BANK OF ENGLAND

Working Paper No. 399

Liquidity costs and tiering in large-value payment systems

Mark Adams, Marco Galbiati and Simone Giansante

July 2010

BANK OF ENGLAND

Working Paper No. 399
Liquidity costs and tiering in large-value payment systems

Mark Adams,[1] Marco Galbiati[2] and Simone Giansante[3]

Abstract

This paper develops and simulates a model of the emergence of networks in an interbank, RTGS payment system. A number of banks, faced with random streams of payment orders, choose whether to link directly to the payment system, or to use a correspondent bank. Settling payments directly on the system imposes liquidity costs which depend on the maximum liquidity overdraft incurred during the day. On the other hand, using a correspondent entails paying a flat fee, charged by the correspondent to recoup liquidity costs and to extract a profit. We specify a protocol whereby one bank in each period can revisit its choice whether to link directly to the system, or to become clients of other banks, thus generating a dynamic client-correspondent network. We simulate this protocol, observing the emergence of different network structures. The liquidity pricing regime chosen by a central bank is found to affect the tiering process and the network structures it produces. A calibration exercise on data from the UK CHAPS system suggests that the model is able to generate realistic predictions, ie a network topology similar to that observed in reality, driven solely by the underlying pattern of payments and the structure of liquidity costs.

Key words: Tiering, liquidity cost, large-value payment system, RTGS, network formation.

JEL classification: C7, G2.

(1) Bank of England. Email: mark.adams@bankofengland.co.uk
(2) Bank of England. Email: marco.galbiati@bankofengland.co.uk
(3) CCFEA, University of Essex. Email: sgians@essex.ac.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England. Besides an anonymous referee, and the Working Paper editor Gabriel Sterne, the authors thank Charles Kahn, John Jackson, Mark Manning and Rod Garratt for important insights. We owe useful feedback to participants to the following conferences, where the paper has been presented: Eastern Economic Association meeting, New York, February 2009; and the 14th conference on Computing in Economics and Finance, Paris, June 2008. We thank colleagues in the Financial Stability area of the Bank of England for comments and support. This paper was finalised on 7 May 2010.

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

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

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

ISSN 1749-9135 (on-line)

Contents

Summary

Interbank payment networks (ie the channels through which banks execute payments), differ widely across countries. In some countries, these networks have a 'star' shape: all (or most) banks are directly connected to a central node, a piece of infrastructure where all payments are executed. In other countries one instead observes 'tiered' structures: a few banks (*first-tier banks*) are directly connected to the central processor, while all other banks are connected to first-tier banks and channel payments through them. This paper studies the forces behind the formation of 'stars' versus 'trees' in payment networks; what it does *not* consider instead is the question of *which structure* is more desirable. This work has therefore a purely explanatory aim, rather than a normative one.

These forces stem from the nature of modern *large-value* payment systems (LVPSs).

Most LVPSs today work in real-time gross settlement (RTGS) mode, whereby each payment must be settled individually by transferring the corresponding value from payer to payee. The main advantage of RTGS is that it eliminates credit risk. However, as payments must be settled in *gross* amounts, the RTGS mode requires large amounts of liquidity – a shortcoming which can however be reduced by co-ordinating payments, so liquidity is 'recycled' between banks.

Another reason why central banks pushed for the adoption of RTGS is that in practice, although not by necessity, RTGS systems use central bank money as medium of settlement. That is, the funds used to settle payments are held in accounts at the central bank. This brings about two benefits: first, the safekeeper of these funds cannot default; second, the central bank is able to monitor and possibly regulate the payment activity.

However, in some countries (including the United Kingdom), many banks are *not* direct members of the national RTGS system, and their payments are *not* settled on the RTGS system. These are the 'tiered' systems mentioned above, where second-tier banks execute payments via *correspondents* in the first tier. Payments between correspondents (due to the correspondents' proprietary and/or client operations) settle on the official RTGS system. But payments between banks with a common correspondent are made on the books of the correspondent itself. *Internalised* by the correspondent banks, these payments thus do not transit across the RTGS system. As a consequence, they are neither subject to the RTGS rules, nor can they be easily monitored by the authorities.

Surveys of UK correspondent banks indicate that internalised payments are a significant fraction – around one third by value – of all interbank payments. The value of payments which correspondents make through the RTGS system on behalf of clients is also large. These latter payments may also create risks, as they are often not pre-funded. That is, correspondents often agree to make them by extending credit to the client. So, when present, tiering is an important feature of a payment system which may have an important bearing on the system's functioning, and on the risks therein.

As mentioned above, one shortcoming of RTGS systems is their potentially high liquidity need. Tiering can be seen as a spontaneous response to this, because a major effect of tiering is to reduce liquidity costs. This is for two reasons. First, internalised payments can be made without liquidity (the *internalisation effect*). Second, by pooling own and client payment flows, the

correspondents may face smoother, better manageable and therefore less costly liquidity needs (the *pooling effect*).

We build a model of tiering choices, with two 'inputs': the cost of liquidity, and an exogenous pattern of payment flows. Starting from these, we formally model the internalisation and liquidity pooling effects. We then show that even such a parsimonious model, when calibrated on real data, generates realistic payment networks. This ability to reproduce some stylised facts suggests that the cost of liquidity *is* an important driver of tiering. This is ultimately controlled by the central bank, so we conclude that a central bank has powerful policy levers to influence tiering patterns. However again: this paper sheds light on *how* these policy levers can affect tiering, but is silent on *how they should* be used to this aim. Such a judgement cannot be expressed here, because several consequences of tiering are not considered in this work. Above all, we disregard any 'risk' to individual institutions and to the system as a whole.

More precisely, our model features a fixed number of banks sending payments to each other. During a day, each bank receives a random stream of payment instructions at a constant rate. Each instruction requires payment of a single unit of currency to another bank. Intraday banks act mechanically: payments are executed as soon as payment instructions are received. Banks instead make decisions about where they want to sit in the 'payment network'. To be more precise, one bank is randomly picked in each period, and is given the choice between becoming a direct member of the RTGS system, or to arrange for a correspondent to execute their payments. If a bank joins the RTGS system, its payment activity generates liquidity costs. If instead it becomes a client of a correspondent, the client bank incurs no liquidity costs, but pays a fee to the correspondent for its service. The correspondent's payment activity changes as a result of taking on a client, and hence so does its liquidity cost. We specify a stylised but realistic 'protocol' for the negotiation of these fees.

By virtue of the internalisation and liquidity pooling effects, total liquidity costs for a correspondent and its customer together are no large than the sum of the standalone costs, thus giving incentives to tier. On the other hand, banks make their decisions sequentially and, depending on their payment activity, they may find it convenient to join different correspondents. Hence, more than one correspondent bank may coexist. After a possibly long (but finite) number of 'days', the system reaches a steady state where a non-trivial network of client-correspondent relationships is formed. We simulate this model, calibrating it to data on the UK CHAPS system, and we look at the resulting networks. As mentioned above, the model produces networks which reproduce some features of the real CHAPS client-correspondent network. We perform some comparative statics exercises, suggesting how the payment network would change, if the central bank changed the price of liquidity.

1 Introduction

Very large amounts of money flow through *large-value* payment systems (LVPSs). In 2008, interbank payments in the United Kingdom's CHAPS system averaged £274 billion (almost $500 billion) a day; the corresponding transactions in the US Fedwire system amount to about twice as much, while in the euro area's TARGET system volumes are roughly three times as large. Considering these amounts, it is natural that central banks and policymakers are interested in the smooth functioning of LVPSs, devoting substantial resources to their study, design and oversight.

These large aggregate flows are only part of the picture, as the structure of LVPSs differ drastically from country to country. In the United Kingdom for example, the main system has only 14 direct, 'first-tier' members, who settle payments on behalf of about 420 other institutions. At the other extreme, the US Fedwire system has a much less tiered structure: over 9,500 banks, some of which are very small, link to the system directly and settle payments on their own behalf. A number of recent studies have charted the topology of payments over these networks in detail: Arnold *et al* (2007) look at the US Fedwire system; Becher *et al* (2008) consider the UK CHAPS, Lublóy (2006) study the Hungarian VIBER, while Inaoka *et al* (2004) look at the Japanese payment system BOJ-NET.

What lies behind these differences in 'tiering', ie in the organisation in correspondent and client banks? Why do certain banks join a LVPS, while others who are eligible to join make their payments via a first-tier correspondent? These questions are important because, first, the network structure of a payment system may affect the stability and efficiency of the system itself. Second, tiering implies that a share of interbank payments does not cross the official, regulated LVPS at all, settling instead on the books of the first-tier banks.[1]

Here, however, we do not attempt to clarify which structure is most desirable from a central bank's perspective. This paper concentrates instead on the following questions: what determines the structure of a payment system? Can a central bank induce the formation of a particular network structure?

To answer these questions, one must consider the incentives to join the first-tier of a LVPS, versus those to remain in the second tier. Direct membership is expensive: first, it imposes fixed costs such as fees and back-office expenses to connect to the system. Second, and perhaps more importantly, a first-tier bank must at all times have sufficient liquidity to support its payment activity. Indeed, most LVPSs nowadays work in RTGS (real-time gross settlement) mode: if bank i owes £2 to bank j, and j owes £1 to i, both i and j must transfer the full amounts to their counterparty, with no netting allowed (with netting, all that would be due is a £1 payment from i to j).[2] This gross system imposes high liquidity demands on the banks, exposing them to the risk of large (albeit temporary) outflows of funds. Management of these flows represents one significant challenge for a first-tier bank.

[1] This share can be large: internalised payments are estimated to be about 30% for the United Kingdom, or over £90 billion daily. For a discussion of risks involved in tiered payment systems, see Harrison *et al* (2005).

[2] Until two decades ago, most LVPSs worked on an end-of-the-day-net basis. Gross systems were introduced worldwide to eliminate the credit exposures that would otherwise build throughout the day.

Jackson *et al* (2006) look at a bank's decision whether to become a direct member of a system, or to use a correspondent. Their findings suggest the existence of economies of scale in correspondent banking, generated by two effects: internalisation of payments and liquidity pooling. Internalisation refers to the fact that, when a bank acts as a correspondent, payments between its clients can be settled on the bank's own books ('on us'), at a zero liquidity cost. Liquidity pooling instead is a dynamic effect: by pooling uncorrelated payment requests from different clients, the liquidity need of a correspondent bank stabilises, implying in turn that the costs of liquidity management are lowered. The sizes of the internalisation and liquidity pooling effects in CHAPS are estimated by Lasaosa and Tudela (2008) in a study which relates the degree of tiering to the liquidity needs of the system.

This paper looks at how 'internalisation' and 'pooling' affect the client-correspondent structure of a LVPS. To do so, we set up a model where a number of banks face a random stream of payment requests. Banks can execute payments on their own, by borrowing liquidity from the central bank at a cost. Alternatively, they may become customers of other banks (correspondents), which can execute payments on their behalf, thus relieving them of liquidity costs. However, correspondents charge their clients a fee to recoup costs and, possibly, to make a profit. Who becomes a correspondent, and who instead remains in the 'second tier' attached to one correspondent or to another, is endogenously determined by a dynamic process. Making some assumptions on how correspondent services are priced, offered and accepted, we look at how the client-correspondent network evolves in time, and ultimately converges to a stable state.

Our model is highly stylised. First, we assume that the timing of payments is outside the banks' control: payments are made as soon as payment orders are received and, in turn, orders are generated by a random process whose intensity (but not the precise timing) depends on banks' choices.[3] In other words, we do not consider active liquidity management on the part of banks, an issue – considered by eg Angelini (1988), Bech and Garrat (2003). As a second important simplification, we ignore all credit risk issues that may emerge between correspondents and their clients. Our paper is therefore different from Chapman *et al* (2008), where instead credit risk is the main driver of tiering, which allows correspondent banks to assume a monitoring role. The interaction of credit risk and tiering are also studied by Harrison *et al* (2005), by Kahn and Roberds (2005) and Lai *et al* (2006). We leave this important issue aside, to focus on the relationship between *i*) the 'geography' of the underlying payments to be made, ii) the liquidity pricing regime chosen by the central bank, and iii) the ensuing network structure of the system.

In a nutshell, our findings are that, if the cost of liquidity is proportional to the amount borrowed, economies of scale in liquidity costs bring about concentration in the correspondent business, generating tiering. If instead convexities (ie dis-economies of scale) are present (as in the case of liquidity lent freely, but against collateral with low opportunity cost), tiering is reduced. In any case, the structure of the tiered network appears heavily influenced by the pattern of the underlying payments, or the 'geography' of the payment industry.

[3] More precisely: a bank's payments orders are generated by a Poisson process, whose parameter depends on the position of the bank in the payment network.

2 Model

The model has a population of N banks, sending payments to each other over a series of days. Banks can either be direct participants in the payment system, or they can hire a correspondent bank to execute payments on their behalf. If a bank participates directly, it needs to obtain liquidity from the central bank. This has a cost, which can be interpreted as either direct central bank charging for intraday overdrafts, or the opportunity cost of posting collateral at the central bank. Instead, banks that hire a correspondent only have to pay a price for the payment service. We look at how correspondent agreements evolve in time, leading to an equilibrium network structure of the payment system.

2.1 Intraday payments

Banks are indexed by i=1, 2,.., N. On any day, banks send to each other unit-size payments according to a 'payments matrix' $P = [p_{ij}]$. The entries p_{ij} are integers representing the number of payments that i sends to j. We assume that incoming and outgoing payments are balanced:

$$\forall i, \quad \sum_j p_{ij} = \sum_j p_{ji} \equiv \frac{1}{2}\lambda_i \qquad \textbf{(1)}$$

where λ_i is a parameter representing the intensity of the payment activity of bank i. In real payment systems, payments need not be balanced on every single day. However, balancedness is an acceptable assumption when modelling an 'average' day. Indeed, this is the simplest way to avoid that some banks are constantly in deficit, and thus eventually disappear.

Payment volumes, and hence values, are fixed by P. However, we suppose that the sequence in which payments are made is random. The only assumption we make here is that the liquidity need of bank i, defined as the sum of payments sent minus payments received up to t, can be described as a symmetric random walk:

$$L_i(t) = \sum_{s \leq t} x(s)$$

where $\{x(s)\}$ is a sequence of independent random variables equal to 1 or -1 with equal probability. The time elapsing between payments sent and received (ie the precise stochastic process generating payments) is unimportant here. For our purposes, two equivalent formulations would be that payments arrival is a Poisson process (with intensity λ_i so that the average volume settled by i is λ_i), or that time is discrete and $L_i(t)$ is a random walk of length λ_i. This equivalence is due to the way we define costs, described in the next Section.

2.2 Liquidity costs and the 'pooling effect'

We assume that, if a bank directly participates in the payment system, it pays a cost related to its daily maximum liquidity need. In the United Kingdom, for example, banks participating to the CHAPS system have to pre-fund their liquidity needs with the central bank. Central bank overdraft charges are some function f, of bank i's maximum liquidity costs during the day [$f(\max_t L_i(t))$, where t indexes time during the day]. Depending on the (random) order in which

payments are made and received, the maximum overdraft varies from day to day. However, banks are risk-neutral so only expected cost matters:

$$C(\lambda_i) = E\left[f\left(\max_t L_i(t)\right)\right] \qquad \textbf{(2)}$$

The appendix shows how $C(\lambda_i)$ can be computed. For certain specifications of the pricing function f, the expression becomes rather intricate. However, we will fix the functional form of f and will run simulations, so all we need is an expression amenable to numerical computation (the appendix shows how to obtain it).

An important fact to note is that $C(.)$ is uniquely determined by λ_i, which is in turn determined by P. The following examples show how C looks like, for an f that we use in the simulations later on.

Example 1

Type-I costs:

$$f(x) = cx, \qquad c > 0 \qquad \textbf{(3)}$$

These costs represent a situation where liquidity is lent against collateral, and this latter entails a constant unit cost. Computation shows that the resulting expected costs C(λ) are increasing, concave, and asymptotically linear, with C'(0)<1 - see Figure 1.

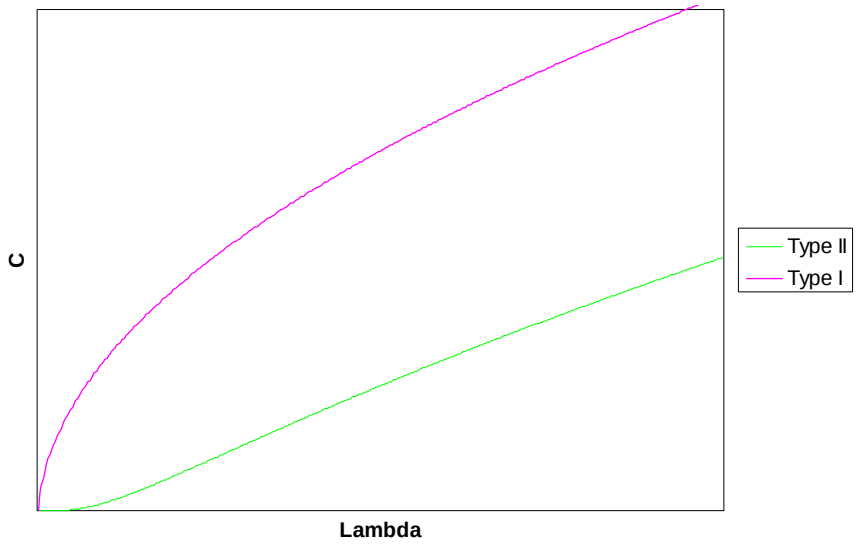
Type-II costs:

$$f(x) = \begin{cases} 0 & for \quad x < K \\ cx, \quad c > 0 & for \quad x \geq K \end{cases} \qquad \textbf{(4)}$$

where K is some positive constant. This case is a stylised representation of liquidity costs in the UK CHAPS system: liquidity is lent against collateral, which carries a constant unit cost. However, the first K units of collateral are free. This is because such K units must be held for prudential reasons (ie independently of payments), and can be used to fund payments.[4] When considering payment choices, the cost of such collateral is therefore sunk, and can be disregarded. Computation shows that the resulting $C(\lambda_S)$ is 'S-shaped' (first convex, then concave as in Case I for higher values) - see Figure 1.

[4] This is the so-called 'double duty' regime.

Figure 1: Liquidity costs in CHAPS-like systems



We use these cost types in the simulations later on (where we will clarify the reason to adopt these precise functional forms). Under both specifications, the concavity[5] of the function C is the so-called 'pooling effect': there are economies of scale in the payment activity.

It should be noted that we do not consider costs other than liquidity costs. In reality, banks have to bear other costs too – for example, IT costs, which generate mainly initial, fixed costs. We concentrate on liquidity costs for two reasons: (a) in reality such initial costs are probably small, compared to the costs of managing payments day to day; and (b) such initial, sunk costs trivially generate economies of scale, whose effect would simply be to create more tiering. But we want to 'factor this out' of our model, in order to concentrate on the effects of liquidity costs – whose effect *a priori* is less clear, and therefore more interesting, and which can be directly affected by policymakers.

2.3 Correspondent banks and 'internalisation effect'

Instead of participating directly in the payment system, a bank j may outsource its payments activity some other bank i. When this happens, bank i acts as the correspondent of bank j (which becomes client of bank i), and the following terms are agreed upon: i supports all liquidity costs deriving from j's payments, and in exchange j pays i a flat fee. A surplus is created by a correspondent agreement: first, some payments may be internalised. Second, there are economies of scale from pooling payments, as shown by Figure 1.

In more detail, suppose bank i is correspondent for a set of banks $S = \{i,j,...\}$. In this case, bank i's liquidity costs are determined by the payments between S and the banks outside S. Instead, payments within S can be settled by changing book entries and require no liquidity. Therefore bank i's cost will be equal to $C(\lambda_S)$, where

$$\lambda_S = \sum_{j \notin S} \sum_{i \in S} p_{ij} + \sum_{i \notin S} \sum_{j \in S} p_{ij} \qquad \textbf{(5)}$$

[5] For Type-II, only above a certain λ.

Note too that λ is subadditive: given two groups A and B, $\lambda_{A\cup B} \leq \lambda_A + \lambda_B$. Indeed, if bank A has no payments from/to B, no payments can be internalised, so $\lambda_{A\cup B} = \lambda_A + \lambda_B$. But, if all payments made and received by B are towards A, then $\lambda_{A\cup B} = \lambda_A - \lambda_B$. In intermediate cases, $\lambda_A < \lambda_{A\cup B} < \lambda_A + \lambda_B$. The subadditivity of λ is the so-called 'internalisation effect'.

Summing up, C is increasing but, due to internalisation, adding a bank to a group S can either increase or decrease the costs of S's correspondent. As noted above (Figure 1) C is convex in a certain range so, even if $\lambda_{S\cup k} > \lambda_S$, it may still be the case that $C(\lambda_{S\cup k}) < C(\lambda_S) + C(\lambda_k)$ – ie a surplus may be realised by adding k to S.

2.4 Network formation

A payments network is a partition of the N banks into distinct sets, each with one correspondent. How do these sets form, ie how does the network evolve? We imagine that one correspondent relationships is renegotiated each day, with banks accepting offers made according to the following protocol (with days indexed by $T = 0,1,2...$).

1. At $T = 0$, all banks are self-settling;
2. at each $T > 0$, one randomly selected bank (say i) receives an offer from each other bank k; this is the fee k would charge i to become its correspondent;
3. i chooses the best (lowest) offer, becoming client of the best offerer;
4. when a bank i becomes a client of another bank, all its clients (if any) go back to self-settling. i must pay a penalty to its previous clients for breaching their contracts.

To clarify, the selected bank i receives offers from all potential correspondents, including banks which are currently clients of another bank (a client k makes an offer to leave its correspondent, to become itself a correspondent for the new group $\{i,k\}$). Of course i may also maintain its role; it will do so when the expected costs of doing so are lower than any other offer.[6]

The offer, or fee charged by the correspondent, is determined according to the Nash Bargaining Rule (NBR). In general terms, the NBR prescribes that, if parties a and b obtain a total profit ω by signing an agreement, they divide it in two shares x_a and x_b as follows:

$$x_a = (1/2)(\omega + O_a - O_b)$$
$$x_b = (1/2)(\omega - O_a + O_b)$$

where O_i is bank i's outside option, ie what it receives if the agreement is not signed. In our story, party b (the client) pays a fee, so the offer is $-x_b$; the correspondent instead takes 'the remainder' of the profit. It should be noted that these offers are 'myopic': banks do not consider that their partners might sign other contracts in the future.

[6] It is simple to see that no bank ever finds it convenient to go back to self-settling.

Example 2

For simplicity, for a group A we write C(A) instead of $C(\lambda_A)$. Suppose that k, a self-settler with no clients, makes an offer to another similar self-settler i. If the offer is rejected, the parties' profit remain -C(k) and -C(i). If instead the offer is accepted, the total profit for both parties is -C({i,k}). The NBR attributes to i a profit (1/2)[-C({i,k})+C(k)-C(i)], ie i is asked to pay q_{ki} =(1/2)[C({i,k})-C(k)+C(i)]. This is the fee offer that k makes to i.

Consider now the general case: *i* receives an offer from *k*. There can be two sub-cases: *i* is a client of a bank (say *w*), or it is correspondent for group *S*.[7] In the first sub-case, *i*'s outside option is $O_i = -q_{wi}$. In the second sub-case, if *i* takes its outside option, it bears a cost $C(S)$ but receives fees totalling $\sum_{r \in S \setminus i} q_{ir}$. So,

$$O_i = \begin{cases} -q_{wi} & \text{when } i \text{ is client of } w \\ -C(S) + \sum_{r \in S \setminus i} q_{ir} & \text{when } i \text{ is correspondent for } S \end{cases} \qquad \textbf{(6)}$$

To determine the joint profits to *i* and *k*, recall that, if a correspondent *i* leaves its group, each of its clients goes back to self-settling. Hence, each 'abandoned' bank r suffers a loss of $C(r) - q_{ir}$. Bank *i* is liable for this, so its defection brings about a penalty equal to

$$X_i = \sum_{r \in S \setminus i} [C(r) - q_{ir}]$$

We suppose that *i* and its new correspondent *k* share this penalty. So when *k*, correspondent for group *P*, makes an offer to *i*, correspondent for group *S*, the profits for the new group are

$$\omega = -C(P \cup i) + \sum_{r \in P \setminus k} q_{ir} - X_i \qquad \textbf{(7)}$$

Hence, the NBR prescribes that *k* charges *i* a fee equal to:

$$q_{ki} = \frac{1}{2}[-\omega + O_k - O_i]$$

with ω defined in **(7)** and O_i defined in **(6)**.

2.5 Dynamic properties

The abstract structure of the model is that of a coalitional game: we have a set of players *N* and, for each subset $S \subseteq N$, a pay-off $C(\lambda_S)$ is defined (by equations **(5)** and **(2)**). To this coalition-form game, we attach a particular protocol (Section 2.4), specifying how coalitions form and dissolve. We do not pursue an abstract analysis of this game. However, the simulations show that the network reaches a stable state in a finite number of steps.[8] That is, no cycles are

[7] When *i* is a self-settler, $S = \{i\}$.

[8] This can be formally proven by noting that new client-correspondent links form *only if* this reduces costs for all other banks that may be involved (recall that a correspondent pays a penalty to its clients, if it rescinds its links with them). Hence, total liquidity cost is a

generated in our protocol, and an equilibrium is reached. What equilibrium networks are possible?

The star network (one bank acting as correspondent for all others), is trivially an equilibrium network.[9] However, it is easy to construct matrices P with two equilibria, both accessible from the same initial condition.[10] Because banks make decisions in a random order, one cannot speak of a unique equilibrium in general. An analytical study of the statistical properties of these equilibria is beyond the scope of this work. Instead, we run the protocol many times using different 'seeds', for each set of exogeneous inputs (a matrix P and a function f). The next section describes the results.

3 Results

In the simulations we fix the matrix of payments P (see Section 2.1) and the cost function f (see Section 2.2). Then, we 'run' the protocol to produce a sequence of networks, starting from a situation where all banks are self-settling, up to equilibrium ie until banks stop changing correspondents. Section 3.1 presents some abstract examples. Section 3.2 calibrates the model using data from the UK CHAPS payment system.

3.1 Liquidity costs, underlying payments and tiering

This section presents some abstract examples, to illustrate the relationship between (i) liquidity costs, (ii) the geography of payments given by matrix P, and (iii) the resulting network structure. We consider the two types of liquidity cost described in Examples 1 and 2, combining them with three payment matrices described in Figure 2.

Lyapunov function for the dynamic system: it falls monotonically, and reaches a minimum in a finite time (as there are only finitely many possible networks).

[9] A lone correspondent internalises all payments, and so it incurs zero costs; as a consequence, its fees cannot be undercut.

[10] If two groups have many within-group payments, and few cross-group payments, they are 'far' from each another, and they can coexist in equilibrium.

Figure 2: Stylised payments patterns

P': disconnected components

0	1	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	0	0	0
0	0	0	1	0	1	1	0	0	0
0	0	0	1	1	0	1	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	1	1	0

P'': complete asymmetric network

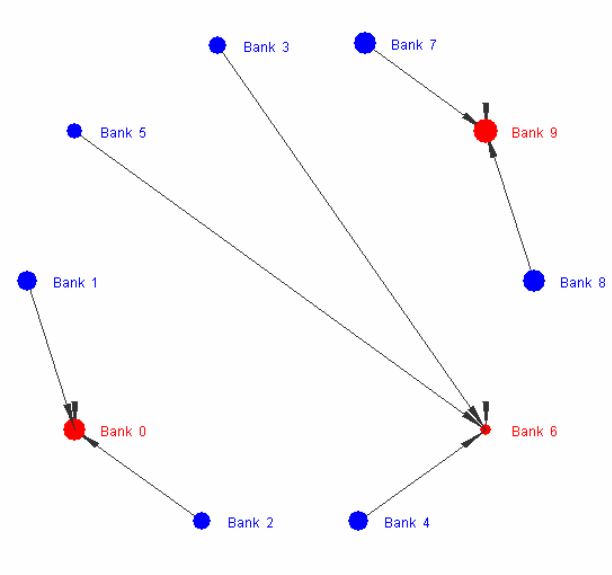
0	1	3	1	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
1	2	0	2	1	0	0	0	0	0
0	1	2	0	2	0	0	0	0	0
0	0	1	2	0	1	1	0	0	0
0	0	0	1	2	0	1	1	0	0
0	0	0	0	0	2	0	2	1	0
0	0	0	0	0	1	1	0	1	0
0	0	0	0	0	0	1	1	0	3
0	0	0	0	0	0	1	1	1	0

P''': complete asymmetric network

0	1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1
1	1	1	1	0	1	1	1	1	1
1	1	1	1	1	0	1	1	1	1
1	1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	1	0

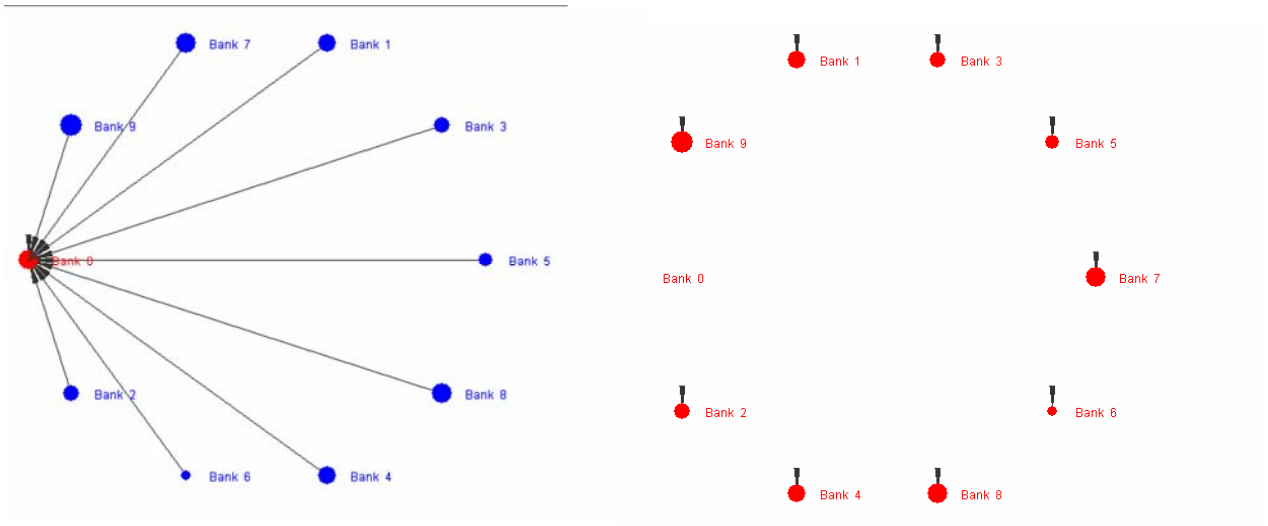
Matrix *P'* represents a disconnected payment network with three distinct payment areas: one comprising banks 1-3, a second comprising 4-7, and a third with the remaining 8-10. Matrix *P''* instead represents a complete, symmetric payment network. Finally, in *P'''* banks have preferential payment partners, but the payment network is completely connected (bank 1 in the yellow area makes payments to eg 3, in the orange area. And bank 6, in the blue area, sends and receives payments from eg bank 5 in the orange area. However, connections between the yellow and blue areas are weak).

Figure 3: Payments *P'*, Type-I and II costs



Matrix *P'* gives rise to the (rather predictable) outcome of Figure 3: one correspondent emerges for each of the three disconnected network components. Given symmetry, the identity of the banks which become correspondent is randomly determined.

Figure 4: Payments *P''*



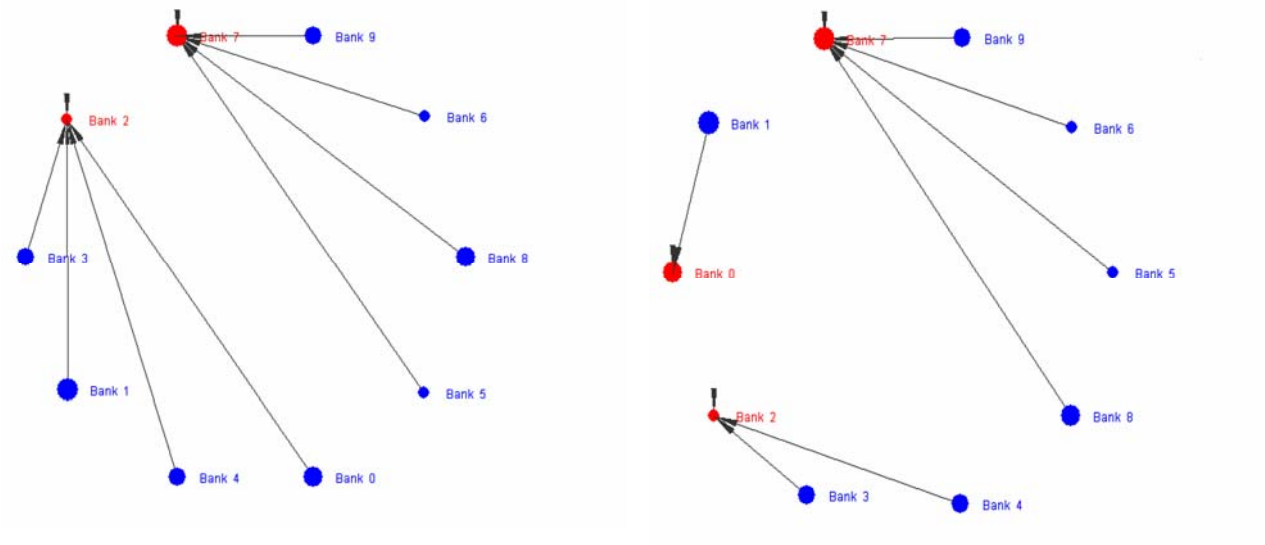
Type-I costs Type-II costs

Matrix *P''* gives rise to either of the two different networks in Figure 4: a 'star' or a disconnected network. Which outcome will arise is determined by the shape of the cost function. Type-I costs induce maximum tiering (star). Instead, under Type-II the network remains disconnected - provided the threshold K (equation **(4)**) falls in a certain interval. Indeed: suppose K exceeds the payments of any single bank, but smaller than the payments that two banks make to other banks (ie $\lambda_i < K < \lambda_{iUj}$). Then, a single bank can settle its own payments at zero cost,

while a group of two incurs a positive cost (recall the meaning of threshold K: liquidity is free up to K). That is, for small volumes there are decreasing returns to scale, and hence no incentives to aggregation. See also Figure 1: for low λs, expected costs increase slowly ie benefits from agglomeration are small.

The outcomes in Figure 4 are clearly not what we observe in reality. Encouragingly, using a more realistic matrix (P''') we obtain the more interesting results of Figure 5.

Figure 5: Payments *P'''*



Type-I costs

Type-II costs

Figure 5 shows that, as in the previous example, Type-II costs generate less tiering than Type-I costs. However, given that the banks of matrix P''' differ in their payments, some tiering occurs. Notably, the shape of client-correspondent network is influenced by the shape of the underlying payment network: three correspondents emerge in each of the three 'areas' of more intense exchange.

3.2 Calibrating the model with real data

In this section we calibrate the model using real data, to test the model's ability to yield realistic predictions.

For the matrix P we proceed as follows. The Bank of England has a complete record of all transactions taking place on CHAPS, which provides the backbone of matrix P. However, CHAPS transactions are by definition those between correspondent banks. Unfortunately, when a transaction is made on behalf of a client, the underlying payer and payee are not recorded in CHAPS. To reconstruct the full matrix we then use the Bank of England 2003 CHAPS traffic survey data set (also used in Becher *et al* (2008)). This survey samples five days of the payments executed on CHAPS, recording both the ultimate payer and payee banks, and the correspondents used. Crucially, the survey also asks correspondent banks to report the percentage of internalised payments. Allocating these payments between banks which use the

same correspondent, in proportion to their outgoing payments over CHAPS as revealed by the survey, we obtain a P to use in the simulations.

We want to run simulations under both Type-I and Type-II costs, the reason being that these two specifications represent two common regimes of liquidity pricing, on the part of central banks. Type-I applies to a situation where a bank pays proportionally to its liquidity usage. Type-II instead applies to systems where liquidity is given against collateral and the opportunity cost of posting collateral is (essentially) zero up to a certain point. A notable example of this is the UK CHAPS system. There, intraday liquidity can be obtained from the Bank of England at a zero interest rate, in exchange for collateral. But, a certain amount of collateral must be held anyway for prudential reasons. Hence, up to the amount of this 'sunk cost', UK banks may obtain liquidity essentially for free.

With Type-I costs the model requires no calibration beside P.[11] However, under Type-II costs we have to choose the threshold K (but only this).[12] We do so considering two different specifications: under the first, K is constant across banks - we call this specification 'absolute threshold'. Under the second, called 'relative threshold', K varies between banks: for bank i it is given by $\alpha\lambda_i$, with $\alpha > 0$ (of course, the two specifications coincide for $K = 0$). It is difficult to determine a realistic estimate of an absolute K: the same threshold could be very small relative to the payment activity of large banks, and very large for small banks. So, the absolute-threshold case is somewhat unrealistic; we use it only as a benchmark case. More realistic is the specification $K_i = \alpha\lambda_i$. Indeed in the UK system, the collateral held for prudential reasons is proportional to a bank's potential liquidity outflows. What is a realistic α, then? Comparing data on collateral holdings and payment activity for UK banks (both available to the Bank of England), one ends up with an estimate in the range 0.1-0.3. This is the range of α that we use in the simulations.

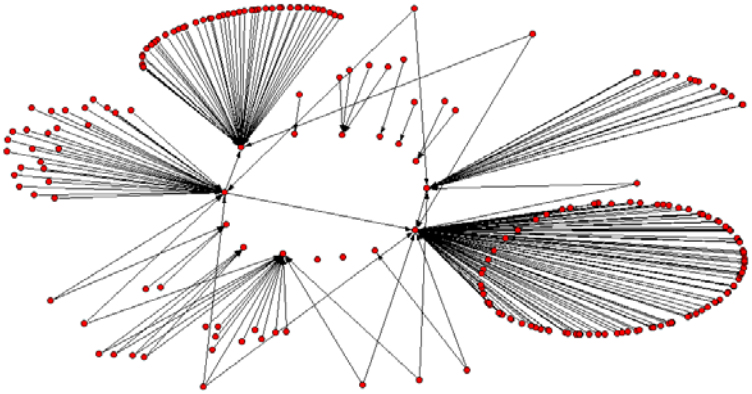
We thus run simulations for a range of parameters, observing the resulting equilibrium network. Recall that banks are called on to make their decisions in a random order so, even for the same set of parameters, simulations may give different results.

[11] With Type-I costs, expected costs (equation **(2)**) are linear in c. So, c rescales all offers proportionally, and is irrelevant for the banks' choices.

[12] With Type-II costs (equation **(2)**) is no longer linear in c. However, it is in f. So we can normalise f (equation **(4)**) by c, and obtain that only K/c (ie K) matters.

Figure 6: Payments *P*, Type-I costs (*K*=0)

Real network



Artificial networks

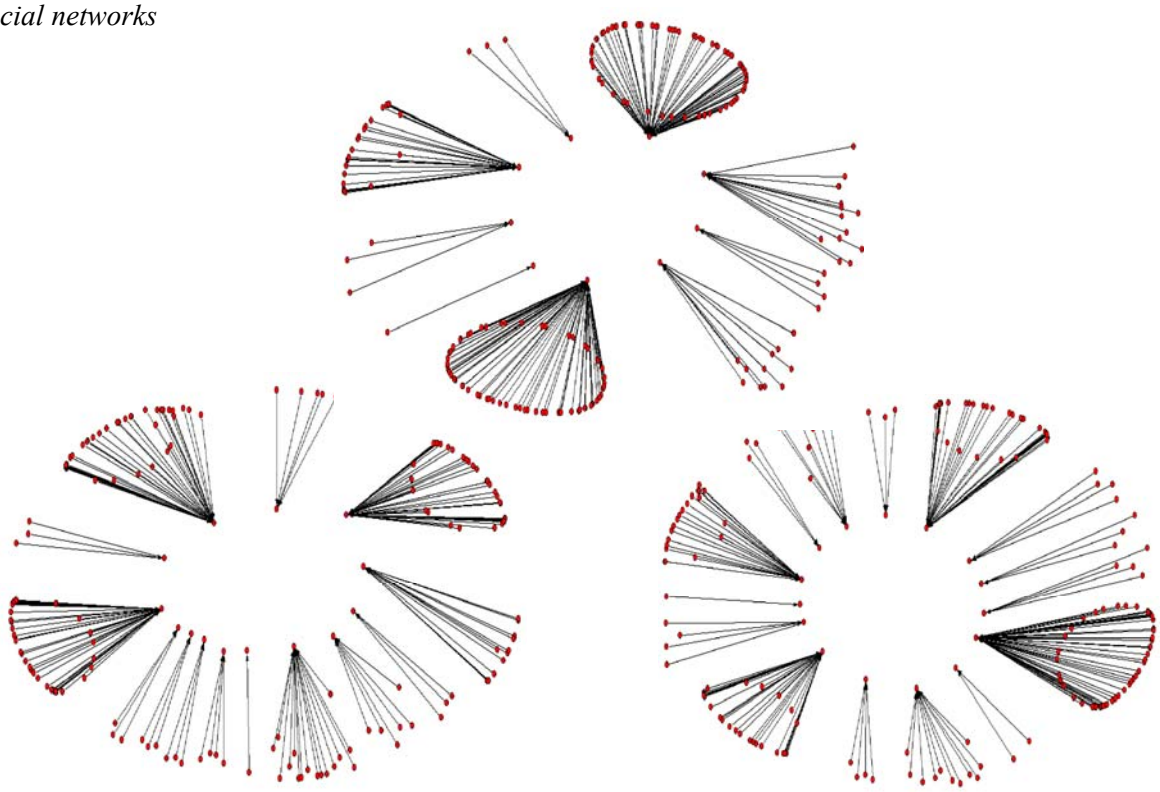


Figure 6 compares the outcomes of three typical simulations (bottom) with a map of the real client-correspondent relationships in CHAPS. If it were not that a few real-life banks have more than one correspondent, the networks in Figure 5 would appear quite similar. The following analysis shows that the similarity goes beyond a mere eyeball test.

Tables 1 and 2 show (average) results obtained in the simulations, for different parameter values. The columns report: the number of direct participants obtained in the simulations; the number of real-life CHAPS participants correctly identified; the proportion of internalised

payments; and Gini coefficients for the shares of payments, and for the number of customers, accounted for by each correspondent. For comparison, Table 3 reports the same statistics for the real CHAPS.

Table 1: Simulations - absolute threshold

K	*N*. corr.	Correct id.	Internal. pay.	Gini pay.	Gini cust.
0 (Type I)	12.3	5.8	28%	0.58	0.58
5	18.2	6.3	25%	0.70	0.64
10	33.2	6.0	41%	0.80	0.71
20	60.4	9.8	29%	0.76	0.61

Table 2: Simulations - relative threshold

α	*N*. corr.	Correct id.	Internal. pay.	Gini pay.	Gini cust.
0 (Type I)	12.3	5.8	28%	0.58	0.58
0.125	12.4	6.5	48%	0.63	0.67
0.15	12.1	5.9	48%	0.62	0.67
0.175	24.2	6.0	26%	0.75	0.75
0.25	35.4	6.4	17%	0.80	0.72

Table 3: Real system

N. corr.	Correct id.	Internal. pay.	Gini pay.	Gini cust.
14	*N*/A	33%	0.61	0.69

With an absolute threshold, the number of direct members of the payment system increases sharply as the threshold rises. By contrast, when the threshold is proportional to the bank's payment volume, increases in α at first do little to encourage direct participation until a critical point is passed (somewhere between thresholds of 15% and 17.5% of payments). This pattern is consistent with the fact that most banks make very small amounts of payments; so, even for small values of an absolute threshold, they are able to settle all payments for free as direct members.

Somewhat surprisingly, the number of real-life CHAPS members identified by the simulations as direct payment system participants is fairly low (typically around 6 of the 14 actual members) and does not increase with rising payment system participation. The banks which are correctly identified as direct participants are, however, the core participants which account for the vast majority of payment flows in CHAPS. Other CHAPS members are only rarely identified as direct payment system members. This suggests that either our model of their liquidity-saving decision is missing something, or that these banks have additional motives for becoming direct CHAPS members (possibilities include historical accident, interdependencies with other payment and securities settlement systems, or a desire to offer sterling correspondent banking services to overseas customers). Alternatively, this could be an artefact of the short period of sample data which we have available to build the matrix P^* - particularly if payments made by some banks show significant seasonality. Or it may be due to possible inaccuracies in the survey itself.

Both Gini coefficients rise along with the number of participants. This is due to the small size of the participants drawn into direct membership as liquidity becomes cheaper. With an absolute threshold, the best fit is obtained with $K \simeq 10$, which however generates too many direct participants. A relative threshold specification performs better: for α between 0.15 and 0.175, we obtain a relatively good fit for both the number of participants, and the Gini coefficients.

Under an absolute threshold, there is no clear trend in the percentage of payments which is internalised, as the threshold is raised. Under variable thresholds, the amount of internalisation initially rises as the threshold is increased from zero. Variable thresholds at medium levels give high amounts of internalised payments, but unlike the case of absolute thresholds these drop rapidly as the number of direct participants increases. This difference in behaviour makes sense, as under absolute thresholds: when the threshold rises, smaller banks will be the first to benefit and so opt to participate directly in the payment system. Under variable thresholds larger banks benefit as well.

Summing up: the fixed and relative-threshold specifications give both a reasonably good fit to the data for low K (the similarity is of course expected, as the two models coincide for $K=0$; the relative goodness of fit is instead the pleasing result). Of the two specifications, we prefer the second, as it represents more realistically liquidity costs in CHAPS. And encouragingly, when K is increased so that the two specifications diverge, the relative-K model outperforms the absolute-K model, which indeed produces the counterfactual result that many small banks become direct members. The best fit is attained for an α between 0.15 and 0.175, which is consistent with estimates of real αs. However, judging from the amount of internalised payments (which appears to depend non-monotonically on α), a reasonable fit is also obtained for a very low α (close to Type-I costs).

4 Conclusions

This paper studies the influence of liquidity costs on the degree of tiering in LVPSs. We formally model the 'netting' and 'liquidity pooling' effects, exploring how these can shape the client-correspondent network of a payment system. The model is extremely parsimonious, requiring essentially two inputs: a matrix of payments (P), and a liquidity cost function (f). Still, when a simple parametrisation is performed using data on the main UK payment system, it produces rather complex networks, which bear an encouraging resemblance with what is observed in reality.

Further work could look at different specifications for the payments' arrival process. Similarly, one could allow for strategic payment behaviour on the part of banks - which would again alter the time profile of payments and liquidity needs, and so be equivalent to a different assumption on the payment arrival process. In a different line, one could incorporate other forms of costs, such as credit risk, into banks' decisions. Finally, the empirical analysis carried out here could be applied to other LVPSs, perhaps applying formal tests to measure the congruence of the 'artificial' networks with the 'real' ones.

Appendix: Computing equation (2)

Recall that $L\_\{i\}(.)$ is a symmetric random walk of length λ (we drop the index i, as it is unnecessary here). So,

$$\begin{aligned} C(\lambda) &= E\left[f\left(\max_t L(t)\right)\right] \\ &= \sum_z f(z)\, p(z,\lambda) \end{aligned} \qquad \textbf{(A-1)}$$

where $p(z,\lambda)$= prob[$\max_{t=0..\lambda} L(t) = z$], and z runs over all possible values that $L_i(t)$ may take as $t = 0...\lambda$. That is, summation runs from $z = -\lambda/2$ to $z = \lambda/2$ (recall that $\lambda = 2\sum_j p_{ij}$, so λ is even).

Now, the distribution function of the maximum of a random walk is well known (it can be obtained using the reflection principle – see eg Norris (1997)):

$$p(z,\lambda) = \left[\binom{\lambda}{z+1} + \binom{\lambda}{z}\right] p^{\lambda}$$

This can be plugged into equation **(A-1)** so, once the functional form of f is specified, an expression for costs can be computed.

As stated at the end of Section 2.3, the function $C(.)$ is increasing for all increasing f. Given what we have said, the proof is rather intuitive: increasing λ lengthens the random walk L. So the pdf of its maxima is skewed towards higher values. As f is increasing, the effect on the sum C turns out to be positive.

Similarly, the existence of a pooling effect (convexity of C) that appears with a linear f and discussed in Section 2.2, can be easily proven by differentiating equation **(A-1)** twice with respect to λ.

In the main text we do assign a specific form to f, and compute C numerically, rather than rely on analytical proofs for the general case.

References

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

Arnold, J, Bech, M L, Beyeler, W E, Glass, R J and Soramäki, K (2007), 'The topology of interbank payment flows', *Physica A*, Vol. 340, pages 380-94.

Bech, M and Garrat, R (2003), 'The intraday liquidity management game', *Journal of Economic Theory*, Vol. 109(2), pages 198-219.

Becher, C, Millard, S and Soramäki, K (2008), 'The network topology of CHAPS Sterling', *Bank of England Working Paper no. 355*.

Chapman, J, Chiu, J and Molico, M (2008), 'A model of settlement networks', *Bank of Canada Working Paper no. 2008-12*.

Harrison, S, Lasaosa, A and Tudela, M (2005), 'Tiering in UK payment systems: credit risk implications', *Bank of England Financial Stability Review*, December, pages 63-72.

Inaoka H, Ninomiya T, Taniguchi, K, Shimizu, T and Takayasu, H (2004), 'Fractal network derived from banking transactions - an analysis of network structures formed by financial institutions', *Bank of Japan Working Paper no. 04-E-04*, April.

Jackson, J and Manning, M (2007), 'Central bank intraday collateral policy and implications for tiering in RTGS payment systems', *DNB Working Paper no. 129*.

Kahn, C M and Roberds, W (2005), 'Payments settlement: tiering in private and public systems', paper presented at the conference 'The future of payments', 19-20 May, available at www.bankofengland.co.uk/publications/events/futureofpayments/kahnroberdsBOE.pdf.

Lai, A, Chande, N and O'Connor, S (2006), 'Credit in a tiered payment system', *Bank of Canada Working Paper no. 2006-36*.

Lasaosa, A and Tudela, M (2008), 'Risks and efficiency gains of a tiered structure in large-value payments: a simulation approach', *Bank of England Working Paper no. 337*.

Lublóy, Á (2006), 'Topology of the Hungarian large-value transfer system', *Magyar Nemzeti Bank Occasional Papers no. 57*.

Norris, J R (1997), 'Markov chains', *Cambridge Series in Statistical and Probabilistic Mathematics*, CUP.