calculations.

Chart 12
Liquidity Usage by Domestic Banks
Daily Average, October 2006



Sources: CHAPS payment database; Bank of England calculations.

Chart 13
Liquidity Usage by Foreign Banks
Daily Average, October 2006



Sources: CHAPS payment database; Bank of England calculations.

as is the case for foreign banks, liquidity must be drawn down to fund the net outgoing payments; hence, one would expect recycling ratios to be lower. However, while the distinction between the payment flows of the domestic and foreign banks at the aggregate level is striking, these results conceal considerable variation within both sets of banks. This is apparent from the intraday patterns of liquidity usage, illustrated in Charts 12 and 13.

While it is clear that the maximum proportion of liquidity used by foreign banks is, on average, higher for much of the day than it is for domestic banks, there is considerable variation within both sets of banks. Some domestic banks are net payers for much of the morning and use a large proportion of their available liquidity early in the day. For these banks, the recycling ratio is relatively low. Other domestic banks, by contrast, use little or no liquidity for much of the day. These

banks are more able to take advantage of funds received in the morning and achieve the highest recycling ratios of all CHAPS Sterling members. By contrast, all foreign banks in CHAPS Sterling are net payers during the morning and approach their maximum liquidity usage early in the day.

So while liquidity recycling appears to be relatively efficient in CHAPS Sterling, the extent to which individual banks benefit from recycling varies considerably. For those banks that are net payers early in the day—including all of the foreign banks and some large domestic banks—recycling ratios are much lower. One could argue that this results from the low cost of liquidity, since the incentive to structure payments so as to reduce liquidity costs may be weak. But many of the banks with relatively low recycling ratios do appear to face liquidity constraints, since they also use a large proportion of liquidity posted. This suggests that *some* banks may be unable to recycle liquidity to the extent that they would wish, which may reflect the simple observation that coordination will result only if all (or a sufficient number) of the banks are similarly incentivised by liquidity pressures to cooperate.

While liquidity recycling appears to be relatively efficient in CHAPS Sterling, the extent to which individual banks benefit from recycling varies considerably.

The variation in recycling ratios also reflects the effect of the other influences on payment timing. The observed patterns of net payments in Charts 10 and 11 are likely to reflect structural differences in the underlying businesses of the participants and their customers, resulting in differences in the distribution of payment instructions and deadlines. If, for example, certain participants (or their customers) routinely borrow in the overnight market while others lend, the payment flows of the two groups will be correspondingly different. To the extent that these structural factors limit participants' discretion over the timing of payments, this may explain the observed variation in the distribution of recycling benefits.[20]

5. Discussion and Conclusions

Our analysis indicates that even though the intraday liquidity regime supporting CHAPS Sterling payments does not give rise to the same incentives for minute-by-minute payment coordination as those in Fedwire, the observed degree of liquidity recycling appears to be high. We have also seen that the intraday profile of payments is comparatively smooth. Taken together, these observations reveal that even if collateral posting is perceived to be costly by some banks—and hence a "liquidity incentive to delay" does exist in CHAPS Sterling—other features of the system help avoid a prisoner's dilemma equilibrium in which the majority of payments are delayed until late in the day. This serves to reduce the maximum liquidity required to make a given set of payments and hence the aggregate value of collateral that needs to be posted. The empirical evidence suggests that payment coordination may also play an important role in Fedwire, although in this case coordination—and consequently liquidity recycling—is strongly concentrated around an end-of-day focal point.

Which features of the system support this high and constant level of liquidity recycling? Centralised coordination devices are likely to play a role; in particular, throughput guidelines counteract any generalised tendency to delay payments until the end of the day. Indeed, the spike in the proportion of payments "offset" before noon is evidence that the incoming funds become an increasingly significant funding source at this time of day, reducing the liquidity cost of complying with the deadline.

Other forms of "decentralised coordination" between members may also be significant. The high visibility of payment flows in the concentrated CHAPS system allows members to monitor their bilateral positions and to take action if counterparts fail to make payments in a timely fashion. The prisoner's dilemma may then simply be resolved through the repeated interaction of the small number of participants, whereby recalcitrant participants are "punished" for failing to provide liquidity to the system. It is arguable that these pressures may be less strong in the more diffuse Fedwire system. While not explicitly revealed by the aggregate data, there is also anecdotal evidence that participants apply bilateral net sender limits with respect to other system participants, thereby promoting the recycling of liquidity between each bilateral pair and enhancing the liquidity efficiency of the system. An empirical question remains as to how often these limits "bite" in practice, but a "small club" like CHAPS is a natural environment for the application of such devices, which help generate a smooth payment profile.

The patterns in the timing and funding of CHAPS Sterling payments described in this article would appear to be risk-beneficial, for individual participants and for the system as a whole. The low opportunity cost of posting collateral and the tendency for all members (domestic and foreign) to post collateral at the beginning of the day help ensure that time-critical payments do not fail for want of liquidity. Moreover, the apparently low opportunity cost of collateral results in many banks providing a liquidity cushion in excess of that

[20] The patterns of funding flows in the overnight market are the subject of ongoing research at the Bank of England.

required to make time-critical payments. This not only increases the resilience of that participant to liquidity shocks, but also contributes to the resilience of the system as a whole.

The efficient recycling of liquidity during the day further contributes to a reduction in liquidity risk by reducing the aggregate liquidity required to make a given set of payments. In theory, this is particularly beneficial for those banks that face a relatively high cost of liquidity. However, we have seen that many of the members that draw down a large proportion of available liquidity are unable to take advantage of incoming funds to the same extent as other members, perhaps owing to the distribution of underlying payment orders from customers. For these banks, the likelihood of needing to post additional collateral during the day may be correspondingly high.

Our analysis can take several interesting directions. In this article, we have formulated a number of hypotheses on the determinants of behaviour in a payments system, suggesting some implications for the efficiency of the system itself. However, we have not been able to disentangle fully the effects of the factors identified; this is perhaps an inevitable drawback given our descriptive approach. Formal analysis, supported and complemented by further econometric work on payment data, may help in this direction by shedding additional light on key issues such as the effect of membership size on payment behaviour and the precise way in which banks achieve coordination (using, for example, bilateral net sender limits). Our analysis of the variations in liquidity recycling intensity also makes a strong case for further analysis of the overnight market, particularly its effect on payment behaviour and liquidity usage. Ultimately, such analysis will contribute to an understanding of how a large-value payments system functions and suggest where and how risks to the system may crystallise.

Appendix: Effects of a Bilateral Net Sender Limit on a Bank's Behaviour

We present a simple, formal illustration of a single bank's problem when choosing an optimal level of liquidity. A bilateral net sender limit is shown to incentivise early liquidity provision and early payment, thereby generating high levels of liquidity recycling. No attempt is made here to develop a fully fledged game-theoretical model of payments. That model would be more complex, because incoming payments would need to be modeled as a strategic choice (by the other banks) instead of as an exogenous random variable. Considering a single bank's decisions in isolation allows us to focus on the marginal effect of a bilateral net sender limit on the incentives to post liquidity and to delay payments. We would nevertheless expect to find this effect in a more complex setting.

Setting

Suppose that our bank receives payment orders exogenously from its customers. To execute these orders, the bank requires liquidity, which can be obtained either by posting collateral or by waiting for exogenous (and random) incoming payments. We assume that our bank faces the following sequence of events, all of which occur within a fixed time interval t (which can be thought of as a metaphor for a trading day, or part of the day such as "the morning"):

$$\begin{array}{ccccccc} t.0 & \_\_\_\_\_ & t.1 & \_\_\_\_\_ & t.2 & \_\_\_\_\_ & t.3 \\ x_t & & w_t & & y_t & & p_t,\ \delta(Q_t)\,. \end{array}$$

- At $t.0$, the bank receives payment orders to the value x_t.
- At $t.1$, the bank decides how much liquidity to raise, w_t, at a cost $\lambda(w_t)$.
- At $t.2$, incoming payments provide the bank with additional liquidity y_t, so the bank has total liquidity of $l_t = w_t + y_t$.
- At $t.3$, the bank makes payments p_t. If $l_t < x_t$, the bank can only pay up to $l_t (p_t = l_t)$, so it "queues" an amount of payments $Q_t = x_t - l_t$. If instead $l_t > x_t$, then $p_t = x_t$ and the bank has spare liquidity. To simplify, we assume that the cost of a backlog Q_t is a function $\delta(Q_t)$, with $\delta(Q_t) > 0$ if $Q > 0$, and $\delta(Q_t) < 0$ otherwise.[21] To simplify further, we assume that if $Q_t < 0$, the bank sells the spare liquidity in the market, immediately realising $\delta(Q_t)$; if instead $Q_t > 0$, then $\delta(Q_t)$ includes all costs derived from delaying payments, in particular, the cost of the extra liquidity with which the bank settles or cancels the queued payments.
- The bank then begins the next period $t + 1$ afresh, with no liquidity and no queues (all costs / benefits stemming from Q_t are accounted for by $\delta(Q_t)$).

We now look at the bank's incentives and its optimal choice. We want to show how a bilateral net sender limit incentivises short queues and thus liquidity recycling. To do so, we compare two cases, one in which there is no bilateral net sender limit and one in which a limit is implemented.

The Bank's Problem: I

No Bilateral Sender Limits

Suppose that incoming payments y_t arrive according to some exogenous distribution $f(.)$, which is independent of t and of the bank's choices. In this case, the bank's problem is actually a single-period maximisation,[22] so we are able to eliminate all time indices. By borrowing w, the bank reduces the expected queue Q, thus abating the expected delay costs $\delta(Q)$ (and possibly transforming them into a gain, if $Q < 0$). However, liquidity is costly, and the bank may hope to make use of receipts from the other bank (via y). In general, the bank will not raise a full $w = x$, and it will rely partly on incoming payments.

[21] A negative queue is a positive amount of liquidity whose positive value is a negative cost. Note that, to rule out the existence of "money-making machines," δ and λ (the cost of liquidity) are mutually restrained. See the example below.

[22] There are no spillover effects between t and $t+1$ because we assume that the bank realises $\delta(Q_t)$ from any queue or spare liquidity, beginning period $t+1$ afresh.

To find the optimal w, the bank looks at the total cost from raising w:

$$C(w) = \lambda(w) + E\delta(Q)$$
$$= \lambda(w) + \int_0^r f(y)\delta(x - w - y)dy,$$

where r is an upper bound on the incoming payments that the bank can expect to receive. Provided that some technical convexity conditions hold, the optimal amount of w^* is determined by solving the standard condition $C'(w) = 0$.

Bilateral Net Sender Limit

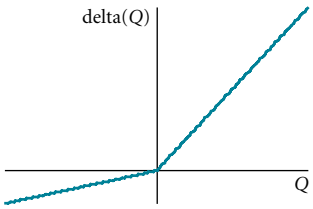
When a bilateral net sender limit is in place, the distribution of incoming payments y is no longer exogenous, but depends on previous payments. As a trivial example, if too few payments are sent out, the bilateral limit is hit and no further incoming y can be expected. More generally, the effect of choosing a particular w_{t-1}, and hence a volume of payments p_{t-1}, spills over to the next period and influences the expected amount of incoming funds y_t. As a consequence, compared with the case in which no bilateral net sender limits exist, every unit of liquidity made at $t-1$ now carries an extra benefit in terms of a liquidity saving at time t (although, of course, only in expected terms and only if it actually allows additional payments to be made at $t-1$). This interperiod spillover tilts the balance in favour of posting more w_t, increasing payments and reducing the queue. If, in addition, the liquidity cost λ increases in time, this mechanism is reinforced: The liquidity saved at $t+1$ by posting more w_t at t is even more valuable. Similarly, if the bilateral limit itself depends on i's past payments (for example, a "bad" payment record may induce the other banks to tighten prudentially their bilateral limit toward i), the incentives to pay early will be even stronger.

The Bank's Problem: II (Example)

We now solve analytically the bank's problem, under particular assumptions about the cost functions and the distribution of y. Suppose the delay cost is

$$\delta(Q) = \begin{cases} CQ \text{ if } Q \geq 0 \\ cQ \text{ if } Q < 0 \end{cases},$$

with $0 < c < C$. The corresponding graph is therefore:



In this case, the costs of a positive queue Q grow faster than the gains from spare liquidity (which are simply the negative costs from a negative queue).

We also assume that liquidity costs are linear: $\lambda(w) = \lambda w$ with $\lambda > 0$. To make the problem interesting, we first impose $\lambda < C$, implying that it is better to post liquidity and make a payment than not to post it and fail to make the payment, and then impose $c < \lambda$, which ensures that it is not optimal to post infinite liquidity. Indeed, if it were $c > \lambda$, then any pound of liquidity would be worth more to the bank than the cost λ, independently of whether it is used to shorten a queue (which gives a benefit $C > \lambda$) or if it results in spare liquidity (whose benefit is precisely c, which must therefore be assumed to be smaller than λ).

No Bilateral Net Sender Limit

Suppose y_t is uniformly distributed in $[0,\ r]$, so its probability density function is $f(y) = \frac{1}{r}$. Then, the bank's cost function is:

(1)
$$C(w) = \lambda(w) + \int_0^r f(y)\,\delta(x-w-y)\,dy =$$
$$= \begin{cases} \lambda w + \frac{1}{r}\left[C\int_0^{x-w}(x-w-y)dy + c\int_{x-w}^{r}(x-w-y)\,dy\right] & \text{if } x-w \le r \\ \lambda w + \frac{1}{r}C\int_0^r (x-w-y)\,dy & \text{if } r < x-w \end{cases}$$
$$= \begin{cases} \lambda w + \frac{1}{r}\left[C\frac{1}{2}(w-x)^2 - c\frac{1}{2}(r+w-x)^2\right] & \text{if } x-w \le r \\ w(\lambda - C) - \frac{1}{2}C(r-2x) & \text{if } r < x-w \end{cases}.$$

We now find the optimal w, to be called w^*.

- Suppose w^* is such that $x - w^* \le r$. In this case, the optimality condition would be

(Opt)
$$\frac{d}{dw}\left(\lambda w + \frac{1}{r}\left[C\frac{1}{2}(w-x)^2 - c\frac{1}{2}(r+w-x)^2\right]\right)$$
$$= \frac{1}{r}(r\lambda + Cw - Cx - cr - cw + cx) = 0$$

yielding an optimal liquidity $w^* = x + \frac{r(c-\lambda)}{C-c}$ and thus a total cost equal to

$$\frac{1}{2(C-c)}(-r\lambda^2 + 2Cx\lambda + 2cr\lambda - 2cx\lambda - Ccr) = C^*.$$

- Suppose instead $r < x - w^*$. Because $\lambda - C < 0$, the total cost decreases in w as long as $r < x - w$ (see equation 1). Hence, we would have a corner solution at $w^* = x - r$, which yields a cost $x\lambda - r\lambda + \frac{1}{2}Cr = C^{**}$.

Now, the difference $C^{**} - C^* = \frac{1}{2}\frac{r}{C-c}(C-\lambda)^2$ is always positive. Hence, the cost-minimizing w is the one found in the first case, supposing $x - w^* \le r$:

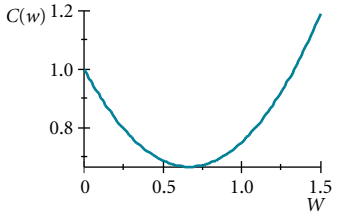
(2)
$$w^* = x + \frac{r(c-\lambda)}{C-c} < x.$$

It should be noted that w^* has the anticipated properties: It increases in x (the amount of payments to make) and in C (the cost of queues), and it falls with $\lambda - c$, that is, with the difference between the cost of liquidity and its benefits as end-of-day spare liquidity.

Finally, substitution of equation 2 into equation 1 yields the optimal (minimised) cost:

(3)
$$C^* = \frac{1}{2(C-c)}(-r\lambda^2 + 2Cx\lambda + 2cr\lambda - 2cx\lambda - Ccr).$$

If we set, for example, $c = \frac{1}{2}$, $\lambda = r = x = 1$, and $C = 2$, the graph of $C(w)$ is:



In this case, the liquidity posted is 66 percent of the payments due ($x = 1$).

Bilateral Net Sender Limit (BNSL)

Suppose again that y_t is uniformly distributed. This time, however, imagine that incoming payments are drawn from $[0, r_t]$, with r_t determined by a bilateral net sender limit:

$$r_t = r_{t-1} + (p_{t-1} - y_{t-1}).$$

In this case, an increase in p_{t-1} due to higher liquidity w_{t-1} pushes up r_t, which in turn affects the minimised costs

(see equation 3). What is the value of such a spillover, from $t-1$ liquidity into time t costs? The answer is

$$\frac{dC_t^*}{dw_{t-1}} = E_{t-1}\left[\frac{dC_t^*}{dr_t}\frac{dr_t}{dw_{t-1}}\right],$$

which we calculate term by term.

The term $\frac{dC_t^*}{dr_t}$ immediately derives from equation 3, which reveals that the expectation term is irrelevant here (λ , c, and C are constant over time):

$$(4) \qquad \frac{dC_t^*}{dr_t} = -\frac{1}{2(C-c)}(\lambda^2 + Cc - 2c\lambda) = Z < 0.$$

To prove the last inequality, recall that $c < \lambda < C$. This implies both $\lambda c < \lambda^2$ and $\lambda c < Cc$, which, summed member by member, gives $2\lambda c < \lambda^2 + Cc$.

Expectations do matter on the second term $\frac{dr_t}{dw_{t-1}}$. In fact, an extra pound worth of w_{t-1} translates into one more payment only if the available liquidity $w_{t-1} + y_{t-1}$ turns out to be less than the payment orders x_{t-1}, with y_{t-1} unknown at $t-1$. Formally,

$$\frac{dr_t}{dw_{t-1}} = \begin{cases} 1 & \text{if } w_{t-1} + y_{t-1} < x_{t-1} \\ 0 & \text{otherwise} \end{cases}.$$

Hence, the expectation is calculated as:

$$E_{t-1}\left[\frac{dr_t}{dw_{t-1}}\right] = \int_0^{x_{t-1}-w_{t-1}} 1 f(s)ds = \frac{1}{r_{t-1}}(x_{t-1} - w_{t-1}) = W > 0,$$

where the last inequality comes from equation 2. Combining the equation above and equation 4, we finally have

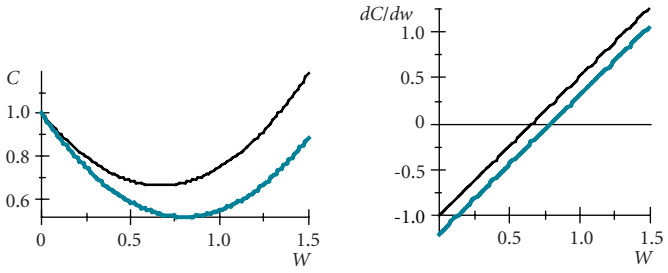
$$\frac{dC_t^*}{dw_{t-1}} = ZW < 0.$$

As anticipated, the spillover is negative, corresponding to an abatement of costs. When the bank internalises these spillovers, these gains (perhaps discounted by a factor β) are added to the optimality condition (Opt), which therefore becomes

$$\frac{dC_{t-1}}{dw_{t-1}} + \beta\frac{dC_t^*}{dw_{t-1}} = \frac{1}{r}(r\lambda + Cw - Cx - cr - cw + cx) - ZW = 0.$$

Hence, the BNSL shifts down the marginal cost schedule. This clearly brings about a higher level of liquidity posting in $t-1$ (in the example below, from 66 percent to 80 percent).

Total and marginal cost of w, with (blue) and without (black) a BNSL:



References

Armantier, O., J. Arnold, and J. McAndrews. 2008. "Changes in the Timing Distribution of Fedwire Funds Transfers." Federal Reserve Bank of New York Economic Policy Review 14, no. 2 (September): 83-112.

Bech, M. L., and R. Garratt. 2003. "The Intraday Liquidity Management Game." Journal of Economic Theory 109, no. 2 (April): 198-219.

Becher, C., S. Millard, and K. Soramäki. 2007. "The Network Topology of CHAPS Sterling Payments." Unpublished paper.

Beyeler, W., R. Glass, M. Bech, and K. Soramäki. 2006. "Congestion and Cascades in Payment Systems." Federal Reserve Bank of New York Staff Reports, no. 259, September.

Buckle, S., and E. Campbell. 2003. "Settlement Bank Behaviour and Throughput Rules in an RTGS Payment System with Collateralised Intraday Credit." Bank of England Working Paper no. 209.

James, K., and M. Willison. 2004. "Collateral Posting Decisions in CHAPS Sterling." Bank of England Financial Stability Review 17, December: 99-104.

Kobayakawa, S. 1997. "The Comparative Analysis of Settlement Systems." Centre for Economic Policy Research Discussion Paper Series, no. 1667, July.

Martin, A., and J. McAndrews. 2008. "Liquidity-Saving Mechanisms." Journal of Monetary Economics 55, no. 3 (April): 554-67.

McAndrews, J., and S. Rajan. 2000. "The Timing and Funding of Fedwire Funds Transfers." Federal Reserve Bank of New York Economic Policy Review 6, no. 2 (July): 17-32.

Mills, D. C., and T. D. Nesmith. 2008. "Risk and Concentration in Payment and Securities Settlement Systems." Journal of Monetary Economics 55, no. 3 (April): 542-53.

Soramäki, K., M. L. Bech, J. Arnold, R. J. Glass, and W. E. Beyeler. 2006. "The Topology of Interbank Payment Flows." Federal Reserve Bank of New York Staff Reports, no. 243, March.