Payment prioritisation and liquidity risk in collateralised interbank payment systems

Robert De Caux[a,*], Markus Brede[b], Frank McGroarty[c,**]

[a] *Researcher, Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ*
[b] *Senior Lecturer, Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ*
[c] *Professor of Computational Finance and Investment Analytics, Southampton Business School, University of Southampton. Southampton SO17 1BJ*

Abstract

Participants in interbank payment systems manage a stream of payment requests of varying priority to minimise their total costs. However, individually optimal strategies may conflict with system-wide optimality and can lead to inefficient equilibria, where banks cannot meet obligations in a timely manner. We construct a model of a collateralised payment system and demonstrate that socially optimal states exist in which banks should delay a proportion of non-priority payments in an internal queue, but banks' strategising behaviour leads to liquidity hoarding and increased systemic cost. We discuss how this behaviour can be reduced using measures available to a regulator.

Keywords: payment systems, agent based, game theory, simulation, policy

* Corresponding author
** Principal corresponding author
Email addresses: rdc1g11@soton.ac.uk (Robert De Caux), Markus.Brede@soton.ac.uk (Markus Brede), F.J.McGroarty@soton.ac.uk (Frank McGroarty)

Acknowledgements

This work was supported by an EPSRC Doctoral Training Centre grant [EP/G03690X/1]

1. Introduction

Providing payment services to allow banking institutions to settle their obligations is one of the key functions of the financial system, and hence it is vitally important that the interbank payment systems are well designed and regulated. The importance of this point becomes even more apparent when considering the exceptionally large payment volumes that have to be processed on a daily basis. The US Fedwire system settles around $2.4 trillion of transactions every day (Federal Reserve Board, 2014), while the UK interbank systems, CHAPS[1] and CREST[2], settle around £575 billion in the same period, which is roughly equivalent to UK annual GDP every three days (Dent and Dison, 2012).

One of the reasons for these incredibly high volumes has been the move to gross settlement. In the past, most payment systems operated on a net settlement basis, but this entailed banks running tremendous counterparty exposures throughout the day. Furthermore, the potential cost of late settlement failure was shown by many studies to lead to heavy systemic risks (Humphrey, 1986; Van den Bergh, 1994). In order to alleviate this problem, many systems now operate a mechanism of Real Time Gross Settlement (**"RTGS"**), ensuring that all obligations are settled with finality, in real-time, via a transfer of funds from the account of the creditor to the account

[1] Clearing House Automated Payment System.
[2] CREST is a securities settlement system.

of the debtor at the central bank.

The increased liquidity demand posed by gross settlement means that intraday liquidity is the lifeblood of these RTGS systems. Ideally this liquidity would be provided free of charge on an intraday basis by the central bank operating as settlement agent for the system, but this would require them to take on an unacceptable level of credit risk. Therefore the central bank must implement a pricing policy to both mitigate this risk but also allow the smooth functioning of the system (Furfine and Stehm, 1998; Freixas et al., 2000). The first policy option is to charge an overdraft fee for any period of negative balance on a participant bank account, as used in the US Fedwire system (Coleman, 2002). The second approach is to demand high-quality collateral on an intraday basis up to the value of liquidity required by the participant. Both CHAPS and the European TARGET2[3] system utilise the latter approach.

In a collateral-based system, the central bank typically provides intraday liquidity to the participant banks at the start of the day. Banks will source liquidity dependent on their projected payment flows, with each balancing a trade-off between the opportunity cost of using that collateral elsewhere and the costs that it will incur due to expected payment delays throughout the day. In a low opportunity cost environment these systems operate efficiently due to a high volume of liquidity being present, but there is a systemic risk if

[3] Trans-European Automated Real-Time Gross Express Transfer 2.

those costs increase. The system incentivises member banks to minimise their own total costs, so that individually optimal strategies, such as free-riding on the liquidity provision of others, may be at odds with the most beneficial behaviour for overall system performance (Afonso and Shin, 2011). Notably, insufficient liquidity sourcing or poor recycling of liquidity within the system can cause cascades of payment failures, leading to delays and system-wide inefficiencies (Angelini, 1998).

For these reasons, intraday liquidity risk is currently receiving a high level of scrutiny. The Basel Committee on Banking Supervision (2012) recently released a paper, detailing a number of measures which banks must report that assess their intraday liquidity requirements under various stress scenarios. However these measures focus on the individual banks themselves and do not fully capture the system-wide effects that a stress situation would cause. An analysis of the equilibrium behaviour of collateralised payment systems under differing conditions is a first step towards understanding these systemic risks. In this paper, we focus on providing such an analysis for CHAPS, but the principal results are applicable to any collateral-based system.

The CHAPS system was until recently very effective in terms of liquidity provision. It had settled into a comfortable equilibrium where banks posted more collateral than they needed to at the start of the day, and all payments were made smoothly with minimal delays. It was also simple for banks to make their decisions on a day-to-day basis as their collateral postings tended not to change, so banks were fully informed as to how others would

act. However, these low liquidity sourcing costs were due to the practice of double duty, whereby collateral that banks were forced to hold as part of their prudential asset buffer could still be used on an intraday basis for posting in a collateral-based RTGS system (Ball et al., 2011). This effectively meant that many banks incurred costs significantly lower than the true opportunity cost of the collateral[4] (James and Willison, 2004). Under the new regulations, banks are forced to hold an additional intraday liquidity buffer (Ball et al., 2011). This implies a significantly larger opportunity cost to banks and hence the incentives for actors in the CHAPS *"liquidity game"* have changed.

Modelling payment flows in these systems is far from trivial, as bank interactions lead to a complex dynamic of queues and cascades. State-of-the-art models in the field such as that of Galbiati and Soramäki (2011) use a combination of multi-agent simulation and game-theoretic analysis to understand the relationship between bank decisions and delay costs. However, they do not yet explore realistic queueing protocols within banks and treat all incoming payments as identical. One important aspect in real payment systems is prioritisation, whereby banks will internally delay certain payments and prioritise others due to both internal and external factors (Becher et al., 2008). In this paper we extend earlier work (Galbiati and Soramäki, 2011) by introducing a novel framework to model this prioritisation behaviour in a multi-agent setting. We then consider a game-theoretic model in which banks

[4] The difference between the unsecured interbank rate and the secure-lending repo rate.

optimise both their liquidity sourcing and their internal queueing strategy with the aim of minimising total expected costs. We discuss the equilibrium behaviours that evolve in this system and how inefficiencies may be reduced by utilising measures that are available to the Bank of England, specifically liquidity-saving mechanisms and throughput requirements.

2. Related literature

Models of interbank payment systems have traditionally taken one of two forms. The first form is simulation based on empirical data, attempting to capture as much detail as possible about the mechanics of the settlement process using a simulator such as the Bank of Finland's BoF-PSS2 (2015). Leinonen (2005) provides a comprehensive overview of such papers, studying liquidity requirements, liquidity shocks and various liquidity saving mechanisms. Similarly, the optimal timing of intraday payments is studied by Angelini (2000), who compares a simulated optimisation model to empirical data from the Italian interbank market. However a major shortcoming of these studies is that bank behaviour is parameterised using rules based on historical data. Such an approach makes it difficult to capture changes in strategic behaviour in the face of unprecedented scenarios. A change in the behaviour of actors is one major source of systemic risk that needs to be investigated when considering new regulation. This is particularly true of the recent paper by McLafferty and Denbee (2013), which simulates the effect of introducing a liquidity saving mechanism to CHAPS. They predict that

up to 30% less liquidity will be required by the system post-introduction, but this is purely based on historical data and does not incorporate any consideration of the change in the behaviour of the participants, which they themselves acknowledge will have a big impact.

The second approach is to treat the interbank payment system as a multi-player game and to use techniques from game theory to study the strategic behaviour of the participant banks (Bech and Garratt, 2003; Willison, 2004; Martin and McAndrews, 2008). This involves defining banks' strategies in terms of liquidity sourcing and the management of payment streams, then calculating incentive structures that characterise the underlying liquidity game. Costs and benefits are a function of the amount of liquidity sourced and the total payment delays incurred[5]. These studies then use game-theoretic reasoning to compare strategies corresponding to Nash equilibria of the liquidity game with strategies a central planner would impose to optimize overall system performance. We term this latter strategy the *"social optimum"*.

The game theoretic approach has the advantage of being able to model bank behaviour in the face of changes to the incentive structure. Notably, behaviour that optimises costs for an individual bank can be counter-productive at the systemic level. Bech and Garratt (2003) demonstrate that if liquidity is sufficiently expensive, banks will choose to *"free ride"* on the liquidity

[5] The delay cost is a function of the time between receipt of a payment order and settlement of that order, with the *"cost"* to the bank being either reputational for failing to make payments on time, imposed for missing a payment deadline or some combination of similar drivers.

of others in order to make their payments, thus resulting in a *tragedy of the commons* scenario (Hardin, 1968) in which there is insufficient liquidity in the system and all participants incur large delay costs. However, this approach can only achieve analytical tractability by simplifying the system dramatically, usually by reducing the number of payment periods and banks to two. In reality, liquidity is recycled many times throughout the course of the day, forming a complex dynamic of queues and cascades. Beyeler (2007) demonstrates that at low levels of liquidity, payment instructions and payment settlements lose correlation leading to the emergence of payment cascades of all sizes, which is a hallmark of an irreducible complex system

In order to combine the positive aspects of both streams, Galbiati and Soramäki (2011) utilise an agent-based approach, which combines detailed stochastic simulation with game-theoretic analysis. Agent-based models have been shown to be a useful tool for analysing other complex economic systems, such as in the work of Chakrabarti (2000) on Foreign Exchange markets. In order to simulate the settlement process, a Poisson-distributed stream of incoming payments is handled by each bank according to its liquidity, creating a series of payments and queues. Payoff matrices for each bank can be estimated from Monte Carlo simulations of the queuing dynamics under a range of cost parameters. Nash equilibrium strategies and the socially optimal behaviour can then be compared analytically on this basis.

One important element that is missing from this work is the idea of payment prioritisation for banks. Angelini (1998) shows that banks have an

incentive to delay payments for as long as possible if liquidity is too expensive in a fee-based payment system such as Fedwire, but the analogous result for a collateral-based payment system suggests that payments should be made instantaneously if possible, as collateral is a sunk cost incurred at the start of the day. However Becher (2008) demonstrates that there are other drivers encouraging banks to prioritise some payments over others. These are both internal factors, such as bilateral net credit limits or counterparty importance, and external factors, such as market timings and throughput restrictions, whereby banks may have to make a certain percentage of payments by a specified time to avoid a fine. McAndrews and Rajan (2000) highlight the concern that a high number of delayed payments may lead to a peak in activity towards the end of the day, which heightens operational risk in the US Fedwire system. One of the measures specified in the Basel Committee paper (2012) is *"the volume and value of time-specific and other critical obligations"*, so that the importance of such priority payments can be assessed across the system as a whole.

Prioritisation has been introduced into the two-period game-theoretic literature by Merrouche and Schanz (2009), where payments are treated in a binary manner as either priority or non-priority. This partition allows the implementation of a *Balance-Reactive Gross Settlement* strategy (Norman, 2010), which is implemented as follows. Each bank chooses a buffer size on a daily basis. All priority payments are submitted immediately to the RTGS system as there is no advantage to not making these payments immediately.

However, if at any point the bank receives a non-priority payment and the balance of its liquidity is equal to or less than its buffer size, the bank will hold the non-priority payment in an internal queue. Once the liquidity balance rises back above the buffering threshold and there are no payments queued centrally within the RTGS, the bank will release as many payments as it can from its internal queue. This strategy can be implemented on a decentralised basis by individual banks or through an RTGS system with centralised queuing and liquidity reservation functionality, such as TARGET2.

Galbiaki and Soramäki (2010) have investigated the addition of priority payments as part of their study on liquidity saving mechanisms. However their model uses an unrealistic queuing system, whereby once a non-priority payment is queued internally, it remains unsettled until the end of the day, regardless of the liquidity balance of the bank[6]. Under these rather extreme conditions, the socially optimal strategy is to delay either all or none of the non-priority payments, dependent on liquidity cost.

The novel contribution of our model is to implement a *Balance-Reactive Gross Settlement* system in a multi-period setting and to incorporate the level of liquidity buffer into the banks' strategic decision-making. This allows us to assess a range of buffering strategies under different parameter conditions, with our analysis below indicating that the socially optimal strategy can vary from the dichotomous regime posited by Galbiati and Soramäki.

[6] This setting represents the special case of infinite buffer size in the model specification we propose in Section 3.

3. Methods

3.1. Overview

We simulate a payment system over a number of days and assess how bank behaviour evolves. The behaviour of bank i on any particular day can be formulated as a tuple $\{l_i(0), b_i\}$, where $l_i(0)$ is the amount of liquidity requested from the central bank at the start of the day and b_i characterises the bank's buffering strategy. In practice, applying a buffer means withholding non-priority payments in an internal queue if the amount of liquidity in the RTGS account is equal to or less than the buffer size. Limiting the banks to a single decision at the start of the day is consistent with behaviour within CHAPS. Large banks will be members of multiple payment systems, meaning that the allocation of collateral must be planned in advance and little unencumbered collateral will be available throughout the day. Similarly, maintaining a consistent intraday strategy for internal queueing is necessary to minimise operational risk unless there is a significant change to the proportion or relative weight of priority payments.

In our analysis we distinguish between processes at two separate timescales. Behaviour is assessed at an *interday* level, where banks can alter their liquidity and buffering strategies in order to minimise their expected cost. This cost is approximated using simulations of the *intraday* payment dynamics for all combinations of strategic bank behaviour, with respect to liquidity l and buffer size b. Our model does not necessarily assume that banks run

such simulations to evaluate their options; an equally valid interpretation would be that banks use historical data and accumulated knowledge about best responses to understand the incentive structure of the system.

3.2. Intraday

The day is modelled as a series of T discrete time-steps. The choice of discretisation is motivated by two main reasons. The first is that in real payment systems, the arrival of orders is likely to be clumpy due to non-continuous feeds from trading desks and other institutions. The second and more important reason is that there is no concept of delay cost beyond a certain degree of granularity. This is demonstrated by the US Fedwire system, where interest on an overdraft is only calculated on the balance at the end of each minute and any intra-period volatility from payment netting is ignored (McAndrews and Rajan, 2000). To model payment dynamics, a series of payment orders is randomly generated throughout the day for each bank using a Poisson process with $\lambda = 1/N$, where N is the number of banks. These orders can be viewed as emanating from other parts of the bank, or from external clients. The payee is fixed to be equally likely from the set of all other banks[7]. Consistent with previous work (Galbiati and Soramäki, 2011), we set T to be 3000, representing time periods of approximately ten

[7] For implementation, it is simpler to generate payments from a central server with $\lambda = 1$, and a constraint that banks cannot pay themselves. It is also simpler to generate the payee on payment, rather than storing that information for any queued payments. Both of these methods cause no loss of generality to the results.

minutes. This means that each institution has 200 payments per day on average.

The simulation is run with fifteen identical banks[8] to represent the members of the CHAPS system who directly face the central bank. We assume that these banks are linked by a complete network, so that payments can be between any pair of banks. Bank homogeneity is not an entirely accurate representation of CHAPS where the majority of payments are made by the top four counterparties (Becher et al., 2008), but simplifies the analysis and is a common assumption in previous models (Galbiati and Soramäki, 2011; Bech and Garratt, 2003).

All payments are of unit size, but we assign a binary partition of payments to be either priority or non-priority. Priority payments carry a higher delay cost than non-priority payments, but are also less frequent. We introduce a further parameter, α, which determines the probability of any particular payment being priority. Therefore in any given day, the expected number of priority payments P and non-priority payments NP for each bank are as follows:

$$\mathbb{E}(P) = \frac{\alpha T}{N}, \quad \mathbb{E}(NP) = \frac{1 - \alpha T}{N} \tag{1}$$

Payment orders are processed by each bank using a mechanistic system

[8] The actual number of banks directly facing the Bank of England in CHAPS has now increased to twenty one, but that does not alter the nature of the simulation.

of queuing, dependent upon the amount of liquidity available at a given time. There may be occasions where payments orders will not be executed instantaneously, even when there is sufficient liquidity to do so, due to the internal buffering.

Let $z_i^P(t)$ represent the number of priority payment orders received by bank i up to time t, and $x_i^P(t)$ the number of priority payment orders executed by bank i up to time t. $z_i^{NP}(t)$ and $x_i^{NP}(t)$ have equivalent definitions for non-priority payments. Therefore at time t, bank i will have two queues of payments:

$$q_i^P(t) = z_i^P(t) - x_i^P(t), \quad q_i^{NP}(t) = z_i^{NP}(t) - x_i^{NP}(t) \tag{2}$$

and a liquidity of:

$$l_i(t) = l_i(t-1) - x_i^P(t) - x_i^{NP}(t) + y_i(t) \tag{3}$$

where $y_i(t)$ represents the value of payments received by bank i up to time t. Bank i is indifferent as to whether it is receiving a priority or a non-priority payment, hence we do not need to distinguish.

We assume that all banks adopt the following payment rules for each new time period:

1. if new payment instructions have been received, add them to $q_i^P(t)$ or $q_i^{NP}(t)$ as appropriate.

2. if $l_i(t) > 0$ and $q_i^P(t) > 0$, make priority payments on a FIFO basis until either of these equations is no longer satisfied[9].
3. if $l_i(t) > b_i$ and $q_i^{NP}(t) > 0$, make non-priority payments on a FIFO basis until either of these equations is no longer satisfied[9].

As the system is closed and all outgoing payments represent incoming payments to other banks, these equations fully describe the settlement process. The discretisation of time means that chains of payments are assumed to occur *instantaneously* in the same time period, which is a realistic assumption given a period length of ten minutes.

Let $\psi_i(k)$ be the k^{th} priority payment request and $\phi_i(k)$ the k^{th} non-priority payment request received by bank i. If $t_{\psi_i(k)}$ and $t'_{\psi_i(k)}$ are the order time and payment time respectively for $\psi_i(k)$, then the total expected delays suffered by bank i for each payment type can be expressed using the following equations:

$$D_i^P = \mathbb{E}(\sum_k t'_{\psi_i(k)} - t_{\psi_i(k)}), \quad D_i^{NP} = \mathbb{E}(\sum_k t'_{\phi_i(k)} - t_{\phi_i(k)}) \tag{4}$$

Both D_i^P and D_i^{NP} are generated by stochastic processes. Expected delays are a function of the strategic behaviour of all participant banks, but also

[9] The model is a stylised version of our description in Section 2, but shows equivalent behaviour. In practice, the P queue will be within the RTGS system and the NP queue will be held internally by the bank. Similarly, all P payments will be released into the RTGS system which will automatically apply FIFO, and NP payments will only be released into the RTGS system if there is sufficient liquidity to settle them.

depend on the frequency of priority payments.

$$D_i^P = f(l_1(0)...l_N(0), b_1...b_N, \alpha), \quad D_i^{NP} = g(l_1(0)...l_N(0), b_1...b_N, \alpha) \quad (5)$$

By simulating the intraday process a sufficient number of times, we are able to construct estimates for D_i^P and D_i^{NP} under different parameter combinations. The analysis is simplified by a key result of Galbiati and Soramäki (2011) which proves that the resulting game is an *aggregation game*. This implies that the outcome for an individual is reducible to a function of its own actions and the *sum* of others' actions. Therefore for any bank, the delay process is not dependent on the exact distribution of liquidity and buffer choices by other participants, but only on the bank's own choices and the average of the other participants' choices.

$$D_i^P = f(l_i(0), L, b_i, B, \alpha), \quad D_i^{NP} = g(l_i(0), L, b_i, B, \alpha) \quad (6)$$

where

$$L = \sum_{j \neq i} l_j(0), \quad B = \sum_{j \neq i} b_j \quad (7)$$

We have reduced the problem to each bank playing against the *"system"* of other banks, and can now represent expected delays for both priority and non-priority payments using just six parameters. Extending Property 2 from Galbiati and Soramäki (2011), we can be comfortable that D_i^P and D_i^{NP} are

both convex with respect to liquidity and buffer choices, which ensures that a unique minimum cost exists.

3.3. Interday

In the following analysis we assume banks to be risk neutral. Hence, on an interday basis, banks will aim to adapt their strategies in order to minimise their expected daily cost C. We change the notation slightly to drop the time dependence of liquidity used in intraday analysis, so $\{l_i, b_i\}$ will now represent bank i's behavioural choices on a particular day, with L and B as before. C is comprised of a cost for the liquidity sourced and costs for the delays of both priority and non-priority payments:

$$C = \beta l_i + \gamma D_i^P(l_i, b_i, L, B, \alpha) + D_i^{NP}(l_i, b_i, L, B, \alpha) \tag{8}$$

Relative to the cost of delaying a non-priority payment for one unit of time, which we can set to be one without loss of generality, β is the cost per unit of liquidity sourced and γ is the cost of delaying a priority payment for one unit of time. We then fix a value for α, and simulate the intraday process under each combination of l_i,b_i,L and B to obtain estimates for D_i^P and D_i^{NP}. As this process is stochastic in terms of the incoming payment distribution, Monte Carlo estimates for averages are used[10]. As a result we obtain two 4-dimensional surfaces for D_i^P and D_i^{NP} which can be amalgamated to form

[10] Empirically, we found that the average of ten sets of 10,000 runs for each combination was effective.

a 4-dimensional surface for C. This surface effectively represents the payoff matrix for the game, where one bank is playing against the aggregate average decision of all the others. C is convex following the convexity of

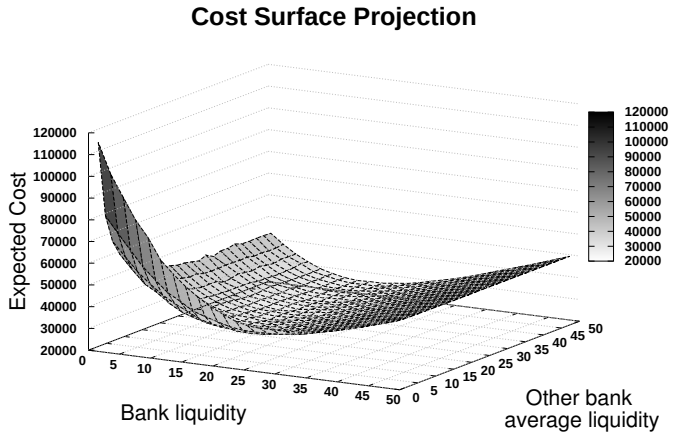

Figure 1. An example cost surface C, with $\alpha = 0.5$, $\beta = 1000$, $\gamma = 20$ and a buffer of zero for all banks

both D_i^P and D_i^{NP}, as demonstrated by the example cost surface in Figure 1 which is projected onto a 2-dimensional surface by setting the buffer equal to zero for all banks. From the combination of the convexity of C and the aggregation property, the social optimum is reached when all banks adopt the same liquidity sourcing and buffering strategies (Galbiati and Soramäki, 2011). Therefore the optimal strategy for any bank can be found at the point on the payoff surface that minimises C whilst satisfying l_i=L/N and b_i=B/N , which is computationally easy to find.

In order to find the Nash equilibria we implement adjustment dynamics based on fictitious play (Brown, 1951). One round of adjustment corresponds to one day in real time and banks adjust their strategies based on Bayesian

learning of their counterparts' behaviour[11], so that after the strategies of each bank have converged to a fixed distribution[12], the proportion of appearances of any particular strategy in a given time period represents its weight in the mixed-strategy Nash equilibrium (Fudenberg and Levine, 1998).

4. Results

In this section, we start by comparing our model to a baseline case where no payments are internally queued. We then systematically alter three parameters - the proportion of priority payments α, the unit cost of liquidity β and the relative cost of priority payment delays γ - and use our simulation to compare the liquidity and buffer choices forming both the socially optimal and Nash equilibrium strategies under each permutation. We also examine the cost associated with each of these strategies.

4.1. The benefit of a buffering strategy on the social optimum

Our first scenario is to test whether a non-zero buffering strategy can improve the social optimum of the system and what impact it will have on liquidity sourcing. We create cost surfaces using the parameter sets $0 \leq \beta \leq 3600$ and $0 \leq \gamma \leq 25$, and set the proportion of priority payments to be 0.5. The range for β is inferred from the typical ratio of liquidity to payments in CHAPS (Galbiati and Soramäki, 2011), while the range for γ is chosen to

[11] See Appendix A for a more complete description.

[12] Although this is not guaranteed to happen under fictitious play, all of our simulations successfully converged.

reflect the diversity in relative importance between priority and non-priority payments. Using the method described in Section 3.3, we find the socially optimal cost for each surface and assess the liquidity and buffering required at that point. For clarity in Figure 2, we then average the results across β.

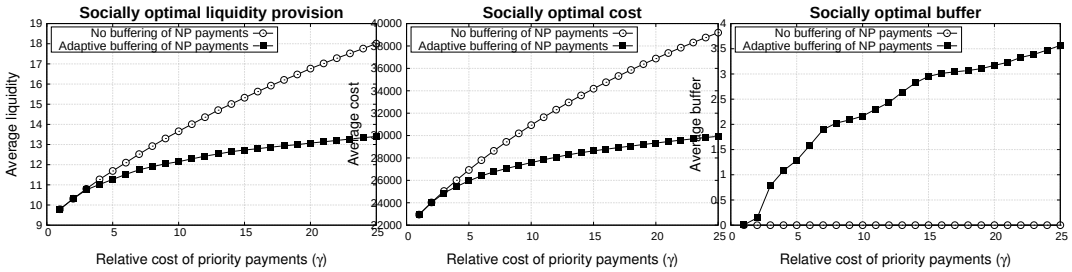

Figure 2. Socially optimal strategies both with and without internal queueing

NOTE: Results for the dependence of liquidity, total cost and buffer size on the parameter γ, the cost of priority payment delays. The black squares show the optimal liquidity and cost that can achieved if no internal queuing is allowed, while the white circles display the optimal values that can be achieved if all banks utilise the best possible buffering strategy for the given parameter combination. The proportion of priority payments $\alpha = 0.5$ and the results are averaged across β.

Figure 2 demonstrates two important points. The first is that buffering can reduce the socially optimal cost of the system as a whole, with the magnitude of the saving increasing with the relative cost of priority payment delays. The second is that adaptive buffering also reduces the total liquidity demand of the system, with the liquidity saving also increasing with γ. Therefore internal queuing can actually aid system efficiency and does not need to be the *"second-best"* approach implied by previous work (Galbiati and Soramäki, 2010).

It is also of interest to see how the optimal buffering strategy varies with γ (right panel). The adaptive buffer grows unsteadily as the cost of priority payment delays increases, with the unsteadiness deriving from the fact that

the buffer can only take a discrete value in our model, precluding a smooth curve. The optimal buffer size and liquidity provision appear to grow at approximately the same rate, indicating that the core (i.e. unbuffered) liquidity in the system remains constant, irrespective of priority payment delay cost.

4.2. Nash equilibrium behaviour

Our second scenario is to evolve the Nash equilibrium strategies based on the parameter combinations in Section 4.1 to see how they compare to the social optimum. We display liquidity and buffer surfaces to visualise how behaviour changes with both liquidity cost and relative priority payment delay cost.

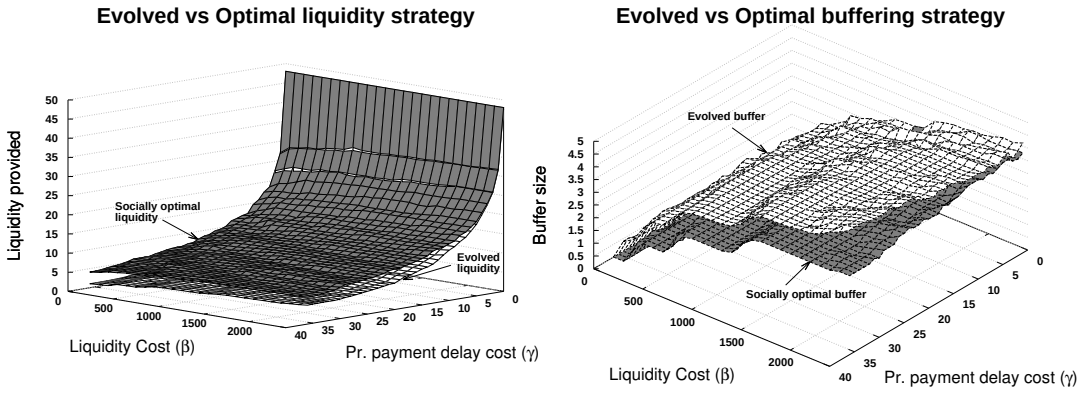

Figure 3. Socially optimal vs Nash equilibrium strategies

NOTE: The dark surfaces represent the socially optimal liquidity and buffering, while the light surfaces represent the liquidity and buffering that are the evolved Nash equilibria of banks participating repeatedly in fictitious play. The proportion of priority payments $\alpha = 0.5$.

Figure 3 demonstrates that banks will consistently *under*-provide liquidity to the system and *over*-buffer their non-priority payments. Liquidity under-provision has been demonstrated when there are no priority payments

(Galbiati and Soramäki, 2011), but it is interesting to see that the behaviour is persistent even as γ increases, when we might expect banks to provide substantially more liquidity in fear of the punitive cost of priority payment delays. Instead, banks increase their level of buffering to protect against this risk, but do so too prohibitively, decreasing the core liquidity in the system. The socially optimal buffer actually decreases slightly as liquidity cost increases due to buffering representing a greater proportion of liquidity in the system, but the Nash equilibrium buffer remains relatively unchanged.

As the socially optimal behaviour will minimise a bank's costs, any variance from it by the Nash equilibrium represents an inefficiency. We can investigate this further by understanding dependence on the proportion α of priority payments. In both Figures 4 and 5, we compare the reference case of all payments being identical (α=0) to a scenario of infrequent priority payments (α=0.1) and very frequent priority payments (α=0.5). In order to approximate continuous system behaviour, we fit our data to continuous functional forms[13].

The right hand graph of Figure 4 makes it clear that over-buffering grows with both α and γ. As either the relative cost or frequency increases, banks have a greater pressure to meet priority payments without delay and consequently adopt progressively more defensive strategies of *liquidity hoarding*.

[13] The discretisation of the liquidity and buffer choices is an artefact of our modelling setup. In reality, banks are able to choose their strategies from a continuous range. We therefore use least squares curve fitting to remove the discrete element from our results.

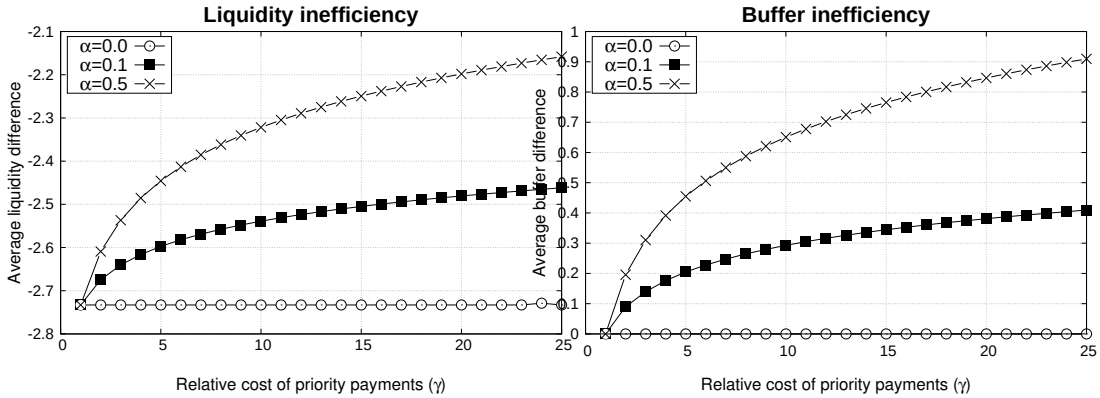

Figure 4. Strategy inefficiencies as a function of relative priority payment delay cost

NOTE: The y axis represents the difference between the socially optimal strategy and the Nash equilibrium strategy. The inefficiencies of the two graphs are negative and positive respectively, representing under provision of liquidity and over-buffering of non-priority payments. Results are averaged across β.

The left hand graph shows that banks *do* provide more initial liquidity to the system as α and γ increase, but it is not enough to compensate for the over-buffering and is still inefficient relative to the social optimum.

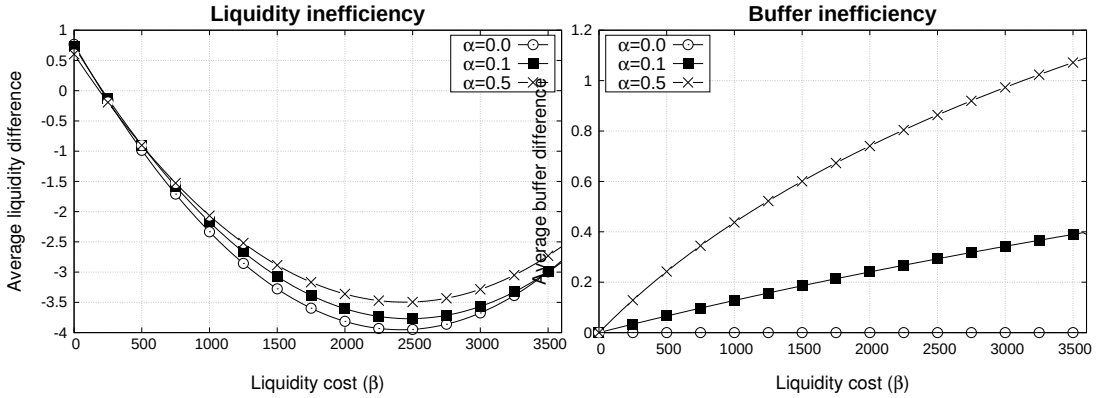

Figure 5. Strategy inefficiencies as a function of liquidity cost

NOTE: As with Figure 4, the y axis represents the difference between the socially optimal strategy and the Nash equilibrium strategy. Results are averaged across γ.

The right hand graph of Figure 5 illustrates a similar dependency to

Figure 4, with over-buffering increasing with both α and β, but this is no longer due to liquidity hoarding. As the price of acquiring liquidity increases, holding a buffer becomes relatively more expensive as it represents a greater proportion of liquidity within the system. Therefore the benefits of making priority payments in a prompt manner are more than offset by the excess queues caused by a reduction in core system liquidity and it is more efficient to reduce buffer size (in parameter terms, the cost of liquidity begins to dominate the relative cost of delaying priority payments as the driver for socially optimal buffering behaviour). However as can be seen from Figure 3, the Nash equilibrium buffering strategies remain relatively unaffected across all values of β, meaning that they become increasingly inefficient as the cost of liquidity increases.

Under-provision of liquidity initially becomes more pronounced as β increases, but starts to improve once the cost of liquidity becomes very high. This is due to liquidity becoming so expensive that the socially optimal amount for the bank to choose begins to approach zero, thus reducing the absolute distance to the Nash equilibrium liquidity. Interestingly the proportion of priority payments has little effect on under-provision, indicating that liquidity cost is the dominant driver of liquidity strategy.

4.3. Breakdown of cost by buffering and liquidity

In our final scenario we assess the importance of both buffering and over-buffering in terms of the bank's expected cost. We calculate the efficiency

of various strategies by comparing them to a baseline, which is the expected cost that the bank would incur by using the socially optimal strategy for both liquidity and buffering[14]. Although buffering and liquidity choices are not independent, we can estimate the impact of buffering by *imposing* the optimal buffering strategy on all the banks and only evolving their liquidity choice. The expected cost difference in that scenario can then be compared to the expected cost difference obtained if both liquidity and buffering are co-evolved. We also compare to a case where liquidity is evolved but the buffer is fixed to zero.

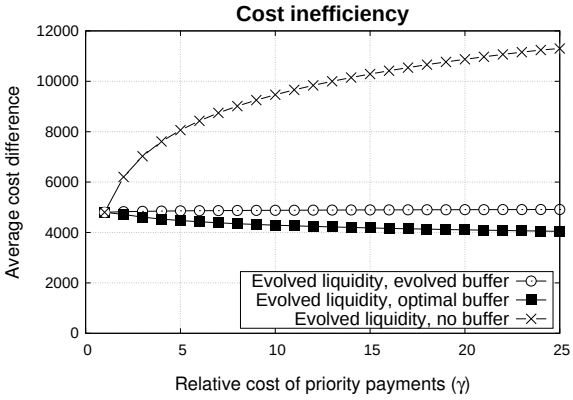

Figure 6. Expected extra cost incurred by bank due to strategy inefficiencies

NOTE: The y axis represents the expected extra cost incurred by a bank relative to a strategy that is socially optimal for both liquidity and buffering. Crosses represent Nash equilibrium strategies where buffering is not allowed at all. Black squares represent equilibrium strategies where the buffer is fixed at the social optimum and only liquidity is evolved. White circles represent equilibrium strategies where both buffer and liquidity are co-evolved. Results are averaged across β.

Figure 6 clearly demonstrates that allowing banks to evolve an internal

[14] This optimum requires that all other banks use the same strategy as well, so must be *imposed* by some central controller.

buffer (circles) leads to substantial savings over the zero buffer case (crosses), even though the buffer that they evolve is too large. The overall cost incurred by the bank when it evolves a buffer is substantially closer to the social optimum than the zero buffer case for all values of relative priority payment delay cost. We can also see that the main driver of cost inefficiency is under-provision of liquidity, as over-buffering (the difference between the circles and squares) only represents up to around 10% of the total excess over the social optimum. However it is clear that over-buffering becomes proportionally more damaging as the relative cost of priority payment delays increases.

Interestingly, enforcing an optimal buffering strategy (squares) leads to an improvement in the cost differential with the social optimum as the relative cost of priority payment delays increases. This is because the bank now has to provide more liquidity rather than over-buffer as a defence against the severe delay costs, thus bringing its liquidity closer to the social optimum and reducing its overall expected cost.

5. Conclusion

In this paper, we develop a detailed model of buffering and prioritisation within a collateralised payment system such as CHAPS. Using stochastic simulation of queuing and buffering dynamics combined with game-theoretic analysis allows us to assess changes in the strategic responses of participant banks to a variety of different regulatory scenarios.

We start by constructing an optimal system in which a benevolent central

planner controls the liquidity and buffering decision of each bank. Considerations of payment prioritisation change the framework of previous studies (Galbiati and Soramäki, 2011; Bech and Garratt, 2003) in an important regard. We show that socially optimal states exist in which banks delay a proportion of non-priority payments by internal queuing. This behaviour becomes more prominent as the proportion or delay cost of priority payments relative to ordinary payments increases. Hence we demonstrate that efficient internal queuing by banks can improve system efficiency while utilising less liquidity.

Having analysed socially optimal choices, we then focus on a game-theoretic analysis of strategic behaviour. We demonstrate that in comparison to the social optimum, the Nash equilibrium is to under-provide liquidity and over-buffer non-priority payments. The severity of this hoarding-like effect becomes more pronounced as the proportion or relative delay cost of priority payments increases. At a system level, the result of strategic interactions is an excess cost and we analysed the dependence of this cost on various parameters. However we also demonstrated that the equilibrium formed by strategic interactions is far superior from a cost perspective to the equilibrium formed if banks are not allowed to buffer at all, hence showing that adaptive buffering by banks can increase system efficiency. Finally, we show that the relative increase in cost attributable solely to over-buffering becomes proportionately larger as the cost of priority payment delays increases.

Our findings allow some general observations regarding the efficacy of two

regulatory options available to the Bank of England. CHAPS has recently introduced a Liquidity Saving Mechanism (**"LSM"**) (Dent and Dison, 2012), whereby an algorithm can settle off-setting payments between counterparties which are submitted to a central LSM queue. The benefit of such an algorithm is that it does not require any liquidity and does not reintroduce settlement risk, as payments are considered unsettled within the queue until they are successfully offset. This incentivises banks to submit payments early to the LSM in order to maximise the chance of them being offset (Willison, 2004; Davey and Gray, 2014). However, the LSM cannot be used in isolation, as not all payments can be offset in a timely manner. Therefore it is combined with the RTGS to form a two-channel hybrid system.

Previous work (McLafferty and Denbee, 2013; Galbiati and Soramäki, 2010) suggests that using an LSM can lead to substantial liquidity savings by dividing payments into two streams, with all priority payments settled through the RTGS and all non-priority payments settled via the LSM. An empirical study by the Bank of England (Davey and Gray, 2014) since the introduction of the LSM demonstrates that banks behaving in this manner has led to liquidity savings, albeit to a lesser extent than anticipated. However, this behaviour removes any potential benefit from efficient queuing by the banks, which we have shown to exist (see section 4.3). It also causes a negative correlation between the LSM and the RTGS, as high LSM usage reduces liquidity recycling within the RTGS. This leads to some *"bad"* equilibrium strategies (Galbiati and Soramäki, 2010) which carry a higher cost than the base case

without an LSM. Indeed, Davey and Gray (2014) also show that the number of payments queued within the system has increased, and it is unclear how the empirical results obtained would change if the system were to become stressed, by either liquidity becoming more expensive or the relative cost of payment delays increasing.

Instead we suggest the following implementation. Any non-priority payments which are internally queued by banks due to their buffering strategy should be placed initially into the LSM. However if a bank subsequently has sufficient liquidity to exceed its buffer and the LSM algorithm has not already settled the queued payments, they should be removed from the LSM and settled instead through the RTGS[15]. This should lead to an improvement in system efficiency without the danger of unexpectedly poor Nash equilibrium strategies, as the positive benefits of internal queuing have not been compromised.

The second regulatory option available to the Bank of England is the implementation of throughput rules, which establish the minimum proportion of a bank's daily settlement flow that must be settled by a particular time (Ball et al., 2011). The current arrangement within CHAPS is for 50% of daily flow to be settled by noon and 75% of daily flow to be settled by 2.30pm. However enforcement of these rules is by peer pressure rather than financial sanctions.

[15] Within the new CHAPS LSM, it is possible to switch payments between the two streams at any time.

Previous work on throughput rule design by Buckle (2003) raises the issue of whether bilateral throughput limits between counterparties would work better than a generic throughput restriction. However this would increase the cost discrepancy between priority and non-priority payments (γ), as payments to certain counterparties would become progressively more important if the bilateral credit position between the two became substantially one-sided. From our results (see section 4.2), we can see that a higher relative priority payment delay cost leads to an increase in over-buffering and a greater system-wide cost.

Similarly, Ball (2011) and Norman (2010) suggest more extensive monitoring of banks' payment behaviour when calculating throughput, in order to discourage strategic delays. However any throughput requirements which focus on the settlement of specific payments effectively ascribes those payments an increased priority, and so the overall proportion of priority payments (α) will increase. This also leads to over-buffering and additional costs (see section 4.2).

Therefore, it appears that a *generic* throughput requirement to homogenise all payments is a better solution, as it will reduce both the relative cost of priority payments (γ) and their frequency (α). Clearly this homogenisation can only be achieved if the incentive to meet the strict throughput requirements is sufficient, as otherwise banks will naturally start to ascribe some payments priority over others. Therefore we would suggest that the flow percentages are enforced on a daily basis, rather than using the current monthly

average which gives the banks opportunity to under-perform with regards to their obligations for long periods. In addition, a stronger enforcement mechanism than peer pressure should be implemented, as suggested by Ball (2011). Davey and Gray (2014) suggest that throughput has been improved by the implementation of the LSM, due to banks submitting their payments earlier in order to maximise offsetting possibilities, but it is unclear whether this does actually lead to increased settlement in all parameter regimes or simply an increase in payment queuing.

A natural extension of our model would be to explicitly introduce an LSM in order to test our suggested implementation. Another potential extension would be to investigate how banks adapt their buffering strategy to changes in the Poisson parameter used for incoming payments. This would allow a study of stress scenarios, where one or more banks cannot make payments for operational reasons and become liquidity sinks. Analysing how the other banks adapt their payment strategies under these conditions is an interesting avenue for further research.

April 14, 2015

References

Afonso, Gara and Hyun Song Shin (2011). "Precautionary Demand and Liquidity in Payment Systems". *Journal of Money, Credit and Banking*, 43(7), 589–619.

Angelini, Paolo (1998). "An analysis of competitive externalities in gross settlement systems". *Journal of Banking & Finance*, 22(1), 1–18.

Angelini, Paolo (2000). "Are banks risk averse? Intraday timing of operations in the interbank market". *Journal of Money, Credit and Banking*, 32(1), 54–73.

Ball, Alan, et al. (2011). "Intraday liquidity: risk and regulation". *Bank of England Financial Stability Paper*, 11, 1–25.

Bank of Finland (2015). "The Bank of Finland Payment and Settlement Simulator". URL `http://pss.bof.fi/Pages/Default.aspx`.

Basel Committee on Banking Supervision (2012). "Monitoring indicators for intraday liquidity management". URL `http://www.bis.org/publ/bcbs225.htm`.

Bech, Morten L and Rod Garratt (2003). "The Intraday Liquidity Management Game". *Journal of Economic Theory*, 109(2), 198–219.

Becher, Christopher, Marco Galbiati, and Merxe Tudela (2008). "The Timing and Funding of CHAPS Sterling Payments". *FRBNY Economic Policy Review*, 14(2), 113–133.

Beyeler, Walter E, et al. (2007). "Congestion and Cascades in Payment Systems". *Physica A: Statistical Mechanics and its Applications*, 384(2), 693–718.

Brown, George (1951). "Iterative Solutions of games by fictitious play". In "Activity Analysis of Production and Allocation", edited by T Koopmans. Wiley, New York, New York, USA.

Buckle, Simon and Erin Campbell (2003). "Settlement bank behaviour and throughput rules in an RTGS payment system with collateralised intraday credit". *Bank of England Working Paper*, 209, 1–37.

Chakrabarti, Rajesh (2000). "Just another day in the inter-bank foreign exchange market". *Journal of Financial Economics*, 56, 29–64.

Coleman, Stacy P (2002). "The Evolution of the Federal Reserve's Intraday Credit Policies". *Federal Reserve Bulletin*, February, 67–84.

Davey, Nick and Daniel Gray (2014). "How has the Liquidity Saving Mechanism reduced banks' intraday liquidity costs in CHAPS?" *Bank of England Quarterly Bulletin*, 5(1), 180–189.

Dent, Andrew and Will Dison (2012). "The Bank of England's Real-Time Gross Settlement infrastructure". *Bank of England Quarterly Bulletin*, Q3, 234–243.

Federal Reserve Board (2014). "Fedwire Funds Services". URL `http://www.federalreserve.gov/paymentsystems/fedfunds_about.htm`.

Freixas, Xavier, Bruno M Parigi, and Jean-Charles Rochet (2000). "Systemic risk, interbank relations, and liquidity provision by the central bank". *Journal of Money, Credit and Banking*, 32(3), 611–638.

Fudenberg, Drew and David K Levine (1998). *The Theory of Learning in Games*. MIT Press, Cambridge, MA.

Furfine, Craig and Jeff Stehm (1998). "Analyzing alternative intraday credit policies in real-time gross settlement systems". *Journal of Money, Credit and Banking*, 30(4), 832–848.

Galbiati, Marco and Kimmo Soramäki (2010). "Liquidity-saving mechanisms and bank behaviour". *Bank of England Working Paper*, 400, 1–27.

Galbiati, Marco and Kimmo Soramäki (2011). "An agent-based model of payment systems". *Journal of Economic Dynamics and Control*, 35(6), 859–875.

Hardin, Garrett (1968). "The Tragedy of the Commons". *Science*, 162(3859), 1243–1248.

Humphrey, David B (1986). "Payments Finality and Risk of Settlement Failure". In "Technology, and the Regulation of Financial Markets: Securities, Futures and Banking", edited by A. Saunders and L. White, pages 97–120. Lexington Books, Lexington, MA.

James, Kevin and Matthew Willison (2004). "Collateral posting decisions in CHAPS Sterling". *Bank of England Financial Stability Review*, December, 99–104.

Leinonen, Harry (2005). "Liquidity, risks and speed in payment and settle-

ment systems: A simulation approach". *Bank of Finland Studies*, E31, 1–354.

Martin, Antoine and James McAndrews (2008). "Liquidity-saving mechanisms". *Journal of Monetary Economics*, 55(3), 554–567.

McAndrews, James and Samira Rajan (2000). "The Timing and Funding of Fedwire Funds Transfers". *FRBNY Economic Policy Review*, July, 17–32.

McLafferty, Joanna and Edward Denbee (2013). "Liquidity Saving in CHAPS: A Simulation Study". In "Simulation in Computational Finance and Economics: Tools and Emerging Applications", edited by Biliana Alexandrova-Kabadjova and Edward Tsang, chapter 7, pages 120–143. IGI Global.

Merrouche, Ouarda and Jochen Schanz (2009). "Banks' intraday liquidity management during operational outages : theory and evidence from the UK payment system". *Bank of England Working Paper*, 370, 1–43.

Norman, Ben (2010). "Liquidity saving in real-time gross settlement systems: An overview". *Bank of England Financial Stability Paper*, 7, 1–11.

Van den Bergh, Paul (1994). "Operational and Financial Structure of the Payment System". In "The Payment System: Design, Management and Supervision", edited by Bruce Summers, chapter 3. International Monetary Fund.

Willison, Matthew (2004). "Real-Time Gross Settlement and hybrid payment systems: a comparison". *Bank of England Working Paper*, 252, 1–30.

Appendix A. Fictitious Play

Bank i's beliefs about the aggregate behaviour of others are held by the matrix

$$\begin{bmatrix} p_i^t(0,0) & \cdots & p_i^t(0,B_{max}) \\ \vdots & \ddots & \vdots \\ p_i^t(L_{max},0) & \cdots & p_i^t(L_{max},B_{max}) \end{bmatrix} \tag{A.1}$$

where $p_i^t(j,k) = P_i(L = j \text{ and } B = k \text{ at time t})$. L and B represent the aggregate liquidity and aggregate buffer of all other banks, while L_{max} and B_{max} are upper bounds for the liquidity and buffer totals respectively. Note that t is now in terms of number of days, rather than intraday time as it was in Section 3.2.

Beliefs are updated according to:

$$p_i^t(j,k) = \frac{(1 + \sum_{s=1,\ldots,t-1} I_{j,k}(s))}{t + L_{max}.B_{max}} = p_i^{t-1}(j,k) + \frac{I_{j,k}(t-1) - p_i^{t-1}(j,k)}{t + L_{max}.B_{max}} \tag{A.2}$$

where

$$I_{j,k}(s)) = \begin{cases} 1 & \text{if } L = j \text{ and } B = k \text{ at time } s; \\ 0 & \text{otherwise.} \end{cases}$$

so that combinations that have appeared many times previously will start to acquire a heavier weighting as time increases.

Bank i is now able to make its behavioural choice for liquidity and buffering at time t, $\{l_i(t), b_i(t)\}$, in the following manner:

$$\{l_i(t), b_i(t)\} = arg_{l_i,b_i} min \sum_{j=0}^{L_{max}} \sum_{k=0}^{B_{max}} C(l_i, b_i, j, k) p_i^t(j, k) \tag{A.3}$$

where $C(l_i, b_i, j, k)$ is the cost incurred by bank i if it follows strategy $\{l_i, b_i\}$ and the aggregate strategy for the other banks is $\{j, k\}$.

It is instructive to note that in our simulation, C is independent of time as the payoff matrix for each bank does not change.