
© 2010 Elsevier

Reprinted with permission of the publisher
An agent-based model of payment systems

Marco Galbiati a,⁎,1, Kimmo Soramäki b

a Bank of England, United Kingdom
b Aalto University School of Science and Technology, Helsinki, Finland

Abstract

We lay out and simulate a multi-agent, multi-period model of an RTGS payment system. At the beginning of the day, banks choose how much costly liquidity to allocate to the settlement process. Then, they use it to execute an exogenous, random stream of payment orders. If a bank's liquidity stock is depleted, payments are queued until new liquidity arrives from other banks, imposing costs on the delaying bank. We study the equilibrium level of liquidity posted in the system, performing some comparative statics and obtaining insights on the efficiency of alternative system configurations.

1. Introduction

Virtually all economic activity is facilitated by transfers of claims by financial institutions. In turn, these claim transfers generate payments between banks whenever they are not settled across the books of a same bank. These payments are settled in interbank payment systems. In 2006, the annual value of interbank payments made in the European system TARGET totalled €675 trillion (≈ $880 trillion). In the corresponding US system Fedwire, the amount was $670 trillion, while the UK system CHAPS processed transactions for a value of €68 trillion (≈ $129 trillion) —BIS (2009). In perspective, these transfers amounted to 24–40 times the value of the respective countries’ GDPs. The sheer size of the transfers, and their pivotal role in the functioning of financial markets and the implementation of monetary policy, make payment systems a central issue for policy makers and regulators.

At present, most interbank payment systems work on a real-time gross settlement (RTGS) modality. That is, settlement takes place as soon as a payment is submitted into the system (real time); also, a payment can be settled only if the paying bank has enough funds to deliver the full amount in central bank money (gross settlement). Because no netting takes place...
offsetting payments, RTGS modality imposes high liquidity demands on the banks, making RTGS systems vulnerable to liquidity risk, i.e. to the risk that liquidity-short banks are unable to send their own payments. This may create delays and possibly cause gridlocks in the system (see e.g. Bech and Soramäki, 2002). The 2007–09 global financial crisis has stressed the importance of liquidity, prompting a lively debate on this issue at central banks and financial regulators. At the time of writing, new regulation is being designed throughout the world, to ensure that liquidity is present in sufficient amounts in key financial infrastructures: first of all, in interbank payment systems.

This paper aims at contributing to this knowledge, offering a model of liquidity demand and circulation in an RTGS system. To our knowledge, this is the first paper that explores this question using an “agent-based” approach. As Ehrentreich (2007) put it, a “major advantage [of such approach] is that [it] allows the removal of many restrictive assumptions that are required by analytical models for tractability”. As we will explain in a moment, this is our main reason to adopt such approach.

The amount and the distribution of liquidity in a payment system is the result of a complex interaction the system’s participants. Indeed, during the day, each bank has to make a stream of payments, that can only be partly predicted. To cover the liquidity needs generated by these payments, banks typically rely on two sources: (a) reserve balances or credit acquired from the central bank and (b) funds received from other settlement banks during the course of the day. The first source can be seen as providing external (to the system) liquidity, while the second is a source of internal liquidity. In normal conditions a bank can draw freely on external liquidity. This, however, has a cost, which gives incentives to economize on its use. Internal liquidity on the other hand carries no cost, but its arrival is out of the bank’s control. Hence, reliance on internal liquidity exposes the bank to the risk of having to delay its own payment activity—something which also is costly. As a consequence, a bank has to optimally decide how much external liquidity to acquire, trying to forecast when and how much internal liquidity it will receive, trading off external liquidity costs against (expected) delay costs. The fact that banks (i) delay some payments, and yet (ii) do not wait till the very end of the day to make all their payments, shows that this tradeoff indeed exists.

Two main difficulties emerge when studying the behaviour of banks in a payment system. First, when modelled in enough detail, liquidity flows in RTGS systems follow complex dynamics (see Soramäki et al., 2007), making the bank’s liquidity management problem anything but trivial. Indeed, recent work by Beyeler et al. (2007) shows that, when the level of external liquidity is low, payments lose correlation with the arrival of payment orders; as a consequence, it is difficult to gauge the precise relationship liquidity and delays, making it hard to determine the optimal usage of external funds. Second, the actions of each bank produce spillover effects on the rest of the system, so no system participant can solve its optimal liquidity demand problem in isolation. As strategic interactions are widespread, banks interact in a fully fledged “game”, jointly determining the performance of the system.

This paper studies this liquidity game, putting particular effort into modelling liquidity flows. We thus build a payments model where external liquidity is continuously recycled among many banks, with delays and costs generated in a non-trivial way by a realistic settlement process. Such realism will force us to abandon a purely analytical approach, and to use simulations to find some elements of the game; in particular, the payoff function, that is the relationship the liquidity choices of each bank, and the resulting settlement delays and costs. We are interested in the equilibria of the liquidity game, or in the choices that banks may be seen to adopt in a consistent fashion. To do so we solve the model adopting a dynamic approach. That is, we assume that banks change their actions over time, using an adaptive process whereby actions are chosen on the basis of past experience. We then simulate the resulting dynamics and we look at the limit, or equilibrium, behaviour. The limit clearly depends on the specific form of the adaptive rule, which we therefore choose in such a way that: (a) it is consistent with some rationality on the part of the banks, in the spirit of the “agent-based” approach and (b) it leads to a meaningful equilibrium. In particular, a convergence point of our dynamics will be a Nash equilibrium of the liquidity game. Given that the outcome of our model is a Nash equilibrium, we could have followed the more traditional approach of directly looking for mutual best replies in the liquidity game. This approach would have lead to the same conclusions. However, it is interesting to see that the equilibrium as an emergent property of the system, a hallmark of agent-based models (Tesfatsion, 2001; Axelrod and Tesfatsion, 2006; Herbert, 2007).

Given its game-theoretic approach, this paper is related to recent work by Angelini (1998), Bech and Garratt (2003, 2006), Buckle and Campbell (2003). Our model differ from these contributions in two dimensions.

First, we consider a related but different issue. Previous works look at the timing decisions of banks: the liquidity management game played by the banks consists in deciding when to make a payment. Instead we concentrate on the decision of how much liquidity to post in a ‘collateralized’ RTGS system. That is, in a system where banks must pre-fund their accounts by pledging collateral with the central bank. Incidentally, most systems nowadays use the collateralized modality, with just
one notable exception: (one branch of) the Canadian system LVTS. This is not to say that the timing dimension is not relevant in these systems; only, it is intertwined with that of the liquidity available. Insofar as frictions make it difficult to acquire new liquidity intraday, the liquidity choice is the first, necessary dimension of a bank's strategy.\footnote{Kahn and Roberds (1998) look at the problem of a bank which needs to choose a portfolio of assets, to ensure it has sufficient funds to face random payment orders. This is very similar to what we do, with the key difference we consider a fully fledged game between banks, while there the focus is on a single bank maximization problem.}

Second, our paper innovates in the method. Existing game-theoretic contributions typically consider settings with a few agents\footnote{Bech and Garratt (2006) have the currently most general framework, able to accommodate a generic number of different banks. Each bank, however, needs to make exactly 1 payment to each other bank, and this can be done either ‘in the morning’ or ‘in the afternoon’. This restricts the variety of possible liquidity flows.} and a small number of periods and payments. This is seemingly necessary to avoid technical difficulties that would emerge in more articulated frameworks (where a timing strategy would take much more complex form if, for example, had to be taken in continuous time and possibly depend on other’s past choices). While these models improve our understanding of the incentives in payment systems, the actual payoff functions may be too simple to describe costs in real payment systems accurately. As we said, in RTGS systems liquidity can circulate many times between many banks, generating dynamics that cannot be captured by these simple, but analytically tractable, models.

Recently, a growing literature has used simulation techniques to investigate efficiency and risk issues payment systems (see e.g. James and Wilson, 2004, the volumes edited by Leinonen, 2005, 2007, and Devriese and Mitchell, 2005). Simulation studies have been widely used to compare alternative central bank policies, or to test the impact of new system features before their implementation in payment systems. A shortcoming of such studies has been, however, that participant behaviour is rarely endogenised. The behaviour of banks has either been assumed to remain unchanged across alternative scenarios, or to change in a predetermined manner, leaving aside (or largely simplifying) the strategic aspects studied by the aforementioned game-theoretic works.

Recognizing the strengths and disadvantages of these two approaches, the present paper tries to build a bridge them, combining the strength of each of them. Indeed, a main innovation of this paper is its mixed analytical–numerical approach. Numerical simulations yield us ‘facts’, which are then used to prove analytical results, solve the model, and derive our conclusions. In a sense, simulations here play the part of experiments in physics: empirical evidence reveals properties, which are then used to derive analytical statements. These latter in turn make sense of the observed facts, making it possible to describe them in a coherent way. Needless to say, our observed facts are not evidence from real systems; they are artificially generated by a model which is a realistic but stylized representation of reality.

The paper is organized as follows: Section 2 provides a formal description of the model, describes some properties of the cost function, and illustrates the ratonnement process towards equilibrium. Section 3 presents the results of the experiments and Section 4 concludes. The analytical results in the Appendices justify our modelling choices, and provide ground and explanation for the simulations.

2. Description of the model

The model considers a sequence of ‘days’ in RTGS. At the beginning of each day, banks choose their opening balances; at the end of the day, they receive payoffs, depending on how much liquidity they acquired, and how many delays they suffered. Banks update such choices day after day according to a ‘learning rule’; eventually, they converge to an equilibrium, on which we focus.

Thus, the main time frame of the analysis is between-days. However, the next section will describe an intra-day model, whose function is to determine the relationship between opening balances and delays; in other words, the ‘payoff function’ of the daily liquidity game.

2.1. Banks and liquidity choices

At the beginning of the day, each of \(N\) banks (denoted by \(i=1 \ldots N\)) chooses its liquidity reserves, say \(l_i(0)\), to be used in the course of the day.\footnote{In the simulations, we assume that \(l_i(0)\) is an integer between 0 and some large \(L\).} To simplify, we assume that these reserves of external liquidity can only be acquired once, at the beginning of the day. Once \(l_i(0)\) is bank \(i\)’s action and the vector \(l=(l_1(0),l_2(0)\ldots l_N(0))\) is an action profile. The next section illustrates how payments are received and executed, generating the outcome of the game.

2.2. Payments and delays

The outcome of the game is determined as follows. The day is modelled as a continuous time interval \([0,T]\). Payment orders arrive according to a Poisson process with parameter \(\lambda = 1\), so the system as a whole receives, on average, \(T\) orders a day. Payments are of unit size. The payor and the payee of these payment orders are determined by (uniform) random draws: for any order, the probability that banks \(i\) and \(j\neq i\) are, respectively, the payor and the payee is \((1/N)(1/N-1)\). Equivalently, each
single bank receives payment orders according to a Poisson process with parameter \( \lambda = 1/N \), and the payee of each such order is determined by a random (uniform) draw. Hence, the model is symmetric, as banks are a priori identical. Symmetry simplifies calculations and the exposition of our findings, but it is not essential. In particular, the key Properties 1 and 2 (below) hold even if the system is asymmetric. Hence, all of our analysis could, in theory, be replicated for asymmetric systems; the Appendix further comments on this.

Payment orders can be seen as generated outside the bank, by a bank’s clients, or within the bank, by some area which is different from the treasury department. Whatever the interpretation, payment orders are exogenous for bank \( i \), which chooses only the initial balance \( l_i(0) \).

Let us call \( z_i(t) \) the number of payment orders received by bank \( i \) up to time \( t \), and \( x_i(t) \) the number of payment orders executed by \( i \) up to \( t \). At \( t \), bank \( i \)’s queue (its backlog of outstanding orders) is therefore

\[
q_i(t) = z_i(t) - x_i(t),
\]

where we set \( z_i(0) = x_i(0) = 0 \). Payments orders are executed using available liquidity. Bank \( i \)’s available liquidity at time \( t \) during the day is defined as

\[
l_i(t) = l_i(0) - x_i(t) + y_i(t)
\]

where \( y_i(t) \) is the amount of payments that \( i \) has received from other banks up to time \( t \). For simplicity, we assume that every \( i \) adopts the following payment rule \(^9\):

if \( l_i(t) > 0 \), execute new and queued payments as FIFO;

if \( l_i(t) = 0 \), queue new payment orders.

Bank \( i \)’s incoming payments \( y_i(t) \) are just other banks’ outgoing payments, so the settlement process is fully described by the above equations.

We are interested in a bank’s delays in payment execution. For a payment order \( \zeta \), let us call \( t^\circ \) the time when it is received, and \( t^\prime \) the time when it is executed. Total (expected) delays for bank \( i \) are then defined as

\[
D_i = E[\sum_{j}(t^\prime_j - t^\circ_j)],
\]

where the summation runs over bank \( i \)’s payment orders. It is clear that \( D_i \) is a function of \( i \)’s and other’s liquidity choices; also, it is random, as payment orders arrive in random order. In the next section, we will define \( i \)’s costs (payoffs) in terms of \( D_i \). We therefore assume that (i) banks only care about total delays (hence the summation), and (ii) banks are risk neutral (hence the expectation).

Finding an analytical form for \( D_i \) is beyond the scope of this paper: as mentioned in the Introduction, even this simple model generates extremely complex dynamics of liquidity \( l_i(t) \) and queues \( q_i(t) \).\(^10\) However, the model can be simulated numerically. A given action profile \( l=(l_1(0),\ldots,l_N(0)) \) pins down the initial conditions of the system. From there, the exogenous arrival of payment orders mechanically generates liquidity fluxes, queues and delays. We simulate this process to obtain a numerical representation of the delay function \( D_i \). The simulations reveal an important property of a bank’s delays:

**Property 1.** Aggregation: for a given number of banks \( N \), bank \( i \)’s expected delays are (essentially) a function of its own liquidity and of the sum of others’ liquidity: \( D_i(l_i(0),\ldots,l_i(0)) = D_i(l_i,\sum_{j\neq i} l_j,N) \).

The Appendix discusses this property, showing that it only depends on the fact that (i) a unit of liquidity moves from bank \( i \) to \( j \) with a given probability and (ii) the number of payments is large, compared to the liquidity posted by each bank.

On (i), we assume that this probability is equal for each \( i \) and \( j \) (symmetry). This is done for simplicity but is not needed for Property 1 to hold, as discussed in Appendix. On (ii), we calibrate the model so that each unit of liquidity is circulated about 20 times in a day. This gives a turnover ratio of around 5% which, besides being realistic, is enough for Property 1 to hold well, as discussed in Appendix.

The aggregation property is essential for the paper. Among other things, it leaves us with a game with only two players: bank \( i \) against “the rest of the system”. So, we can e.g. plot the (average\(^11\)) delays of a bank in a 3-D space: this is Fig. 1.

Delays have also a second property:

**Property 2.** Convexity: bank \( i \)’s delays are a convex function of own and others’ liquidity.

Property 2 (rather clear from Fig. 1) is formally proven in Appendix. There, we also show another fact on the curvature of \( D_i \) (from now on, we often omit index \( i \) \(0 < \frac{\partial^2 D_i}{\partial l_i \partial l_j} < \frac{\partial^2 D_i}{\partial l_i^2} \)). Important for analytical results later on, this means that the surface in Fig. 1 is ‘more convex’ along the ‘my liquidity’ dimension, than along the ‘others’ liquidity’ dimension. The Appendix makes it clear that none of these curvature properties relies on the assumption that the system is symmetric.

Fig. 2 illustrates role of Properties 1 and 2: together, they imply Theorems 1 and 2 (still to be presented).

---

\(^9\) This rule is optimal for the cost specification given in the next section: banks need to pay upfront for liquidity, so they have no incentive to delay payments if liquidity is available. Under other specifications (as heterogeneous payment delay costs) this would not be the case.

\(^10\) Queues do not form only when \( l_i(0) \) is high. Then, \( \Delta x = \Delta z \) so executed payments essentially follow a Poisson process which mirrors the arrival of payment orders. But when liquidity is low, the Poisson process is disrupted by the fact that banks run out of liquidity, and payments are queued.

\(^11\) And standardized by the length of the day.
At the end of the settlement day, banks receive payoffs that depend on (i) the liquidity posted at the beginning of the day, and (ii) the delays generated by the settlement algorithm illustrated in the previous section. More precisely, we assume that acquiring initial liquidity \( l_i(0) = l_i \) imposes a liquidity cost equal to \( l_i \) (from now on we use \( l_i \) instead of \( l_i(0) \); at this stage we are not no longer interested in the intra-day timing of payments and intraday liquidity balances. These are all subsumed in the resulting delays). Beside liquidity costs, we assume that a bank incurs a cost proportional to its total daily delays \( D_i \), illustrated in the section above. Hence, a bank’s total costs are

\[
C(l_i, l_j) = l_i + kD_i(l_i, l_j) \quad (2)
\]

where \( l_{-i} \) is a vector indicating the choices of all \( j \neq i \). Given its linearity in delays, \( C \) inherits the key properties of the function \( D \) : aggregation and convexity. This is obvious also from Fig. 3, plotting \( C \) for two levels of \( \lambda \); “low” (blue) and “high” (red). Clearly, when \( \lambda = 0 \) it is \( C = D \), while as \( \lambda \) grows \( C \) tends to a linear function (a plane in 3D).

2.4. Equilibrium

Liquidity choices determine banks’ costs, according to Eq. (2). As costs depend on the choices of everyone, banks engage in a game. To find the equilibrium of this ‘liquidity game’, we use the so called fictitious play tâtonnement process (Brown, 1951). Largely studied in evolutionary game theory, fictitious play is a specification of how players change their actions in time, learning from experience. A precise description of this process is in Appendix; the reason to adopt this particular dynamic is twofold. First, despite its simplicity the fictitious play rule is in a sense rational and thus not too unrealistic, corresponding to Bayesian updating of beliefs about others’ actions.\(^{12}\) Second, fictitious play can indeed be a useful tool to compute equilibria. Indeed, when fictitious play converges to a stable action profile, this is a Nash equilibrium of the underlying game.\(^{13}\) Summing up, fictitious play can be seen either as a computational device, or as a “story” with an appealing economic meaning.

A key question is whether the game has a unique equilibrium and, if not, which equilibrium is uncovered by fictitious play. Theorem 1 below strengthens our equilibrium selection by stating that, even though our model may have different equilibria,

---

\(^{12}\) See e.g. Fudenberg and Levine (1998) p. 31 for details.

\(^{13}\) It is also well-known that fictitious play may fail to converge. However, this is not the case here, as shown by the simulations. This might be a general result for aggregation games, but we have not been able to show this analytically. Interestingly, convergence in aggregation games was shown by Kukushkin (2004) and by Mezzetti and Dindo (2006) for dynamics similar to fictitious play.
all of them are characterized by the same total level of liquidity $\sum l_i$. Hence, in this respect, it is irrelevant which equilibrium is picked by fictitious play.

**Theorem 1.** All equilibria feature the same total liquidity.

This results makes it possible to speak about the equilibrium liquidity level, and so allows us to perform comparative statics (i.e. to change the model's parameters and look at how the liquidity level changes).

3. Results

3.1. Liquidity demand and efficiency of the equilibrium

We start with a base case scenario with 15 banks and 200 payments per bank. The number of banks is chosen to have a system that “looks like” the UK system CHAPS. The number of payments is instead arbitrary. In all of the simulations banks interact in a complete network, i.e. each sends payments to every other bank in the system.

First, we obtain a liquidity demand function, relating the total (equilibrium) liquidity $\sum l_i$ to the unit price of liquidity $\lambda$, for unit delay costs $\kappa$ normalized$^{14}$ to 1. Following the 2007–2009 financial crisis, liquidity regulation is widely discussed at central banks and financial regulators. It is widely recognized that an important effect of such liquidity regulation will be to change the ‘cost’ of liquidity for banks. Hence, it will impact the amount of liquidity that banks would spontaneously hold in payment systems.$^{15}$. Fig. 4 shows our first result (on the vertical axis, the amount of liquidity posted divided the number of payments made). When $\lambda = 0$, liquidity is free and therefore any amount of liquidity that ensures zero delays is an equilibrium. We then cap liquidity demand to 50—which in practice ensures zero delays.$^{16}$ As $\lambda$ increases, the equilibrium liquidity drops following a power function,$^{17}$ until it reaches zero. At this point (not show in the picture) liquidity is so expensive that even the maximum amount of delays is less costly than a single unit of liquidity.

Is such equilibrium liquidity demand efficient? The simple answer is: no. External liquidity is used by all banks, while an individual institution only cares about private costs and benefits. These positive externalities in liquidity provision imply that, in equilibrium, banks try to free-ride on each other and end up under-providing liquidity.

We can attempt and quantify the extent of this inefficiency by comparing the equilibria to what a ‘benevolent’ central planner could achieve. Straightforward in words, this comparison is computationally daunting: with e.g. 15 banks and 30 possible liquidity choices each, proceeding by brute force would require computing the costs for $30^{15}/15! = 1.0973 \times 10^{10}$ different liquidity profiles (although the task could be simplified by the aggregation property). The following result, proven in Appendix, greatly simplifies the analysis:

**Theorem 2.** System-wide costs are minimized assigning the same initial balance to each bank.

This drastically reduces the range of choices to consider; in the above example, the central planner has to choose between 30 equi-liquidity allocations.

---

$^{14}$ Recall that a bank's total costs are of the type $C(\lambda) = \lambda l + \kappa D(l, l_i)$.

$^{15}$ This is just one of the possible influence channels of liquidity regulation. Depending on the form this takes, regulation may substantially change the ‘rules of the game’ in payment systems and in other areas of banking operations.

$^{16}$ Strictly speaking, only an infinite amount of liquidity would ensure this, as the Poisson assumption allows for any number of payment orders to arrive within the day. However, for all practical purposes, posting one fourth of the expected number of daily payments (we have 3000 expected payments system wide, and 15 banks) brings expected delays to zero.

$^{17}$ The dots in Fig. 4 are directly computed equilibrium values. The continuous red line is a fitted function of the type $y = ax^{-b} + c$. The fit yields $R^2 = 0.997$. 

---
The blue line in Fig. 4 shows the liquidity choices of the central planner in a system with 15 banks. This is consistently above the equilibrium values, apart from when the price of liquidity is extremely high, in which case both planner and equilibria have \( l = 0 \) (not shown in the picture). So, independent banks inefficiently under-provide liquidity. The size of such inefficiency (measured either as the ratio between the blue and red line, or as ratio of the corresponding costs) first grows with the price of liquidity, then falls. We do not attempt a full calibration study. However, knowing that CHAPS typical netting ratio (i.e. the ratio liquidity/payments) is around 6–10%, we can infer a realistic range for the (standardized) price of liquidity \( \lambda / k \). In this range, banks appear to provide 3% to 10% less liquidity than the planner, suffering costs in the range 35–75 basis points. These figures are to be taken with caution, given that the model is a simplified one (above all, we suppose the system is symmetric and payments have unit size), but they suggest that sizeable gains could be reaped if banks were to post more liquidity.

3.2. Relative efficiency of different-sized networks

Is a system with more participants preferable to a smaller one? This question can be considered from different points of view: from a risk/financial stability perspective, or from an efficiency perspective. Here, we concentrate on the second aspect.

We then run experiments varying two dimensions of the model: the number of banks \( N \), and the number of payments. In a first experiment we vary \( N \), keeping constant the number of payments per bank. In a second experiment we vary \( N \) again, keeping now fixed the total number of payments of the system. In a third experiment we fix \( N \) and vary the number of payments. In all cases we assign equal volumes to each bank to retain symmetry and simplify the analysis.

The first scenario can be seen as a stylized comparison of different countries. The second scenario corresponds to the case of a system becoming more or less ‘tiered’. In countries like the UK (but not only), only the largest banks are directly connected to the interbank payment system. All other banks make their payments using these banks as intermediaries. This pyramidal organization, called ‘tiering’, is believed to carry some risks, so a key policy question is whether broader access to payment systems should be encouraged. Our second experiment thus contributes to this debate. The last experiment looks at how a system may react to e.g. a surge in payment activity.

In all experiments efficiency is measured as: (a) the amount of liquidity used up per payment, (b) the average delay of each payment, (c) the cost of each payment and (d) the ‘efficiency gap’, defined as the ratio between the equilibrium cost, and the socially optimal cost (i.e. the cost obtained by a ‘benevolent planner’).

---

18 For example, fewer participants could imply that the failure of one bank implies disruption of a larger share of payments. On the other hand, fewer participants might also mean safer participants, making it non-trivial to draw financial stability conclusions.

19 A slightly technical point: if a bank joins a payment system, instead of acting via an intermediary, the number of payments submitted to the system typically increases, instead of remaining constant. This is because all the between the intermediary and the intermediated bank, or between two clients of the same intermediary, were previously settled on the book of the intermediary (‘on us’), while now are settled via the payment system. This is often referred to as ‘dis-internalization’ of payments. We find that larger systems are less efficient than small ones, even without taking into account dis-internalization. Hence, adjusting the model to more faithfully reproduce a de-tiering scenario would only strengthen our findings.
3.2.1. Size effects I—constant individual bank’s volume

Here we vary $N$, keeping the number of payments per bank constant. Fig. 5 shows the delays of a single bank, for various $N$s (each surface is for a different $N$). When $N$ is increased, liquidity becomes less effective in reducing delays. Indeed, for any combination of ‘my’ and ‘others’ liquidity, a bank in a larger system experience higher delays.

The following is an intuitive explanation of this fact (as we said, liquidity flows are too difficult to be described analytically, so we can only rely on intuition). Consider a bank $i$ and suppose $N$ is increased from, say, 2 to 3. Because the number of payments per bank is kept constant and equally distributed over all banks, both outgoing and incoming expected payments remain constant for $i$ in any time interval. However, the variance of $i$’s incoming payments increases: at each $t$, a bank $i$ can now receive 0, 1 or 2 payments instead of only 0 or 1. As the relationship between liquidity balances and costs is convex, increasing the variance of the balance increases the expected value of delays.

But what happens in equilibrium? On the one hand, larger systems appear to use up more liquidity per payment (Fig. 6), and to give rise to higher costs per payment (Fig. 7).

On the other hand, the simulations also reveal that the ‘efficiency gap’ is wider for smaller systems. A corollary is then that regulation would be most beneficial for small systems, because it is these latter which deviate most from the social optimum. The reason for this is that the externalities produced by a single bank are stronger in small systems—as then a single bank has a bigger role in the whole system.

3.2.2. Size effects II—constant total volume

In this second experiment we keep total system volume constant, distributing it over a varying number of banks (in equal shares to maintain symmetry).

The results are rather clear cut: as shown by Figs. 8 and 9, systems with fewer members are more efficient in terms of liquidity absorption, delays and costs (even though, as in the previous case, the efficiency gap turns out to be larger for smaller systems). Using the same logic as in p. 13 Section 3.1, we can narrow down the range of relevant liquidity prices. We thus find that, if CHAPS were to double the number of its members, delays would increase significantly (perhaps by around one fifth) despite the fact that liquidity would increase (by about half). As a consequence, costs would also increase (by around half, according to our model). As we do not perform a precise calibration study (for example we assume symmetry in the system, which is clearly not the case in CHAPS), these results should be taken as suggestive of orders of magnitude only.

Our explanation of why smaller systems are more efficient is the following. A reduction of the number of banks from $N$ to say $N'$ can be seen as taking place in two ‘steps’: first, a re-assignment of the payments as to involve $N'$ banks only; second the elimination of the banks left with no payments. The second stage is neutral on costs and delays, as the eliminated banks are

---

20 If $Z$ is the number of a bank’s outgoing payments, the total outflow out of all $j \neq i$ is $(N - 1)Z.$ By construction, $i$ captures a fraction $1/(N - 1)$ of this flow, i.e. $Z,$ which is constant by construction.
‘dummy’. Instead, the first stage brings about liquidity savings, due to the so-called liquidity ‘pooling effect’. Pooling generates economies of scale in payment activity; it is discussed more at length in Appendix.

3.2.3. Size effects III—constant N, varying volumes

How does the system react to a surge in volumes? Our last series of experiments addresses this question, keeping N constant and varying the number of payments for each bank.
As volumes increase the system becomes congested: for any given combination ‘my liquidity’/‘others’ liquidity’, delays increase with volumes, showing a similar pattern as in Fig. 5 above.

Such ‘congestion’ is a mechanic, rather predictable fact. Coming to the equilibria, the simulations show more surprising results. We find that, as volumes increase, the system may become more or less liquidity-efficient, depending on the price of liquidity. When liquidity is cheap, an increase in volumes increases liquidity efficiency, while the opposite is true

Fig. 8. Liquidity demand and system size (II).

Fig. 9. Equilibrium costs and system size (II).
when liquidity is very expensive. Also, an increase in volumes turns out to accelerate settlement when liquidity is cheap, and to slow it when liquidity is expensive. Overall, equilibrium cost per payment fall with volumes, at all liquidity prices. However, the inefficiency gap rises with volumes—at least in the relevant range of liquidity prices described in the previous sections.

It is difficult to single out a ‘reason’ for these results because the equilibrium performance of the system is determined by only indirectly by the system’s mechanics, through the banks’ choices. Understanding how the payoffs of a game changes may be easy; understanding how the equilibrium changes is typically much more difficult.

4. Conclusions

In this paper we build and simulate an agent-based model of an RTGS system, paying special attention to the complex liquidity flows exchanged by the participating banks. We show that a complete, symmetric RTGS system can be described as an aggregation game, whose convenient features allow to compute the equilibrium behaviour of the system, and to perform various comparative statics exercises.

Our first result is a liquidity demand function, relating the liquidity demand of the system to the price of liquidity (standardized by the cost of delays). We find that this demand follows a power function.

Then, we consider the question of whether such liquidity demand is efficient. We find that cost-efficiency could be enhanced if banks were to commit more liquidity than they do in equilibrium. Calibrating our model, we find that this inefficiency is not extreme, but may be rather sizeable. This might constitute a rationale for imposing measures that encourage liquidity provision.

Thirdly, we look at whether systems with a large number of banks are preferable to systems with few banks. When a given total volume is concentrated in a small number of banks, the system is more efficient in a number of respects: in terms of liquidity per payment, of average delays, and of cost per payment; we attribute this to the emergence of ‘liquidity pooling’, described by previous studies.

Finally, we look at the impact of volumes. Perhaps surprisingly, a surge in volumes brings about an acceleration in settlement, and a reduction in total costs per payment. This proves the existence of strong cost-economies of scale in payment activity. We underline that this is not a mechanical fact, but is obtained taking into account the banks’ strategic behaviour.

We wanted to explore liquidity choices in a model which featured realistic, complex liquidity flows. This barred the possibility to reach full analytical results. However, extracting information from the simulations we could establish analytical properties of the settlement process, which provided the foundation of a rigorous agent-based modelization (with agent learning the equilibrium via a ‘fictitious play’ belief adjustment).

Appendix A

A.1. Fictitious play

Consider a sequence of daily games (settlement days) running from t=0 to potentially infinity. The actions chosen on day t are a vector \( p_t = (p_{t1}, p_{t2}, \ldots, p_{tl}) \); Fictitious play assumes that, over the sequence of days, every \( i \) forms a belief of what others will play next, choosing \( l_i^t \) as a best reply to such belief. That is:

- \( l_i^t \) is belief at time \( t \) is a vector \( p_i^t(\cdot) = (p_i^t(1), p_i^t(2), \ldots) \), where \( p_i^t(x) \) is the probability that \( i \) attaches to \( \sum_{j \neq i} l_j = x \) being played at \( t \).
- a bank updates its belief according to the following rule:
  \[
  p_i^t(k) = \frac{1 + \sum_{s=1}^{t-1} l_i^s} {t + A}
  \]
  where \( A = NL \) (\( N \) being the number of banks, \( L \) the maximum liquidity each can post), and \( l_i^t(s) = 1 \) if \( \sum_{j \neq i} l_j = k \), and zero otherwise.
- at \( t \), bank \( i \) chooses \( l_i^t = \text{argmin}_l \sum_{x=1}^{L} C_i(l|x)p_i^t(x) \) where \( C_i(l|x) \) is the cost incurred by \( i \) playing \( l \), if the others play \( \sum_{j \neq i} l_j = x \).

---

21 Here \( l_i^t(t) \) denotes the action \( l_i^t(0) \) chosen at time zero in day \( t \). We are not interested in the intra-day timing now, but rather in sequence of days, so we slightly change notation.

22 On the first day (\( t=0 \)), all banks believe that each \( \sum_{j \neq i} l_j \) is equally likely: \( p_i^0(k) = 1 / A \). Then, the more frequently a \( \sum_{j \neq i} l_j \) is played, the more frequently it is “believed” to be played again.
A.2. Property 1

The first property of delays is

\[ D(l_1, l_2, \ldots, l_N) = D\left(l_i, \sum_{j \neq i} l_j, N\right). \]

That is: a bank's delays are a function of its own actions, and of the sum of others' actions. Of course, \( D \) also depends on \( N \): changing the number of banks does change the 'game' and its outcome.

The equality in Property 1 is only approximate, but such approximation is very good. We show this by both theoretical and empirical arguments.

A.2.1. Theoretical explanation of Property 1

The 'explanation' behind Property 1 is the following. Each unit of liquidity moves from any \( i, j \) with a given probability. So, within a day, the distribution of liquidity in the system can be described by a Markov process, which converges to a stable distribution. As we have a large number of payments in the day, liquidity is quickly redistributed, flushing out the initial conditions. So delays are essentially determined by the total liquidity in the system or, to a better approximation, by (a) the total liquidity and (b) 'my' liquidity—which is particularly important in the 'transition phase' towards the stable distribution.

It is then clear that, the faster intraday liquidity reaches its stable distribution, the better Property 1 holds. So, how fast is convergence? We remind that, in a Markov process, the convergence to the stationary distribution \( \pi \) is given by \( \lambda \), the second eigenvalue of the transition matrix \( M: \pi(t) - \pi = O(\lambda^t) \) as \( t \to \infty \).\(^{23}\) In our case (symmetric system) \( \lambda = 1/(N-1) \), which means that, after each liquidity unit is circulated \( t \) times, the (expected) distance between \( l \) and \( \pi \) can be measured by \( 1/(N-1)^{\frac{1}{2}} \). We calibrate the simulations so the ratio payments/liquidity is in the region of 20 (to give a turnover ratio of 5\% broadly in line with e.g. the UK CHAPS). This means that, by one fifth of the day the approximation factor is \( (1/(N-1)^{\frac{1}{2}})^{0.2} \). This varies between about 0.004 for \( N=5 \) and 0.00003 for \( N=15 \). So, by one fifth of the day, liquidity is already very close to its stable intraday distribution, i.e. initial conditions are already negligible.

What is the role of symmetry here? It can be shown that a symmetric \( M \) has smallest second eigenvalue among all possible \( M \)s, so convergence is fastest \((\lambda \) vanishes faster as \( t \to \infty \). That is, convergence is slower with asymmetric \( M \)s, and Property 1 holds less well. However, for the parameters we consider \((N \) and turnover) convergence is still 'fast'. Indeed, if we set \( N=13 \) and pick \( M \) randomly (e.g. drawing \( M_{ij} \) from a uniform distribution, and normalizing it to make \( M \) a transition matrix), \( \lambda \) is typically low: it exceeds \( 1/2 \) in less than \( 0.1\% \) of the cases. This means that, in more than \( 99.9\% \) of the possible \( M \)s, the approximation factor is already \((\lambda^{0.2})=0.062 \) by one fifth of the day. This suggests that, even without symmetry, initial conditions are washed out quickly, so Property 1 holds well.

A.2.2. Empirical check of Property 1

Here we check Property 1 by means of simulations. To do so, we consider a generic bank \( i \). We then take an amount of liquidity, \( A \), and we distribute it across the \( j \) \((j \neq 1)\) in twenty different ways.\(^{24}\) These 20 assignments range from \( l_j = 1/AV \forall j \neq i \) to \( l_j = A, l_i = 0 \forall j \neq k, l_i \), spanning the full spectrum of variance in liquidity distribution (from zero variance i.e. equi-distribution, to maximum variance, or concentration at just one bank). For each assignment, and for each liquidity choice by \( i \), we run several simulations to compute the corresponding expected delays for bank \( i \). We then perform a simple analysis of the variance, to assess if knowledge of the particular liquidity assignment conveys any information about expected delays. The results are shown in Fig. 10 and Table 1.

Fig. 10 shows that there is no systematic relationship between bank's \( i \)’s delays and the variance of other's liquidity distribution (the 20 allocations are on the right-hand, horizontal axis, ranked in by decreasing variance). The variable 'own liquidity' (left-hand horizontal axis, 50 liquidity levels) is far more explanatory. Table 1 shows this reporting the results of the analysis of variance. The percentage of variance of 'Delays' explained by the precise allocation of others' liquidity is on average\(^{25} \) \( 3\% \) of the total variance. And, this is true for very different values of \( A \). That is, the distribution of others' liquidity appears to be irrelevant at any liquidity level.

This leads us to conclude that the precise distribution of others' initial liquidity does not matter for a bank's delays. In other words, a bank's delays are more affected by ordinary 'daily randomness', than by the precise distribution of liquidity across counterparties.

---

\(^{23}\) In a Markov process, \( \pi(t) = (0)M^t \). By the spectral theorem for stochastic matrices, one can write \( M = VDV^{-1} \), where \( D \) is a diagonal matrix containing \( M \)'s eigenvalues (standardized so \( D_{11} = 1 \) and \( D_{ii} < 1 \) \( \forall k \)), and \( V \) is a matrix whose columns are the corresponding eigenvectors. So \( \pi(t) = VDV^{-1} \). While \( D_{11} = 1 \) \( \forall t \), all other entries of \( D \) converge to zero as \( t \to \infty \), so that \( \pi(t) - \pi \). The upper bound for the speed of convergence is thus given by \( D_{22} \), the second eigenvalue (which being the second largest falls to zero less quickly).

\(^{24}\) We repeat the experiment for different values of \( A \), obtaining essentially the same results.

\(^{25}\) Across choices of 'own' liquidity.
A.3. Property 2 (curvature properties)

The second property of the delays is that $D(l_i, l_i/C_0)$ is convex. This comes out rather clearly from the simulations (see Fig. 1), but can also be shown analytically. Beside convexity, we also need another result about the curvature of $D$, to then prove Theorem 2. Convexity and this extra result are the content of the following Lemma (whose overall role had been illustrated in Fig. 2):

**Lemma 1.** (i) $D(l_i, l_i/C_0)$ is convex; (ii) $0 < \partial^2 D/\partial l_i \partial l_i < \partial^2 D/\partial l_i^2$.

**Proof.** To prove (i), we show that the Hessian of $D$ is positive semidefinite i.e. has a positive determinant. To this, we show that: (a) $d^2 D/dl_i^2 \geq 0$; (b) $d^2 D/dl_i^2 \geq 0$ and (c) $(d^2 D/dl_i^2)(d^2 D/dl_i^2) \geq (d^2 D/dl_i dl_i)$. For a given bank, $y(t)$ is the amount of payments received up to time $t$. Call $\bar{x}(t)$ the amount of payment orders received up to $t$. Define $\beta(t) = l_i + y(t) - \bar{x}(t)$. If positive, $\beta(t)$ equals the balance of the bank at $t$. If negative, $\beta(t)$ is the queue of the bank at $t$. Hence, total delays at the end of the day $T$ are given by the integral of the negative part of $\beta$. This is illustrated in Fig. 11—where the delays are the red area.

Consider now an (infinitesimal) increase in $l_i$. This amounts to an infinitesimal upward shift of $\beta(t)$, and therefore a reduction in delays (red area). As Fig. 11 clearly suggests, delays are reduced by the (measure of) the boundary set $\delta = \{t : \beta(t) \leq 0\}$ (dark line). In turn, as $\beta(t)$ is shifted up, this set $\delta$ ‘shrinks’, so the reduction in delays gets smaller. That is: the derivative of the delays gets larger. Formally: delays are given by

$$D(l_i) = \int_0^T 1_{\beta(t) \leq 0} \beta(t) \, dt$$

where $1_{x \leq 0}$ is the indicator function (equal to 1 when the subscript condition is satisfied, zero otherwise). Deriving we obtain

$$\frac{d}{dl_i} D(l_i) = \frac{d}{dl_i} \int_0^T 1_{\beta(t) \leq 0} \beta(t) \, dt = -\mu(\delta)$$

where $\mu(\delta)$ is the measure of $\delta$. As $\beta(t)$ increases in $l_i$, we have $\delta(l) \leq \delta(t)$ for $l \geq t$ (as $\beta$ ‘goes up’, $\delta$ ‘shrinks’). Which means $(d - \mu(\delta))/dl_i = d^2 D/dl_i^2 \geq 0$, proving (a).
Essentially the same argument also shows (b): the only difference is that now it is no longer $d\beta(t)/dl_i = 1\forall t$: an increase in $l_{-i}$ 'shifts' $\beta$ up only from a certain $t > 0$ (from the time when $i$ receives that extra amount of liquidity it would have not received otherwise). So $d\Omega/dl_i = \mu(\delta')$ for some $\delta' < \delta$. But the rest of the argument is unchanged. In particular, $\delta'$ too shrinks in $l_i$ and $l_{-i}$ so (b) is proven.

Before showing (c), we prove the second part of the Lemma, i.e. statement (ii). That both second derivatives are positive was shown above. The rest of the statement, $\partial^2 D/\partial l_{-i}^2 < \partial^2 D/\partial l_i^2$, can be written $d/dl_{-i}\mu(\delta) < d/dl_i(\mu(\delta))$ using the definitions introduced above. Fig. 12 suggests that $(d/dl_i)\mu(\delta) = \sum_{k,B|k} = 0 \text{ abs}(1/dB(\cdot)/dt|_{t=k})$ and similarly $(d/dl_{-i})\mu(\delta) = \sum_{k,B|k} = 0 \text{ abs}(1/dB(\cdot)/dt|_{t=k})$. The second sum includes fewer terms than the first. Hence it is smaller, so (iii) is proven.

To finally show (c), first observe that $(\partial^2 D/\partial l_i^2) < (\partial^2 D/\partial l_{-i}^2)$ for all $t \geq t'$. Thus, it shrinks the sets $\delta$ and $\delta'$ by the same amount (by definition $\delta' = (t: \beta(t) \leq 0)$ and $\delta = (t: \beta(t) \leq 0 \land t \geq t')$). That is, $(d/dl_{-i})\mu(\delta) = (d/dl_i)\mu(\delta')$ which is $(\partial^2 D/\partial l_i^2) = (\partial^2 D/\partial l_{-i}^2)$. Using (ii) we obtain $(\partial^2 D/\partial l_i^2) = (\partial^2 D/\partial l_{-i}^2) < (\partial^2 D/\partial l_i^2) = (\partial^2 D/\partial l_{-i}^2)$. Using (iii) we obtain $(\partial^2 D/\partial l_i^2) = (\partial^2 D/\partial l_{-i}^2)$ which is statement (c). □

From this proof, it is rather clear that these curvature properties have nothing to do with the symmetry assumption of our model.

A.4. Theorem 1 (uniqueness of equilibrium liquidity)

Theorem 1 in Section 2.4 states:

All equilibria feature the same total liquidity.

This result is central, as it allows to speak about the equilibrium liquidity, even though the game may possess many different equilibria. Hence, it makes it possible to perform comparative statics exercises. To prove the theorem, recall the following notation:

- $D(l_i, l_{-i})$ is the amount of delays suffered by $i$ at strategy profile $(l_i, l_{-i})$, and total payoff is $C = \alpha l_i + \kappa D(l_i, l_{-i})$.

Given Property 1 (aggregation), $C$ is a function of two variables. So, from now on by $l_{-i}$ we mean a scalar, $l_{-i} = \sum_{j \neq i} l_j$, instead of a strategy vector. We need some notation:

- $\Gamma_i(l_{-i})$ is bank $i$'s best reply to $l_{-i}$.
- $A_i = l'_i - l_i$, to be used when $l_{-i}$ and $l_{-i}$ are clear from the context. Similarly, $A_{-i} = \sum_{j \neq i} (l'_j - l_j)$ and $A = \sum_{i \in N} (l'_i - l_i)$.
The argument proceeds in two steps.

Step (1) For each \( l_i \) and \( l_{i'} \), we have \( |A''_i| \leq |A_{i'}| \) and \( A''_i A_{i'} < 0 \).

By Lemma 1, a bank’s cost function is convex in own liquidity. Hence, the optimal liquidity \( l_{i*}^* \) satisfies the first order condition \( \partial D / \partial l_{i*} = -\lambda / k \), and the standard envelope theorem applies

\[
\frac{dl_{i*}^*}{dt} = \frac{\partial^2 D}{\partial l_{i*} \partial t} \frac{\partial D}{\partial t}
\]

From Lemma 1, (iv) we obtain \(-1 \leq \frac{dl_{i*}^*}{dt} \leq 0\) and so, integrating, the statement of Step (1).

Step (2) If \( l^* \) and \( l^* \) are equilibria, then \( \Delta l^* = \sum l_i^* - \sum l_i^* = 0 \).

To reach a contradiction, suppose that \( l \neq l^* \) are equilibria but \( \sum l_i' > \sum l_i \), i.e., \( \Delta > 0 \). If it were so, there should be a non-empty set of banks \( S : A_k > 0 \) for all \( k \in S \). By Step (1), for each such \( k \) we can write \( A_k = -(A_{-k} + \varepsilon_k) \) with \( \varepsilon_k \leq 0 \) (\( l \) and \( l^* \) are equilibria, so we omit star superscripts). The total change in liquidity between the equilibria is therefore

\[
\Delta = \left[ \sum_{k \in S} A_k \right] + \left[ \sum_{k \in N \setminus S} A_k \right] = - \left[ \sum_{k \in S} (A_{-k} + \varepsilon_k) \right] + \left[ \sum_{k \in N \setminus S} A_k \right]
\]

Now, given a set \( N = \{x_1, x_2, x_3, \ldots \} \) it is clear that \( \sum_{i \in N} x_i = (|N| - 1) \sum_{i \in N} x_i \). If we then partition \( N \) into two sets, \( S \) and \( N \setminus S \), we can also write \( \sum_{i \in N \setminus S} x_i = (|S| - 1) \sum_{i \in S} x_i + |S| \sum_{i \in S} x_i \). So the above expression can be written as

\[
\Delta = - \left[ (|S| - 1) \sum_{k \in S} A_k + |S| \sum_{k \in N \setminus S} A_k - \sum_{k \in S} \varepsilon_k \right] + \left[ \sum_{k \in N \setminus S} A_k \right] = - \left[ |S| \left( \sum_{k \in S} A_k + \sum_{k \in N \setminus S} A_k \right) - \sum_{k \in S} \varepsilon_k \right] + \left[ \sum_{k \in N \setminus S} A_k \right] = (1 - |S|) \Delta + \varepsilon \rightarrow \Delta = \frac{\varepsilon}{|S|}
\]

But \( \varepsilon \leq 0 \), so this contradicts \( \Delta > 0 \). \( \square \)

A.5. Theorem 2 (efficient liquidity choices)

Theorem 2 in Section 3.1 states:

System-wide costs are minimized assigning the same initial balance to each bank.

Proof. Follows from the convexity of the cost function, and from the aggregation property. Indeed, let \( \mathbf{x} = (l_1, \ldots, l_n) \) be the vector of liquidity choices which minimizes total costs, with \( \sum l_i = l \); this gives cost \( C(l) \) to each \( i \). To reach a contradiction, suppose without loss of generality \( l_1 \neq l_2 \). We now show that allocation \( \mathbf{l} = ((l_1 + l_2)/2, (l_1 + l_2)/2, l_3, \ldots, l_n) \), obtained redistributing liquidity between 1 and 2, decreases overall costs. First, the costs for all \( i > 2 \) are unchanged—this is
immediate from the aggregation property. Second, the sum of 1 and 2’s costs is decreased. Indeed, under ℓ the costs of 1 and 2 are
\[ C(ℓ) = C(ℓ) = C_0(1) + C_0(2) = \frac{(l_1 + l_2) + (l_1 + l_2)}{2}, \]
but \((l_1 + l_2)/2, L-(l_1 + l_2)/2\) is evidently a linear combination of \((l_1, l_{-1})\) and \((l_2, l_{-2})\), so by convexity of \(C\) we have
\[ 2C((l_1 + l_2)/2, L-(l_1 + l_2)/2) \le C(l_1, l_{-1}) + C(l_2, l_{-2}) \]
that is, \(C(ℓ) + C_0(ℓ) \le C(l_1, l_{-1}) + C(l_2, l_{-2})\). Allocation ℓ leaves unchanged the costs for \(i > 2\), but reduces costs for 1 and 2—a contradiction. □

A.6. Equilibria uncovered by fictitious play simulations

Most of the equilibria found with the simulations have banks switching between two or more actions, depending on the evolution of their beliefs. This is due to the fact that, in the simulations, liquidity choices are discrete, so game may have mixed-strategy equilibria (with continuous strategy spaces and continuous payoff functions we would have pure equilibria). For example, at the lowest delay price level banks oscillate between \(i=0\) and 1, chosen with probabilities 8.5\% and 91.4\%, respectively. As banks become sufficiently confident that other banks chose \(i=1\) each, the best reply is \(i=0\). As the probability of others choosing \(i=0\) is thereby increased, banks switch back to \(i=1\). In this case, the game is a classic “hawk–dove” game. If nobody commits any liquidity, all experience very high delays as no payments can be settled. If everyone commits one unit of liquidity, payment settlement can take place. From an individual banks perspective, however, a better outcome would be not to commit any liquidity while others do.

As the cost for delays is increased, the probability of banks committing no liquidity is reduced gradually until, at delay price of one, a pure equilibrium emerges, where each bank chooses \(i=1\). At higher cost levels banks either reach a pure equilibrium, or a mixed equilibrium where they mix between a narrow range of different liquidity levels.

A.7. Pooling effect

The ‘payment pooling effect’ is a liquidity- (delay-) saving effect appearing when different banks merge their payment activity into one. Along with ‘payment internalization’,\(^{26}\) pooling generates economies of scale in payment activity.

Jackson and Manning (2007) observe the pooling effect in an empirical study, describing it as the reduction in the (per-payment) liquidity needed to settle without delays. Galbiati and Gianstone (2010) formally describe the pooling effect, by showing that the minimum of a random walk (the liquidity needed to execute payments with no delay) is a concave function of the length of the walk itself (the number of payments to be made and received). So, increasing the scale of activity of a bank, liquidity needs increase less than proportionally.

The pooling effect can be stated in a ‘dual’ formulation: given a level of liquidity, the delays of a bank increase less than proportionally with the volume settled. This we find by simply comparing the simulated delay function when each bank’s volumes vary.

References


\(^{26}\) Which is due to the fact that, when be settled on the book of a correspondent bank, payments are ‘cancelled’ from the payment system and thus need no liquidity to settle.