Non-monotonic Tradeoffs of Tiering in a Large Value Payment System

Borradores de ECONOMÍA

Por: Carlos A. Arango-Arango
Freddy H. Cepeda-López

Núm. 946
2016

tá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Col
BANCO DE LA REPÚBLICA COLOMBIA

Non-monotonic Tradeoffs of Tiering in a Large Value Payment System

Carlos A. Arango-Arango Freddy H. Cepeda-López*
Banco de la República de Colombia

Abstract

Even though international authorities encourage open and wide access to large value payment systems, the optimal level of access, or tiering, is still an open question. In the case of real-time gross settlement systems (RTGS), the level of access, or tiering, may be limited by the tradeoff between: (i) potentially higher liquidity needs of a larger pool of direct participants settling in real time and (ii) the lower counter-party credit risks that result from a lower number of second-tier participants entering in uncovered bilateral credit positions with correspondent banks. Previous literature has evaluated this tradeoff through simulations finding monotonically increasing liquidity savings and increasing credit risk exposures as the level of tiering in the system rises. In contrast, we find that in the Colombian RTGS case liquidity savings increase but then decrease with higher tiering showing a hump shape. Our results provide insights into the effects of tiering when participants are too-big or too-connected to tier.

Key Words: Tiering; RTGS payment systems; liquidity; Credit risk; network topology.

JEL Codes: E42, G21, E58, C15.

1 Introduction

The operators of large value payment systems (LVPS), mostly central banks, are always confronted with the question of what should be the level of access to their systems and to the liquidity facilities they provide. Principle 18 of the "principles for financial market infrastructures" of the BIS encourages fair and open access to the systems by financial market participants and other financial infrastructures to stimulate competition, interoperability and efficiency (BIS-CPSS, 2012). The tradeoffs between the costs and risks involved in the provision of payment services and the benefits to participants in terms of liquidity efficiency determines the level of participation in a payment system. Systems that are based on multilateral netting are low cost in terms of liquidity needs but high risk in terms of the credit exposures and systemic risks when one or more participants default. Real-time gross settlement systems (RTGS) are low risk as they contain well the risk of default but are costly in terms of the heavy liquidity requirements imposed on participants.

*Carlos A. Arango-Arango: +(571) 343 1793, carangar@banrep.gov.co. Freddy Cepeda-López: +(571) 343 0644, fcepedlo@banrep.gov.co. We thank Carlos E. León, Charles Khan, Clara L. Machado, Dionisis Philippas, Fabio G. Ortega, Hernando Vargas, Joaquín F. Bernal, Kimmo Soramäki, Pamela A. Cardozo and Radoslav Raykov for their insightful comments and suggestions. We also thank participants at the World Symposioum on Banking and Finance (Dubai, 2016) and at the Modelling and Simulation of Payments and Other Financial System Infrastructures conference (Bank of Canada, 2016). We also thank Carlos A. Cadena for his invaluable research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Colombian central bank.

Within a particular payment system design, however, central banks can choose the allocation of costs and risks by adjusting their set of prices, both of access and liquidity management in the system, and their minimum operational and financial entry standards. The more stringent access to an LVPS is, the more tiered the payment network becomes, with more financial institutions operating as indirect participants (IPs), relaying on correspondent institutions that are direct participants (DPs) in the system[1] to complete their payment flows. In an RTGS, a higher degree of tiering may reduce the needs for liquidity of DPs, which take advantage of the payment flows of their clients. Yet, a more tiered payment system may increase credit risk as DPs and their clients (IPs) are exposed to the failure of one another. Finally, it may increase the concentration risk of the system as fewer DPs may process increasing payment volumes, rising the risks of too-big-to-fail (TBTF) or too-connected-to-fail (TCTF).

Lasaosa and Tudela (2008) and Arculus et al. (2012) are, as far as we know, the only studies in the literature that try to estimate these tradeoffs. Using data from CHAPS, Lasaosa and Tudela (2008) use the Bank of Finland payments simulator (BoF-PSS2)[2] to quantify the liquidity needs and bilateral credit risk exposures of IPs and DPs, by simulating increasing levels of tiering in an already concentrated payment system (13 DPs out of 420 financial institutions in the UK economy). They use their results to make extrapolations for a hypothetical less tiered CHAPS system, finding that the higher liquidity costs of a less tiered system are not economically significant and advocate for a less tiered system on the grounds of reducing concentration risks in CHAPS and bringing in financial intermediaries considered potential systemically important. Arculus et al. (2012) make the same calculations for the case of the Australian payment system RITS. They find that the increasing liquidity needs of a less tiered RITS depend on the system design. A pure RTGS would put more strain on liquidity needs as participation increases, yet these needs drop considerably once RITS liquidity-saving mechanisms are introduced in the simulations. The authors see these results as supporting the fact that the actual wide access to RITS is due to the lower participation costs induced by these mechanisms.

We contribute to this research by studying tiering in the case of a national payment system in an emerging economy: the large value payment system CUD (Spanish acronym for the Deposit Accounts System) operated by the Colombian central bank. CUD started operations in 1998 as an RTGS system, and currently settles daily averages of 7,570 transactions, whose values represent about 4.8% of GDP (BdR, 2015). The key attributes that make CUD a relevant case of study are its low levels of tiering, since it currently supports 150 participants out of 290 financial institutions operating in the country,[3] and the heterogeneity of DPs among both banking and non-banking financial institutions. Furthermore, 128 out of the 150 participants have access to central bank intraday and overnight liquidity through repo operations to facilitate their payment obligations.

In this article we calculate the potential liquidity savings of increasing tiering in the Colombian payment system using simulation methods and contrast them with the additional credit and concentration risks. We decompose liquidity savings into those coming from pooling and those from internalization of payments. Savings by pooling result from the additional liquidity that settlement banks may get from combining their payment flows with those of their clients. In this way settlement banks may fund more payments from receipts rather than using other sources of liquidity. Savings by internalization result from the bilateral

[1] In this document the term direct participant (indirect participant) is used interchangeably with correspondent or settlement institution (client), and first-tier (second-tier) participant.

[2] See http://www.suomenpankki.fi/en/rahoitusjarjestelman_vakaus/BoF-PSS2/Pages/default.aspx

[3] This total includes commercial and investment banks, brokerage firms, financial cooperatives, insurance companies and mutual funds among others.

settlement of payment obligations between the second-tier participants and their correspondent institutions in the correspondent institution's books.

The results show that the Colombian payment system could become significantly tiered without much growth in concentration risk in the system. Similarly, counter-party risk arising from exposures of settlement banks to IPs remains low compared with settlement institutions' capital and therefore the likelihood of systemic contagion due to the failing of an IP remains low as well. However, tiering beyond the smallest 90 DPs (about 50% of all DPs in CUD) makes both risks grow exponentially reaching levels that could become systemically large in terms of solvency. We find that increase in credit risk introduced in the system's DPs by new IPs is weakly correlated with the size of their new clients, measured by their values settled in the system. In addition, liquidity savings start to be economically meaningful with the tiering of the largest 40 direct participants. Indeed, these savings could reach 13.4% of the liquidity needs currently faced by CUD participants.

In contrast with previous findings, our results show evidence that liquidity savings do not necessarily increase with tiering. Depending on the nature of their business and their payment flows, the matching between second-tier clients and correspondent institutions could result either in liquidity savings or more stringent liquidity needs. In fact, we find that systemwise, tiering simulations in the Colombian case show a hump shape with increasing liquidity savings as larger DPs are tiered, up to a point where additional tiering comes with decreasing liquidity savings for those remaining as DPs in the system.

This is due mainly to the fact that as tiering increases both the size and interconnection of tiered participants rise. As it is found in other interbank marketsCraig and von Peter (2014), CUD is shown to have a hierarchical structure with large participants serving as hubs of connectivity in the network. Tiering small participants would bring liquidity savings to correspondent institutions as they are already part of the correspondent institution's hub. Yet, tiering more interconnected institutions may bring small liquidity gains to their correspondent institutions. Indeed, these corespondent institutions may inherit part of the liquidity needs that their clients had from their dense payment obligations with the rest of the direct network before being tiered.

The paper is organized as follows. Section two discuses the main issues related with tiering in payment systems. Section three provides descriptive figures and facts related to the Colombian large value payment system. Section four presents the simulation methodology. Section five provides the simulation results and section six concludes.

2 Tiering tradeoffs

Most financial systems in the world face some degree of tiering in the access to their critical financial infrastructures. Usually, the level of tiering is a result of the tradeoff between provision costs and efficiency gains, which determine the convenience of accessing such systems. In the case of an RTGS, settling on a real-time gross settlement basis avoids the accumulation of large credit positions among participants typical of differed multilateral netting systems, and therefore potential systemic credit risks are well contained (BIS-CPSS, 2005). However, an RTGS tends to be costly in terms of liquidity because settlement occurs in real time and funds have to be granted at the moment payment orders are submitted to the system. To mitigate the liquidity costs of an RTGS system, central banks could provide access to collateralized or uncollateralized (interest-bearing overdrafts) liquidity or introduce liquidity saving mechanisms (Martin and McAndrews, 2008).

Those financial institutions that do not find it profitable to have direct access to an RTGS usually process their payments through a correspondent institution which has direct access to the system. In contrast with DPs that clear and settle in a multilateral fashion, IPs settle bilaterally. Correspondent banking could become large and complex as the number of IPs grow and the activity becomes more concentrated as correspondent firms become bigger and more specialized.

According to Lasaosa and Tudela (2008), as a payment infrastructure becomes more tiered and correspondent banking becomes more prominent, the liquidity needs of the payment system tend to decrease due to pooling and internalization effects. Savings from pooling result from the additional liquidity that settlement banks may get from combining their payment flows with those of their clients. In this way settlement banks may fund more payments from receipts rather than using other sources of liquidity. Savings from internalization arise from payment obligations between second-tier participants and their correspondent institutions. The settlement of this obligations in the correspondent institution's books allows for arrangements less liquidity-demanding than in real-time gross settlement.

Yet, a higher degree of tiering may increase various types of risk and could even threaten the systemic stability of the payment system. Counterparty credit risk may build up as correspondent institutions usually offer overdraft facilities to IPs when making outward payments on their behalf. These intraday lines of credit may become overnight exposures in stressed situations (Flannery, 1996). Conversely, when IPs hold positive account balances with DPs, they are exposed to the risk of DPs' default.

However, the credit risk exposure that correspondent institutions face from a default of their IP clients brings strong monitoring incentives (Kahn and Roberds, 2009). As second-tier banks often rely on uncollateralized intraday credit lines from their settlement banks, these bilateral relationship induces high risk management standards imposed by the correspondent institutions to IPs. Becoming DPs in a financial infrastructure may curtail such commercial relationships.

Higher levels of tiering could also induce higher hierarchical levels in the network structure, with high levels of dependence on correspondent banking for liquidity, clearing and settlement services. In particular, IPs would be more vulnerable to an operational disruption or stressed situation of their correspondent institution, with limited options to respond promptly and use other channels to continue sending payments and accessing liquidity. Also, it is more likely to have systemic risks (e.g. "liquidity sink") due to failures of TBTF or TCTF DPs, as correspondent banking increases.

In addition, a more tiered system is more prone to legal risks as bilateral payment obligations between IPs through a correspondent institution will not have the same level of protection against legal claims to their finality compared with a payment that goes through a designated systemically important payment system (Bernal-Ramirez, 2012).

Finally, a more tiered payment system may give DPs a market advantage over second-tier institutions. IPs may face incentive problems with the conditions and fees correspondent institutions may place on them. If direct access is limited to a few it could be prone to inefficiencies drawn from monopolistic market structures. Also, there is a tension between IPs and correspondent institutions in terms of the access to business-sensitive information by an agent that could potentially be an IP's competitor (Lai et al., 2006). Harrison et al. (2005) points out that second-tier clients could be exposed to substantial liquidity needs and may lead settlement banks to limit or require costly collateral to provide intraday credit. In this regard the client may have a disadvantage as IP, as it is liquidity dependent on the correspondent institution and has no access to central bank liquidity, specially in times of distress. Evidence of this is shown in Jackson and Manning (2007). Bernal-Ramirez (2012) suggests that payment systems and financial infrastructures could actually benefit from a larger pool of direct participants by speeding up payments, reducing costs, and maintaining a more even competitive field.

In deciding to become a DP in a payment system, a financial institution faces a tradeoff between the costs of liquidity and collateral requirements as well as membership fees and the benefits associated with becoming part of a privileged club-type payment services platform. There are a number of costs related to direct participation, including: employing staff to manage counterparty risks, payment flows and liquidity needs; procuring and maintaining computer and communications systems to handle the flow of payments; and the cost of holding sufficient liquidity and collateral to enable settlement (BIS-CPSS, 2005). These costs mean that it may be unprofitable for some financial institutions to become DPs, especially those that make relatively few payments. Economies of scale may enable settlement banks to offer payments to an IP at a lower unit cost than direct participation. For systems which settle in central bank money, the costs of holding the collateral necessary to access central bank intraday liquidity are among the factors against direct participation, although correspondent institutions may well charge for the provision of liquidity to second-tier customers. Jackson and Manning (2007) find that systems would tend to be concentrated due to the economies of scale in payment processing.

A model of correspondent banking without central banks, as formulated by Chapman et al. (2013), shows that the competitive tiering equilibrium increases with the degree of economies of scale in the provision of the system and falls in the level of public information available about the creditworthiness of potential participants. Therefore, in a competitive market, it is efficient to let the banks that have an advantage either informational or in their cost structure to be the DPs and provide correspondent banking services to second-tier participants.[4]

These observations are also consistent with the view that inter-bank markets have a tiered structure. Earlier research on the federal funds market suggests, for example, that small banks tend to turn over surplus funds to large banks that distribute or invest the funds (Ho and Saunders, 1985; Allen and Saunders, 1986; Bech and Atalay, 2010; Craig and von Peter, 2014).

It is worth mentioning that little work has been done about the implications of different levels of tiering in a LVPS and the efficiency of monetary policy, specially in times of financial stress. Also, little has been done on how the decisions to be a DP or an IP would depend on a joint decision within a financial group. Also, we could not find theoretical work on tiering and network effects. Some of the literature on systemic risk suggests that the likelihood of a systemic impact of a participant's default may have a U-shaped relationship with the degree of connectivity in the inter-bank market: falling as participants are more connected, which allows for better risk and liquidity sharing but, at some point, increasing as contagion becomes more likely in case of default of a DP. As more tiering could induce higher concentration and connectivity among DPs in the system, there could also be an optimal level of tiering from a systemic-risk point of view. These unresolved issues leave the optimal level of tiering in a LVPS still an open question.

[4] Tiering is the natural way agents adjust to the set of costs and risks that the predominant payment institutions offer. As history has shown, after any new innovation in payment systems, agents have developed practices and businesses that try to moderate the costs implied by direct participation and corresponding banking has been at the center of such developments (Kahn et al., 2016).

Table 1: Main financial institutions with direct access in CUD (2014)

Institution type	Number	Main purpose
Commercial Banks	22	Provision of deposit and loans, including mortgages
Financial companies	21	Provision of deposit and loans focused on goods and services commercialization (e.g. leasing)
Financial Corporations	5	Provision of deposit and loans focused on medium term industrial financing, akin to an investment bank
Financial Cooperatives	5	Provision of deposit and loans focused on the cooperative sector
Trust companies and mutual funds	28	Provision of investment vehicles with the purpose of investing in securities and other assets according to the risk profile of investor
Brokerage Firms	22	Provision of brokerage services with the purpose of buying or selling securities; allowed to trade for its own account
Pension Funds	4	Provision of investment vehicles with the purpose of investing for retirement
Special Official Institutions	10	Official (government owned) financial institutions with special objectives

Source: modified from León et al. (2011)

3 Tiering in the Colombian Large Value Payment System

The large value payment system in Colombia CUD is a systemically important infrastructure operated by the central bank.[5] Since its inception in 1998, CUD has operated as an RTGS system. It provides clearing and settlement services to other systems such as securities depositories, retail payment systems, the foreign exchange Clearing House and the Central Counter-party Clearing House. In this array of infrastructures, the one that generates most of the payment flow activity is the central bank central securities depository (DCV), mostly dedicated to the clearing of bonds issued by the Colombian government.

As of 2014, CUD supports 150 participants out of 290 financial institutions operating in the country. All financial institutions and some special official institutions are allowed to have direct access to CUD. Table 1 spells out the most relevant types of financial institutions with direct access to CUD. In 2014, CUD settled daily averages of 7,570 transactions, representing in value about 4.8% of GDP (BdR, 2015). The most active institutions in CUD are the commercial banks (CB) and brokerage firms (BF), which account for around 75% of all payments. Of the 150 direct participants in 2014, 14 made about 71% of all payments: 12 credit institutions 61% and 2 brokerage firms 10%.

There are several reasons for the wide access to CUD. First, monthly fees are a function of the value

[5] Accordingly with the principles stipulated in BIS-CPSS (2012).

of payments settled by each participant. Second, there is relatively "cheap" intraday liquidity available in the system because of high reserve requirements imposed on credit institutions. In fact, the ratio of disposable reserves to liabilities subject to reserve requirements is quite high at 7.8% in 2014. Third, as pointed out by Arculus et al. (2012), lower liquidity costs are induced by liquidity-saving multilateral-queuing algorithms, which, in CUD, are activated five times a day. Finally, practically all types of financial institutions (except insurance companies, capitalization companies and general deposits warehouses) have access to the collateralized intraday lines of credit (intraday repos), convertible into overnight repos, for their own purposes or on behalf of third parties, which adds to the benefits of being a direct participant.

Of all payments settled in CUD, 49% are associated with portfolio operations against public debt deposited in DCV.[6] Direct transfers among government's accounts are close to 15%. Unfortunately, for the gross part of the remaining share of payments in CUD (35%), we do not have a way of identifying what percentage is done on behalf of second-tier participants. Therefore, we can not make simulations to study further entry of second-tier IPs into CUD, and instead focus on increasing the level of tiering in the system.

Our research contributes to the understanding of tiering in the Colombian payment system by looking at the potential gains in terms of liquidity savings the system could achieve by increasing second-tier participation and weigh it against the potential rise in counter-party risks and changes in the structure of the payment network.

4 Tiering simulations

We follow the methodology in Lasaosa and Tudela (2008) and Arculus et al. (2012) and use simulation techniques to study differences in system risks and liquidity needs under different tiering scenarios. We use the Bank of Finland's payment and securities system simulator ((BoF-PSS2), which takes payments data (real or artificial) as input and processes it following predefined settlement rules and liquidity configurations. It produces, among its outputs, account balances and summary statistics about settled and unsettled payments.[7]

Our payments data is made up of payment flows performed in CUD during the month of April 2013. This month is chosen as it is representative of the average monthly flows through CUD and is not affected by seasonal events or financial distress. During such month, direct participants (excluding the central bank and government accounts) made a daily average of 7,320 payments worth COP$ 31.7 trillions. For each payment, we have information about the sender and the receiver of the payment, the payment amount, the exact time at which it was settled and whether the transaction was carried out in one of the five cycles of the liquidity saving mechanisms.

One main indicator used across the article is ML_i which represents the maximum liquidity needs of a DP in CUD to settle opportunely all its obligations given the observed payment timing. This indicator is

[6] Only few of these payments (less than 5%) are the result of portfolio operations in DCV made by DPs on behalf of IPs.

[7] The BoF-PSS2 is the payment and securities settlement simulator provided by the Bank of Finland. The simulator allows researchers to model multisystem setups that can be a combination of payment, securities and settlement systems, as well as central counterparties among other financial market infrastructures (FMI). The BoF-PSS2 simulator can be used to analyze several issues including settlement, liquidity and credit risks in FMI; systemic and counterparty risks in FMI; identification of critical counterparts; policy change impact evaluation; network analysis; liquidity dependency analysis; relationship analysis of monetary policy and liquidity needs for settlement of payments; evaluation of sufficiency of liquidity buffers and margins; mergers´ effects on liquidity needs; system performance benchmarking, testing and development of liquidity saving algorithms; and system development and prototyping.

calculated as the value of the lowest intraday balance B_{it} reached by a participant in a payments simulation,[8] assuming that participants start with zero balances and have no access to other sources of liquidity different from the net flow of incoming and outgoing payments:[9]

$$ML_i = \mid \min(0, \min_t(B_{it})) \mid$$

Our benchmark scenario consists of a simulation that replicates the payment flows performed during the sample period. Our simulations then work through an increasing degree of tiering. We can identify the payments sent and received by DPs that become IPs, giving us the basis for the simulations under different scenarios. To estimate the liquidity needs of the remaining DPs (ML_i) in each scenario, we drop all transactions between the settlement institution and the new client as well as those made between the tiered participant and the other clients of the settlement institution.

We increase the degree of tiering by turning the smaller DPs, by payment shares in the benchmark scenario, into IPs. The assignment of an IP to a correspondent DP in each scenario is made based on the total value of incoming and outgoing payments in the sample period (based on the adjacent matrix of payment flows[10]). IP_i at scenario s is assigned to the correspondent institution CI_{is} with whom IP_i has the highest value of gross payment flows in the benchmark scenario, among the set of DPs available in scenario s, DP_s:

$$CI_{is} = \{DP_j \in DP_s \mid P_{ij}^I + P_{ij}^O \geqslant P_{ik}^I + P_{ik}^O \quad \forall \quad DP_k \neq DP_j \in DP_s\}$$

where P_{ij}^I and P_{ij}^O are, respectively, the sum of incoming and outgoing payments made in the sample period between participants i and j in the benchmark scenario.

This is the criteria used in Lasaosa and Tudela (2008) and Arculus et al. (2012) to match participants in their tiering analysis. We take this practitioner´s approach to see if the results also hold in the Colombian case. As a robustness check, we also performed all the calculations based on discretionary payments, i.e. those that correspond to the transfer of funds for which the responsibility to settle does not depend on the rules of a clearing and settlement infrastructure, but on the willingness and time preferences of participants. This alternative was thought of as a good proxy for proximity based on trust. Yet, the results are similar to those using the "total payments" criteria and, therefore, we present the results of the later approach for comparison with the previous literature.

As we explain in detail below, it turns out that the results for the Colombian case are markedly different from those found for the UK and Australia. We show evidence that liquidity savings do not necessarily increase with tiering. Indeed, we show that the size and degree of connectivity of those DPs that are being tiered and the nature of their payment flows with their correspondent institutions influence the results. These findings suggest that matching in terms of gross payment flows may be misleading with regards to the potential liquidity savings that could be achieved in a correspondent institution's relationship with a client, which opens new venues for further research.

[8] The word "simulation" in this paper does not refer to the use of probabilistic or Monte Carlo techniques or to the use of a particular statistical model. It refers to the calculation of different statistics such as intraday account balances under certain assumptions, based on actual daily intraday payment flows taken from the sample period.

[9] This indicator corresponds to the upper-bound concept defined in Leinonen and Soramäki (1999).

[10] An adjacent matrix is a way to summarize connections among nodes in a network. Each column (row) represents a node, and each element of the matrix a_{ij} provides information about the edge between nodes i and j. In the simplest representation a_{ij} takes the values zero or one, indicating whether there is an edge between nodes.

We consider as tierable institutions, all current participants in CUD except for what we call core institutions: credit institutions that are critical to various aspects of the central bank's operations and mission. In particular, they include commercial banks for which it is mandatory to have an account with the central bank and market-maker institutions chosen by the ministry of finance for the public debt market. As BIS-CPSS (2013) states, central banks limit the access to central bank money, through central bank accounts, as they usually do not want to compete in the provision of banking services with the private sector. In this sense, the access of commercial banks to central bank accounts is well justified for policy reasons, as in most countries they operate on the basis of fractional banking and, due to their intrinsic fragility, require a safety net that includes lender-of-last-resort facilities at central banks.

We perform 132x20x3 simulations in total, corresponding to 132 tiering scenarios run over 20 days and three types of simulations that aim at estimating pooling and internalization liquidity savings. For each simulation, we calculate daily B_{it} and ML_i for each DP at different tiering scenarios. From these figures we calculate different statistics such as sample period averages, maximums and minimums of ML_i per DP. We also calculate aggregates of liquidity needs across DPs, as the sum of sample period averages for ML_i.

There are three basic assumptions to bare in mind in our analysis of tiering based on simulations: (i) we assume that the timing of payments does not change when a DP becomes client of a correspondent DP; (ii) we assume that those financial institutions that become second-tier participants will keep their customer base and therefore will not loose (gain) payment flows by turning into IPs; and (iii) we assume, as is done in these literature, that tiered participants can only resort to one an only one correspondent institution to have access to the LVPS.

These assumptions could bias the results as tiering should induce both IPs and DPs to adjust the timing and other attributes of their payment obligations and there may be efficiency and risk motivations for having commercial relationships with more than one correspondent institution. As Lasaosa and Tudela (2008) point out, it is hard to determine in what way the results would be biased using the simulator without making arbitrary choices about timing and payment flows and alternative matching arrangements between tierable institutions and DPs. We believe biases due to possible changes in timing and other payment attributes are bounded by the fact that more than 80% of CUD payments are high-priority rather than discretionary obligations (Cepeda and Ortega 2015).

Otherwise, our work underscores several issues regarding the intricacies of evaluating tiering under the IP-DP matching assumptions used in the literature, with respect to its impact in the structure and attributes of the network. Yet, it is beyond the methods used here to inquire about optimal participant strategies towards tiering.

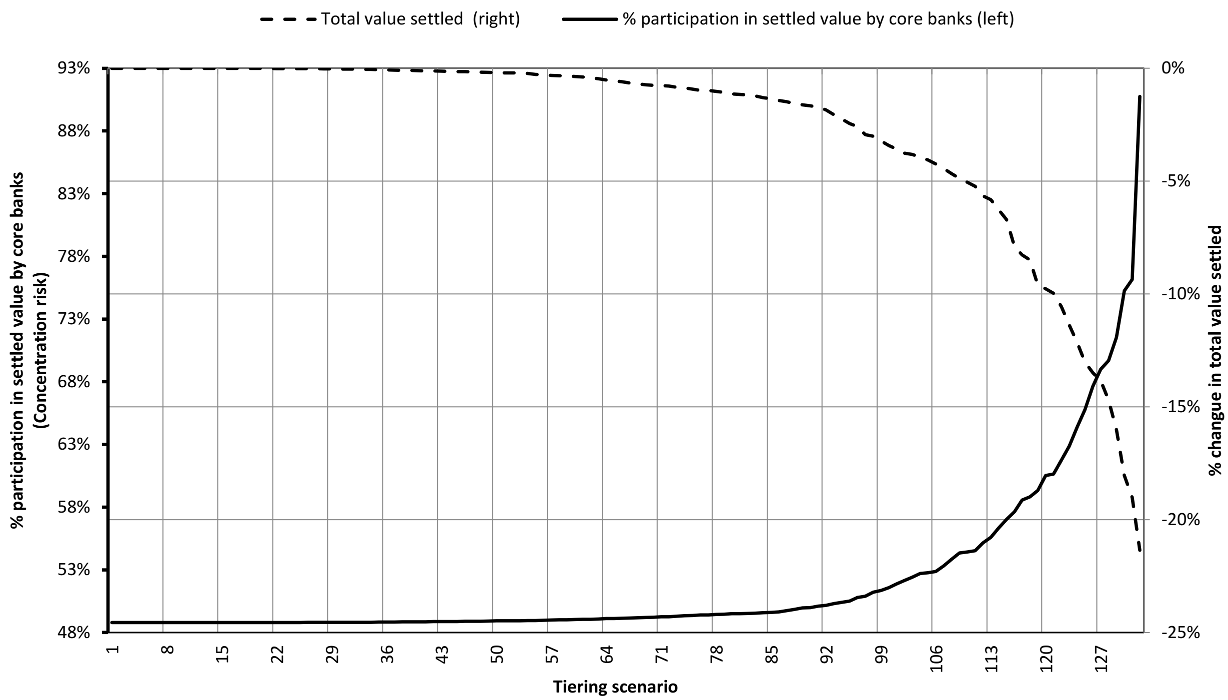
5 Results

5.1 Concentration risk and network configuration

A greater degree of tiering means fewer DPs in the system, which would process not only their own payment flows but likely an increasing number of payments from IPs as more participants become their clients. The simplest way of estimating how concentration changes in the system is by payment shares. The payment share of a DP j in the system is defined as:

$$Sh_j = \frac{Value\ Sent_j + Value\ Received_j}{CUD\ total\ sent\ and\ received}$$

Figure 1: Increase in tiering, settlement amounts and concentration risk in the system

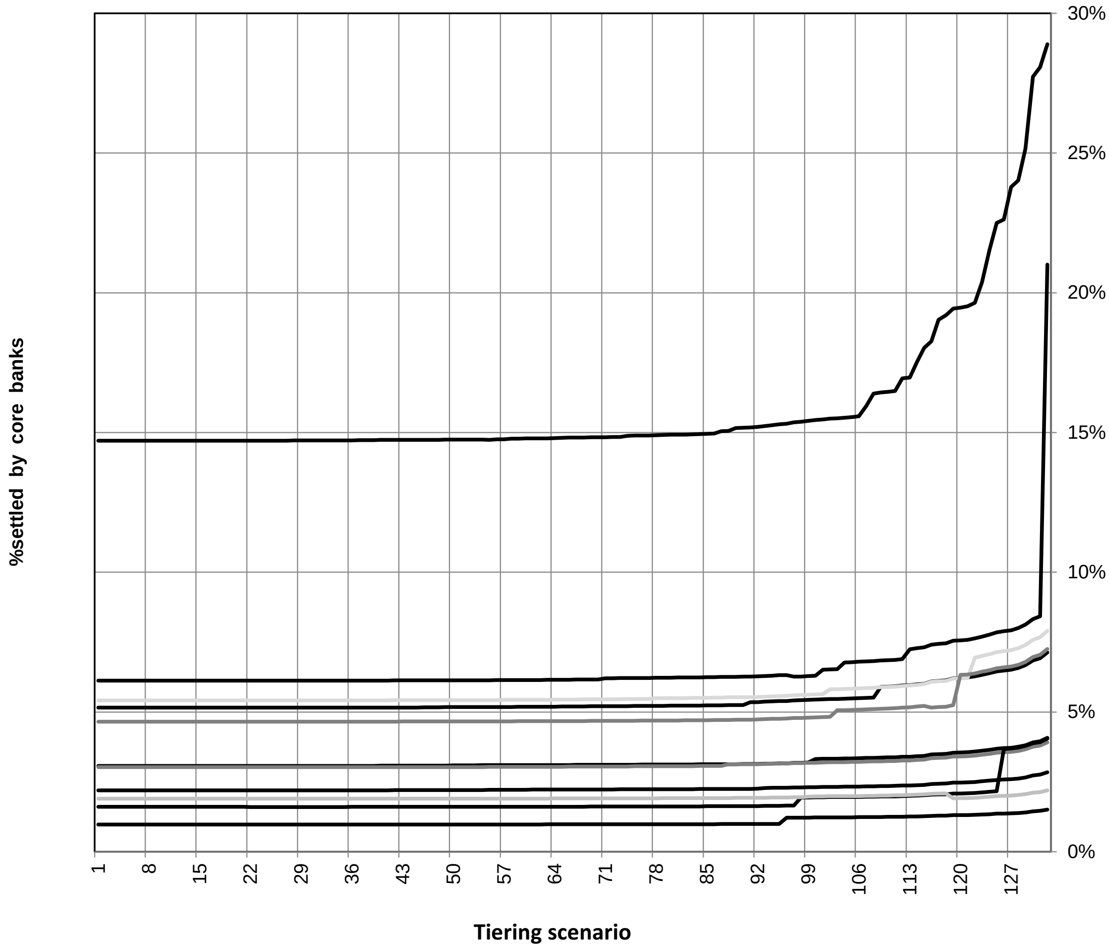


The tiering simulations show that concentration risk only appears an issue as we move beyond the 100th tiered scenario (Figure 1). Indeed, the payment share of the core 11 participants only increases few percentage points, from 49% to 52% up to that point. Beyond this point, payment concentrations start to grow exponentially, making core banks increase their participation in 23 percentage points as larger participants become IPs in the simulations.

However, concentration risk differs among remaining settlement institutions. As Figure 2 shows, most of the increase in concentration risk is biased toward one core DP (#6), which indicates potential growing too-big-to-fail risks (TBTF) associated with particular institutions due to tiering. This is due to the different degrees of centrality that core participants have in the system. Those DPs more connected in terms of the adjacent matrix of payment flows would have closer relationships with a larger number of IPs and, therefore, a larger pool of clients. The second factor could be due to the level of internalization of payments between settlement institutions and their clients. Those correspondent institutions that tend to send payments to other DPs on behalf of their clients rather than clearing in their own books would increase their payment shares and therefore increase their concentration levels in the system.

In this section we go beyond this basic level of analysis in the literature and look at the properties of the network of DPs (DPN) as well as that of the entire community of participants, both DPs and IPs (TN). A payment system network, or graph, is defined in network theory terms as a set of nodes or vertexes (e.g. DPs in CUD) and links, edges or connections between nodes (represented by payment flows in payment networks). Based on this description, we can study the topological properties of the DPN and the TN when

Figure 2: Increase in tiering and concentration risk of core settlement institutions



Note: Each line corresponds to one of the core institutions, with higher levels of concentration reflecting more prominence in payments. Labels are not shown as the name or type of the institution can not be reveled.

we increase the degree of tiering in the system.[11]

Looking at the DPN at different degrees of tiering is of relevance since it portraits the linkages between the set of participants that are relevant to the operations of the Central Bank's LVPS and its liquidity and lender-of-last resort facilities.

Concentration risk also renders well when translated in terms of different degrees of connectivity and tiering in the DPN. We calculate the density of the DPN for each tiering scenario defined as the average of the number of adjacent edges of DPs nodes in the system divided by the total number of edges in the network of DPs. We find that the network of DPs becomes more connected as the level of tiering increases, reaching the maximum level (1), meaning that the DPN becomes complete, when only the core financial institutions remain as DPs. In this sense, as tiering increases, connectivity of the remaining DPs in the system increases to the point in which each may have a too-connected-to-fail (TCTF) position in the DPN (see Figure A.1 in appendix A available online).

Also, the average geodicity distance in the system falls. This statistic measures how far DPs are from each other in terms of the number of edges involved in a shortest path connecting them in the network. We find that in the less tiered scenarios the average geodicity distance among DPs is about 1.8 but as the tiering level increases, naturally this measure falls towards 1. This is corroborated by local clustering, which shows that the average probability that two neighbors of a vertex are themselves neighbors increases, specially after scenario 75th.[12] The combination of higher degree levels and lower geodicity distances translates into a DPN with lower levels of network centralization, measured in terms of degree and closeness, as the level of tiering increases (see Figures A.1, A.2 and A.3 in appendix A available online),[13] but with each of the DPs becoming equally densely connected.

Finally, although most network statistics show a tighter level of connectivity as tiering increases, its nature may vary in unexpected ways. We calculate the correlation between degree and local clustering across DPN's vertexes in different tiering scenarios. A negative correlation provides evidence that the system has a hierarchical structure where those nodes that have higher degree tend to have lower clustering levels and vice versa. Even though we find an increase in negative correlation as tiering increases, the process is non-monotonic, with a window of scenarios where tiering may actually reduce the hierarchical level of the network (see Figure A.2 in appendix A available online).

These results confirm that the DPN is a modular scale-free network as pointed by León and Berndsen (2014): that is, it departs from an homogeneous network formed from random matches where all nodes tend to be equally connected and it is modular in the sense that there is some hierarchy where groups of participants have a high density of links among them, with a lower density of links between groups or modules. These features make the DPN robust to random shocks, yet fragile to targeted attacks (Haldane, 2009). Moreover, the DPN conserves this modularity even as it tends to a complete graph at higher degrees of tiering. However, the nature of DPs that are made second-tier participants (e.g. those central in terms of incoming or outgoing payments) may substantially decrease the level of modularity in the DPN.[14] Changes

[11] See León and Berndsen (2014) for a detail account of network theory applied to payments.

[12] See Newman (2010) for further details on clustering.

[13] These two measures of centralization are based on the difference between the degree and geodesic distance of each vertex and that of the vertex most connected (highest degree or lowest geodesic distance). Usually the sum of these node differences are compared with that of a star-shaped network which would be the most central configuration with all nodes single-linked to a central node (Freeman, 1979;Wasserman and Faust, 1994). These measures provide a way to compare the overall degree of centralization between different graphs or networks and differ from centrality measures that focus on the connectivity of a particular node, such as the normalized degree and the geodesic distance measures.

[14] Keep in mind that for a complete graph such correlation would not be computable as there would not be variance either

in the modularity levels of a network are important, specially for the DPN's system CUD, as they limit cascades and isolate feedbacks (Haldane and May, 2011; Babus and Hu, 2015).

In the TN, on the other hand, we find that it tends to become less connected, in terms of degree, and its geodesic distance increases as participants that once had the possibility to connect directly to other participants now are forced to interact through a correspondent institution. We also find that the level of clustering drops with tiering as the average probability that two neighbors of a vertex are themselves neighbors decreases because IPs become, by construction, satellites of correspondent institutions (see Figures A.4 and A.5 in appendix A available online). [15]

Interestingly, the way tiering is conducted in our scenarios flips the modularity properties of the TN. Indeed, the correlation between degree and clustering moves from being negative, when small DPs are tiered, to positive early on at scenario 25. This is because tiering creates an increasing number of vertexes in the TN, i.e. the IPs, with low-degree and low-clustering levels, as their linkages collapse to only one edge with their correspondent financial institution. This is clearly illustrated in the video clip of Figure 3, which presents the graph of the TN changing with tiering up to scenario 132. Note however, that there is a tension between a larger numbers of IPs and the consolidation of strong hubs of correspondent institutions (star-shaped in the argot of network theory) with a clear high-degree low-clustering structure traits typical of modular networks.

In some sense, tiering can be associated with modularity induced by a process of homeostasis, a feedback process discussed in Assenza et al. (2011) that produces modularity by weakening certain types of interactions and encouraging others. For example, tiering could be induced by increasing the costs of accessing a LVPS directly or increasing the liquidity costs or risk management requirements of being a DP. A natural unraveling of the network would happen into institutions specializing in correspondent services and IPs, both searching for best matches in terms of liquidity savings from internalization and pooling, intermediated by the management of counterparty risk in these relationships (Adams et al. 2010; Chapman et al., 2013; Babus and Hu, 2015).

In fact, the hump shape we find in the hierarchy plot of the TN as tiering increases reflects the relative higher weight played in the correlation between degree and clustering by these high-degree low-clustering hubs such as #133 and #134 as tiering approaches scenario 132 in the video clip of Figure3 (see Figure A.5 in appendix A available online). This shows how the business nature of a participant may influence the shape of the graph. The tiering of large participants that are already playing an important role as correspondent DPs, due to the already high degree of tiering, changes the local properties of those remaining as DPs that inherit the clientele of tiered former correspondent institutions.[16]

These transformations are also reflected in the level of centralization of the TN. We find that the way IPs are assigned to DPs, based on the gross payment flows, induces a reduction in centralization in terms of degree and closeness (see Figure A.6 in appendix A available online). This is due to the way this measures are constructed, as they summarize the degree (geodesic distance) levels of all nodes relative to the node with the highest degree and shortest geodesic distances. In this sense, degree centrality falls because, in relative terms, some nodes lose and others gain connectivity as IPs are connected to those with which they have strong payment ties, which not necessarily are the most central DPs (See video clip of Figure 3) .

on degree (all vertexes will have degree equal 1) or local clustering.

[15] This may not be the case if one allows for more than one correspondent relationship between IPs and DPs.

[16] These results depend on the tiering assumptions about which DP would be matched with IPs. It could be possible that instead of making second-tier participants clients of a sole DP they could enter into correspondent relationships with more than one DP or with the most central DPs, ending up with different measures of local degree and clustering. We abstract from this possibility in our analysis.

Figure 3: Evolution of the TN graph with tiering

The video clip is available at: `https://vimeo.com/159448680`. The video shows the evolution of the network of all participants as tiering increases, featuring a dense network of direct participants that evolves into a network of core direct participants that are hubs of multiple indirect participants.

However, as tiering approaches large DPs, playing central roles as correspondent institutions, it becomes more likely that they will be matched with correspondent hubs, increasing their centrality relatively to all other nodes, producing an increase in degree centralization. The fall in closeness, on the other hand, seems to be dominated by the fact that tiered institutions increase their distances as they are forced to interact with other nodes through a correspondent institution, as they do not seem to gain much closeness by having access to multiple nodes through correspondent hubs. Probably the reason is that the actual CUD (initial tiering scenario) is already very hierarchical[17] with central hubs enabling high levels of closeness for participants that are made second-tier participants along the tiering process.

Note that the notion of TCTF in a world of strong correspondent banking changes from the usual one targeted only to DPs in systemically important payment systems (as, for example, in León et al., 2013): you may have a low-connected, low-centralized TN, but with a DPN composed of TBTF, TCTF DPs and a hierarchical structure with a potentially highly vulnerable second-tier network of participants.[18]

5.2 Credit risk

We estimate correspondent institution credit risk to IP's default as the maximum intraday liquidity that those clients actually require to settle all transactions on a gross basis and in real time as DPs in the benchmark scenario, ML_i^o. In contrast to the DPs, which have to post collateral to access the repo intraday lines of credit of the central bank or get uncollateralized credit from other DPs, correspondent institutions usually grant their clients intraday overdraft facilities. Therefore, the maximum intraday liquidity needed by tiered participants in the benchmark scenario as DPs in CUD could be considered as an upper bound of the unsecured lines of credit these participants would need from their correspondent institutions if they become IPs.[19]

Figure 4 shows that, in average, a correspondent institution could be exposed to between COP$ 1 billion and COP$ 10 billions during normal times. Yet, the exposures could be between COP$ 200 billion and COP$ 1 trillion in times of tight liquidity. The latter figures could represent between 2.8% and 14.8% of the Tier I capital of the core credit institutions. Note however, that the building of credit risk as we increase tiering seems to be well contained up to scenarios 70th to 90th because no mayor players are tiered yet.

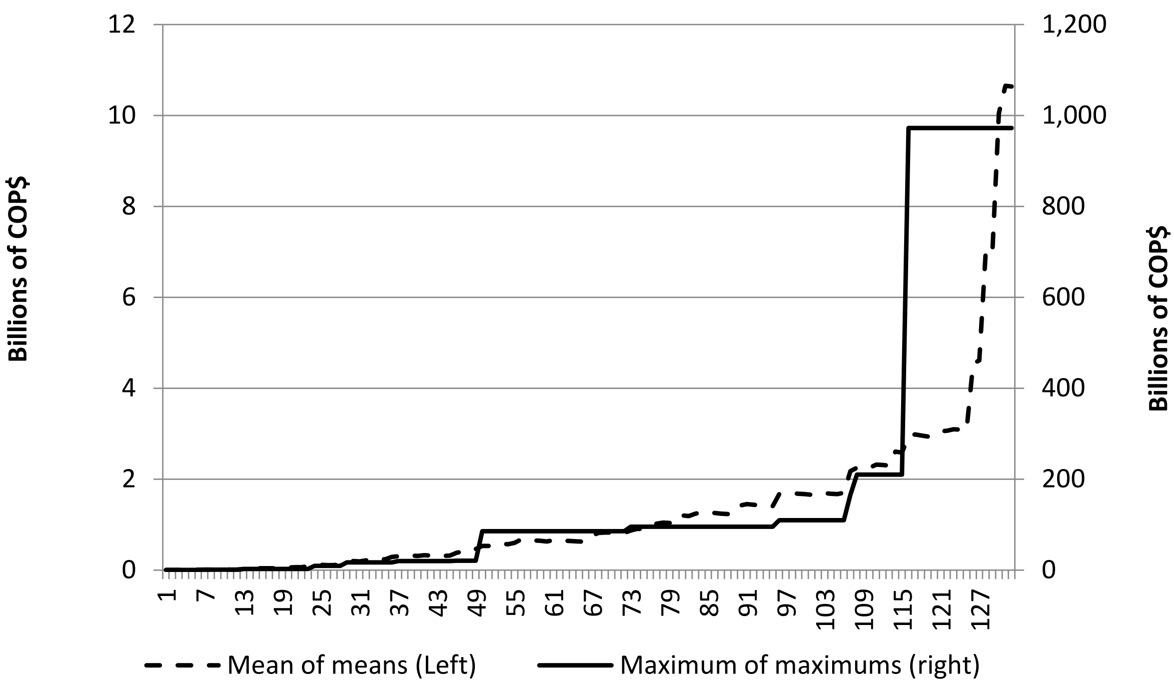
Table 2 shows that credit risk exposures faced by DPs increase with tiering as measured by the median of the potential loss distribution. It also shows that the loss distribution becomes more dispersed with tiering.

[17]See León et al. (2014).

[18]If the matching of DPs and IPs where based on which DPs in each tearing scenario are more central (e.g. higher degree), degree centralization may increase rather than decrease with tiering.

[19]Note that the notion of credit risk does not entail calculating the net payment flow positions of the client and its correspondent bank. It is the maximum intraday liquidity position needed by the tierable institution to fulfill its obligations with other participants, including its correspondent institution. Hence, it does not depend on whether the tierable institution has an overall positive or negative payment flow with the correspondent institution.

Figure 4: Potential overdrafts provided by correspondent institutions to their clients



Note: The figure provides estimates of possible overdrafts correspondent institutions may extend to their tiered clients. It is based on the calculation of the maximum liquidity needs tierable participants hold in the benchmark scenario ML_i^o. We calculate this daily figure for each tierable participant in the sample period to get means and maximums. Then, at each scenario, we calculate the mean-of-means and the maximum-of-maximums among the participants that have been tiered up to that point.

Table 2: Credit risk faced by correspondent institutions (selected scenarios, COP$ millions)

Quartil	Scenario 30 ML^o	Scenario 60 ML^o	Scenario 90 ML^o	Scenario 132 ML^o
Min	0.00	0.00	0.00	0.00
25%	1.49	6.04	14.64	30.99
50%	23.95	112.98	241.36	628.80
75%	205.02	876.87	1,309.03	3,810.32
Max	1,093.15	4,662.54	16,683.86	404,338.09

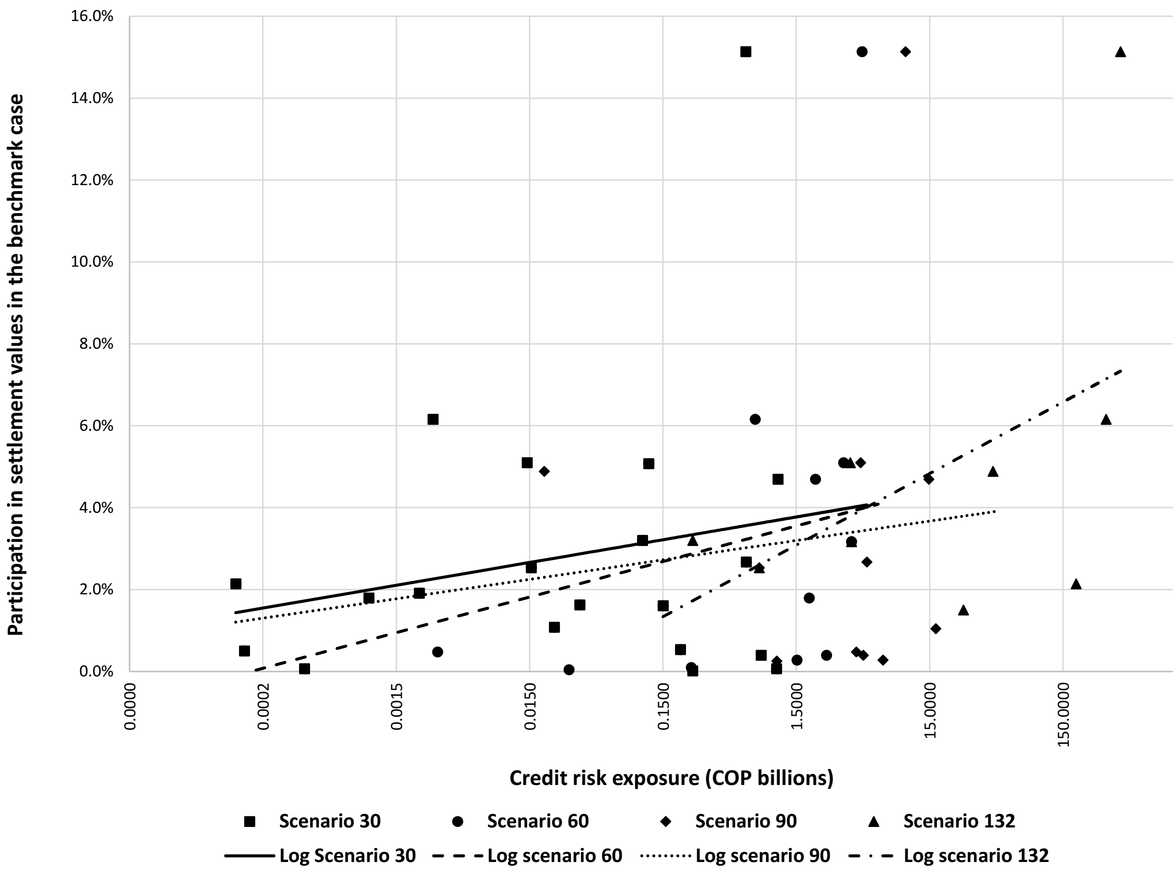
Note: The scenarios in the tables are chosen to illustrate how the distribution of credit risks faced by correspondent institutions as a result of clients' default evolves with tiering. We calculate the level of potential overdrafts provided by correspondent institutions to their clients at each tiering scenario, based on ML_i^o. Based on the maximum overdraft given to a client by a correspondent istitution, we get the distribution of overdrafts and the statistics in the table. Since the risk distributions at higher levels of tierning stochastically dominate those at lower levels, we only present selected scenarios.

Again, the table shows that a mayor degree of tiering could be done in CUD without significant impact on credit risk exposures as they are no more than COP$ 4.6 billion and the median is COP$ 112 millions when 60 institutions are made second-tier participants. These amounts represent no more than 0.5% of the Tier I capital own by correspondent institutions.[20] However, credit risk sharply increases as we tier the largest DPs (in terms of their payment shares) of our set of tierable institutions. Indeed the median potential overdraft granted by correspondent DPs more than triple as we move between scenarios 90 and 132, amounting to credit risks that could be as large as 9.9% of Tier I capital own by core DPs in normal times.

We also find weak evidence of the correlation between credit risk exposures and the correspondent DP's size. As Figure 5 shows, there is a positive correlation between the credit exposure of correspondent institutions and their participation in settlement values in the benchmark case. The estimated trends indicate a significant more-than-proportional growth in credit risk relative to DP's payment shares. The figure also indicates that this correlation becomes steeper as we increase tiering. In contrast with Lasaosa and Tudela (2008) findings for UK system CHAPS, this may indicate that in the Colombian case there is less room for economies of scales in the management of liquidity in correspondent banking. Yet, abstracting from the

[20]Based on financial institutions' balance sheets recorded at the Colombian central bank.

Figure 5: Credit risk and size of correspondent institutions



Note: The figure presents the credit risk exposure of correspondent institutions on overdraft lines-of-credit to their clients at four selected scenarios. The X axis is presented in logarithms and the lines are fitted logarithmic regressions on the observations for the set of correspondent institutions in each scenario.

outlier values of the largest DPs in the core group at the top right side of the figure, the relationship between credit exposures and DP's size would be less clear. Large credit exposures could be faced by relatively small correspondent institutions that may not carry enough capital buffer to stand a client´s default. In fact, in some of the small correspondent institutions, their exposure could be as high as 6.5% of their Tier I capital.

In addition, we estimate potential credit risks faced by clients in case of default of a correspondent institution ME_i. This is calculated as the maximum positive balance that a tierable candidate has as a DP in the benchmark scenario:

$$ME_i = (\max(0, \max(B_i))).$$

We take the client with the highest ME_i for each correspondent institution as a measure of the severity of the losses that a default by correspondent institutions can inflict upon individual indirect participants.[21].

Table 3 shows that client credit exposures also increase and the distribution of potential losses becomes more disperse with higher tiering levels. Surprisingly, the distributions of client risk exposures to correspondent institution's defaults stochastically dominate the distributions of potential losses that correspondent institutions face in case of client's defaults. This feature of tiering in the Colombian case is important as increasing indirect participation would make these institutions quite vulnerable to systemic risks derived from DPs' financial distress. In fact, these credit exposures could reach levels of 15.6%, and, in extreme cases, 119.3%, of the capital own by tiered institutions.

5.3 Liquidity savings

In an RTGS system, like CUD, participants require substantial liquidity to make payments. Such liquidity comes from reserve requirements, incoming payments, short term repo lines of credit available from the central bank, or uncollateralized inter-bank loans; the later two cases being the most costly because of the interest paid and, in the case of repo operations, the opportunity cost of the securities posted as collateral.[22]

To compute the liquidity savings of different degrees of tiering, we compare the ML_i needs of direct participants in the benchmark scenario with their ML_i calculated in each tiering scenario. Figure 6 shows the potential liquidity savings in the system when we increase the degree of tiering. Compared with the benchmark scenario, significant liquidity savings only start to show up after the 100th scenario and could reach a maximum of 13.4%. In contrast with the results in Lasaosa and Tudela (2008) and Arculus et al. (2012), where liquidity savings grow monotonically with higher tiering levels, in the Colombian case there is a significant reversal in liquidity savings starting at scenario 125. Furthermore, there are significant differences in the change of liquidity needs among the remaining core DPs as tiering increases.

Indeed, the same figure shows that some correspondent institutions could see their liquidity needs raise about 46% as early as at scenario 97. Whereas for others, the liquidity savings could be 33.8% compared with the benchmark case. At the peak of liquidity savings the average daily needs fall by COP$ 584.8 billions, which are equivalent to the average daily liquidity provided by the central bank via intraday repos in the sampled month.

Figure 7 also shows the relationship between liquidity needs and payment system activity. It shows that liquidity savings are a function of the type of institutions that become second-tier. As relatively small

[21]The default of a correspondent institution may have larger implications in a tiered system as it may affect simultaneously various of its clients and may produce domino effects in other direct and indirect participants. It is beyond the scope of this article to explore such risks.

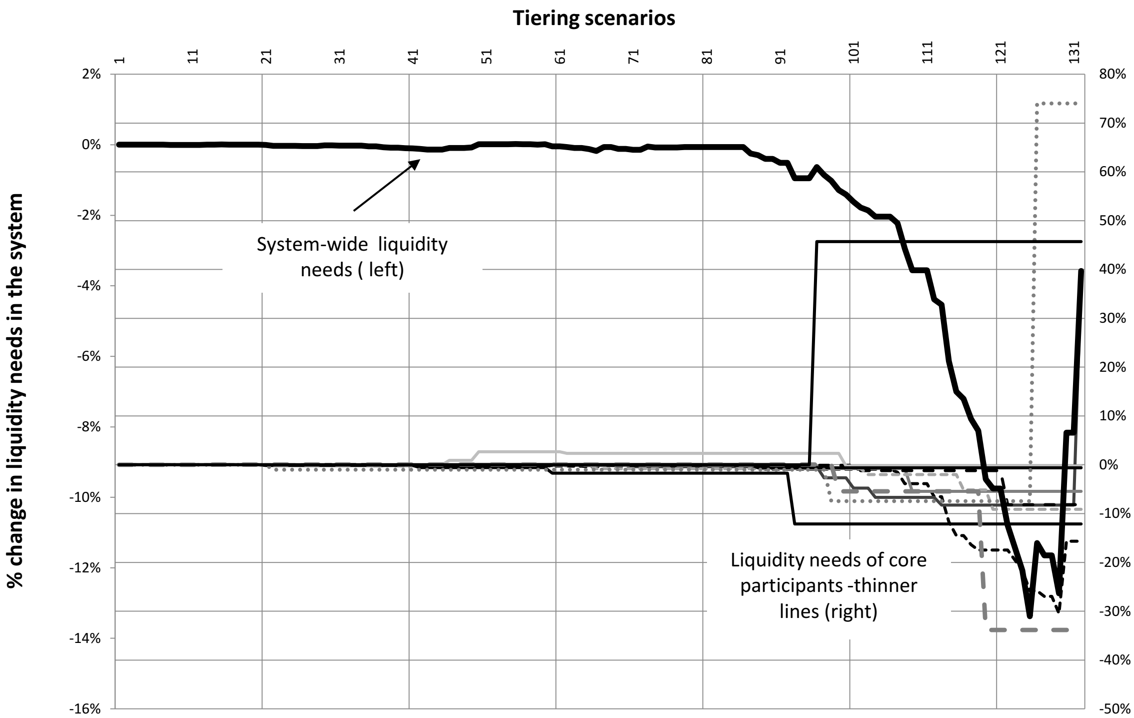
[22]See a detail discussion of these issues for CUD in Bernal-Ramirez et al. (2012).

Table 3: Credit risk faced by clients of correspondent institutions (selected scenarios, COP$ millions)

Quartil	**Scenario 30 ME**	**Scenario 60 ME**	**Scenario 90 ME**	**Scenario 132 ME**
Min	0.00	0.00	0.00	0.00
25%	44.57	139.09	428.13	1,208.56
50%	129.65	977.79	2,388.24	6,289.55
75%	391.92	2,365.90	6,426.37	33,021.24
Max	1,245.36	5,525.94	19,735.46	219,737.42

Note: The scenarios in the tables are chosen to illustrate how the distribution of credit risks faced by clients because of default by correspondent institutions evolves with tiering. We calculate the level of funds that an indirect participant may lose in the event of a default by its correspondent institution ME_i. Based on the maximum ME_i among the clients of a correspondent institution, we get a measure of how much its clientel may be exposed to tis default. From this calculation for each direct participant we get the distribution of losses by indirect participants and the statistics in the table.Since the risk distributions at higher levels of tierning stochastically dominate those at lower levels, we only present selected scenarios.

Figure 6: Liquidity needs by tiering scenario



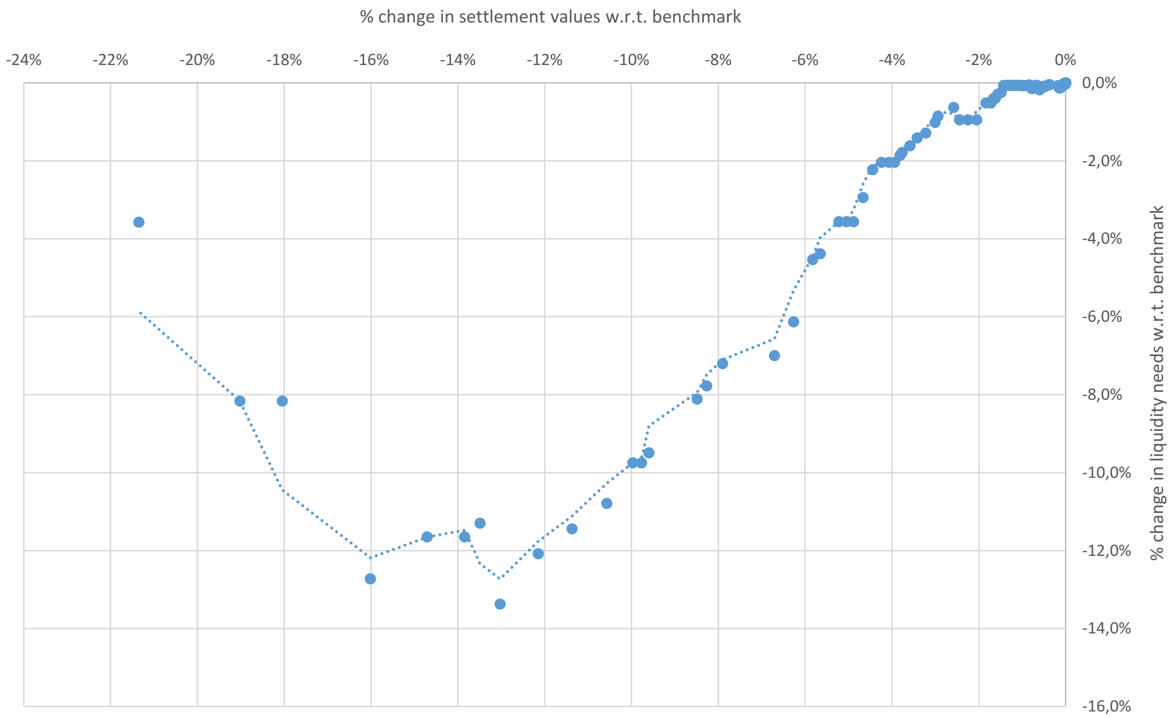
Note: System-wide liquidity savings are calculated as the difference between the sum of the average liquidity needs among direct participants in the benchmark scenario and the sum of their average liquidity needs in each tiering scenario. Labels for each line are not shown as the name or type of the institution can not be reveled.

participants in terms of their payment shares in the system are made IPs, the savings in liquidity are less than proportional to the decrease in settlement values. The relationship becomes proportional as we tier middle size participants but it turns out negative as the largest tierable participants become second-tier. This shows that caution should be exerted when doing extrapolations such as those in Lasaosa and Tudela (2008), as the relationship tends to be quite non-linear.

Disentangling the sources of liquidity savings is worth doing given the additional insights into using tiering as a LVPS efficiency-risk policy. As mentioned before, total liquidity savings are the result of internalization and pooling of payments received from clients by correspondent institutions. Savings from internalization come from the settlement in the correspondent institution's books of those payments due between the correspondent institution and its clients as well as those due between clients of the same correspondent institution. Savings from pooling come from correspondent institutions combining their payment flows with the payment flows that they conduct in the system on behalf of their clients. In this way correspondent institutions may fund more payments from receipts rather than using other more costly sources of liquidity.

As Lasaosa and Tudela (2008) and Arculus et al. (2012) explain, we can only get approximate estimates of pooling and internalization liquidity savings based on the methods used here. To do so, we run an additional

Figure 7: Settlement values and liquidity needs of correspondent institutions



simulation for each scenario where we eliminate all those payments susceptible of being internalized, i.e. those among DP's clients and between the clients and their correspondent institutions in the benchmark scenario. In addition, payments made between IPs and other DPs before tiering are made payment flows of correspondent institutions with other DPs in each scenario. With this new artificial set of payment flows we simulate intraday balances for each DP, B^b_{it}. The difference between the sum across DPs of their daily-average maximum liquidity-needs in the benchmark scenario $ML^a_i$and the sum of their daily-average maximum liquidity-needs calculated with the artificial set of payment flows $ML^b_i = \mid \min(0, \min_t(B^b_{it})) \mid$ provides us with an estimate of the total liquidity savings TLS, at a particular scenario.

To isolate pooling from internalization in the total liquidity savings we also simulate intraday balances, B^c_{it}, with a second set of artificial payment flows. In this case, we eliminate only those payments susceptible of being internalized, i.e. those among DP's clients and between the clients and their correspondent institutions before tiering. In other words, we allow internalization to affect a correspondent institution's intraday balances but not pooling. Therefore, the difference between the sum across DPs of their daily-average maximum liquidity-needs in the benchmark scenario ML^a_i and the sum of their daily-average maximum liquidity-needs calculated with the artificial set of payment flows $ML^c_i = \mid \min(0, \min_t(B^c_{it})) \mid$ gives us an estimate of liquidity savings due to internalization ILS, at a particular scenario. The difference between TLS and ILS is our estimate of liquidity savings due to pooling (Figure 8 presents an example of how the artificial payment flows are generated to calculate $ML^b_i$and ML^c_i).

We find that the liquidity savings from a more tiered payment system in Colombia come from both internalization and pooling. Note also that most of the non-monotonicity in liquidity savings starting at scenario 125 comes from the more stringent liquidity needs of DPs as a result of pooling their payments with those payment obligations their clients had with other DPs before being tiered. This is due mainly to the fact that as tiering increases both the size and interconnection of tired participants rise. As is common in LVPS, CUD is shown to have a hierarchical structure with large participants serving as hubs of connectivity in the network. Tiering small participants would bring liquidity savings to correspondent institutions as they are already part of the correspondent institution's hub. Yet, tiering more interconnected institutions may bring small liquidity gains to their correspondent institutions. Indeed, these corespondent institutions could inherit part of the liquidity needs that their clients had from their payment obligations with the rest of the direct network before being tiered.[23]

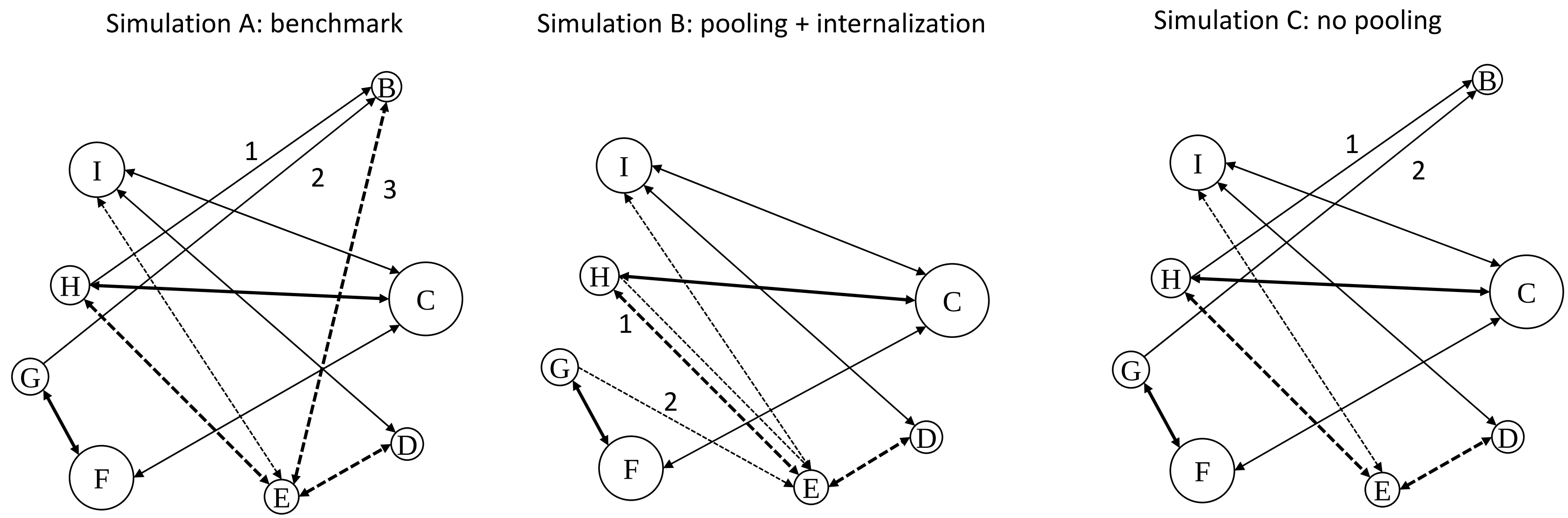
Figure 10 shows tiering brings winners and losers in terms of liquidity savings. Either if we look at total savings or savings due to internalization or pooling, there is a fair amount of DPs in some scenarios that would see their liquidity needs increase with tiering. Note that the number of DPs with higher stringent needs tend to be more diversified as we move through scenarios 1 to 61 or so, beyond which it become concentrated among fewer DPs. However, in most tiering scenarios those that enjoy liquidity savings both from internalization and pooling dominate those that suffer more stringent liquidity needs.

In contrast with other studies, the fact that some DPs actually may suffer an increase in their liquidity needs shows that the criteria to tier and match clients with correspondent institutions is critical to the results.

Figure 11 shows the case of a correspondent institution matched with one of its best clients in terms

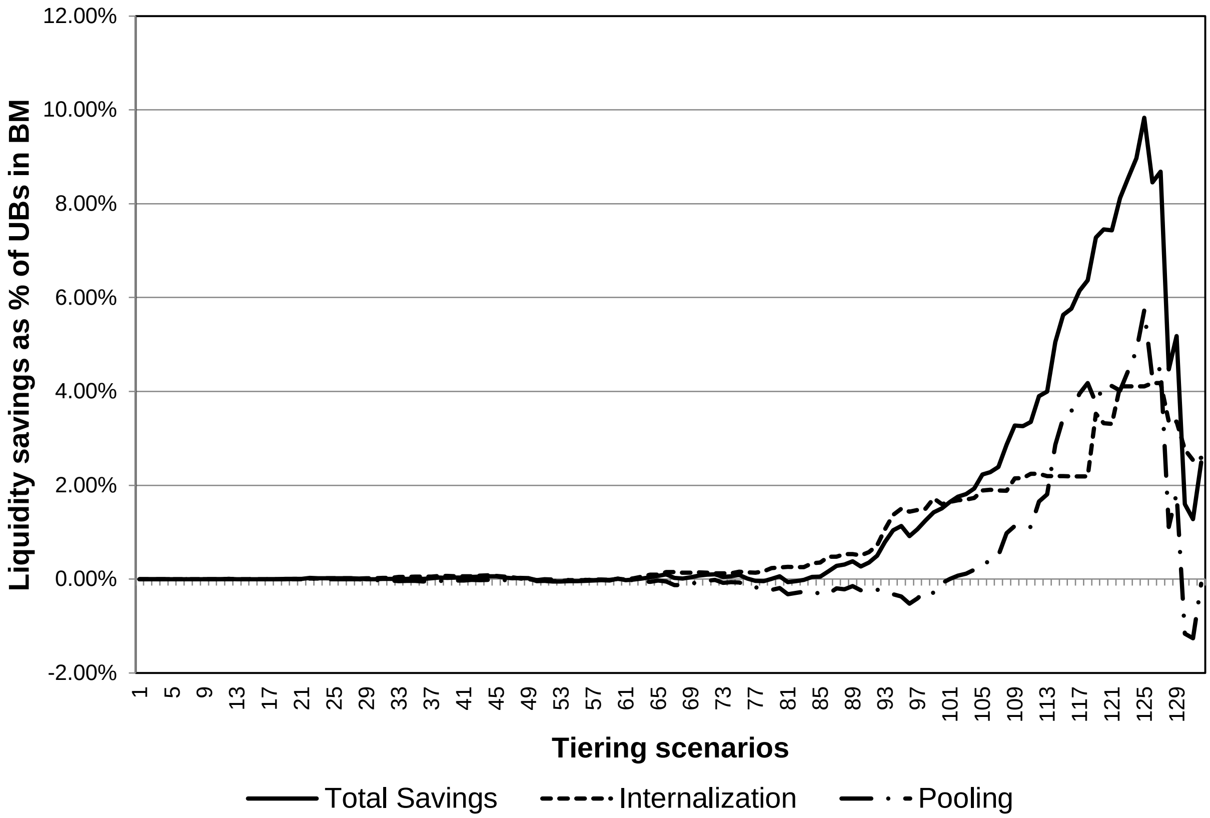
[23]The assumption that the timing of payments remains fixed through tiering scenarios may bias the gains from pooling downwards when a large proportion of the payment flows are discretionary rather than of high priority. In such case, correspondent institutions could rearrange the timing of their clients' payment obligations to better match their intraday liquidity needs, gaining higher liquidity savings from pooling than under the fixed-timing assumption. However, we believe the bias is moderated given that in the case of CUD, about 80% of the payments are of high priority (Cepeda and Ortega, 2015).

Figure 8: How artificial payment flows are created to calculate ML_i^b and ML_i^c



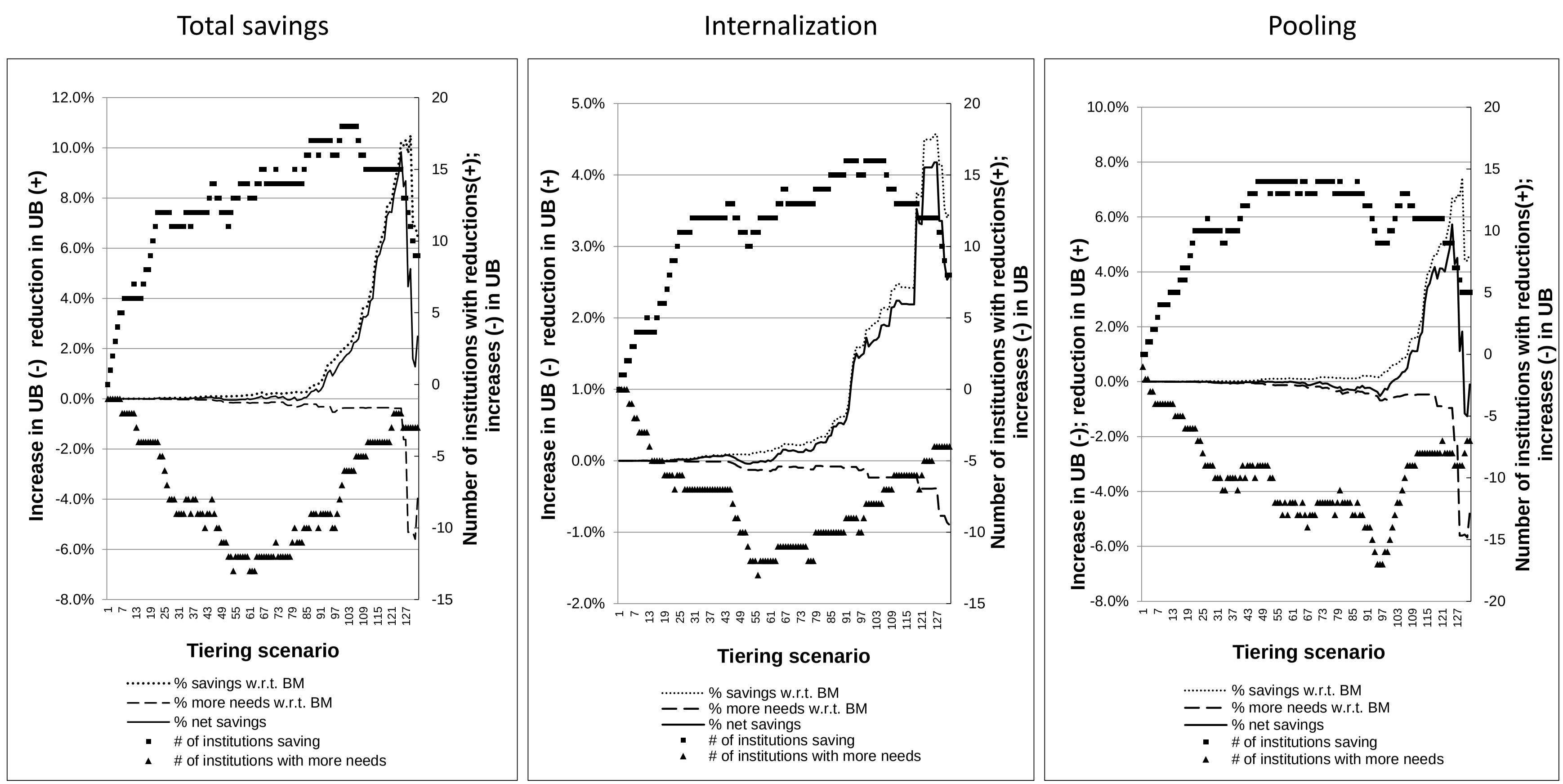
Note: The network example in the figure illustrates the manner in which payments are redefined to calculate $ML_i^b$and ML_i^c. The graph on the left side shows the payment flows before tiering. The dotted arrows are the intraday payment flows of direct participant E before tiering is made. These payment flows are the ones used to calculate his ML_i^a. Let direct participant B be the next to tier and E her correspondent institution. To calculate ML_i^bwe redefine payment flows 1 and 2 made from participants G and H to B in the left graph and make them E´s payments to G and H (made on B´s behalf), as shown by the dotted arrows in the central graph. Note that payment flows represented by 3 are not part of E´s intraday payments to the network (now settled in E´s books) and, therefore, are not considered in the calculation of ML_i^b. To calculate ML_i^cwe drop the bilateral payment flows between B and E prior to tiering (represented by 3 in the left graph) but leave payments 1 and 2 unaltered so that E faces internalization effects due to 3 but not pooling effects from payments 1 and 2.

Figure 9: Total liquidity savings and their pooling and internalization components (as percentage of the UBs in the benchmark case)



Note: UB holds for "upper bound" or maximum intraday liquidity needs ML_i as defined in the text. BM holds for "benchmark scenario." Each line is based on the sum of daily averages of intraday liquidity savings across direct participants in each scenario.

Figure 10: Winners and losers of a more tiered payment system by change in UBs



Note: The figures show winners and losers as we decompose the total changes in liquidity needs into internalization and pooling. The lines represent the aggregate change in liquidity needs among winners (the doted line), among lossers (the broken line) and the net savings (the solid line). UB holds for upper bound or maximum intradey liquidity needs ML_ias defined in the text. "w.r.t. BM" holds for "with respect to the benchmark scenario."

of gross payments before scenario 119 is run. In this case, the correspondent institution saves liquidity, as measured by its intraday ML, with strong pooling an internalization savings. In contrast, Figure 12 presents the case before scenario 125 where the correspondent institution inherits the strong liquidity needs of its client, ending up with higher liquidity needs. In this last case, the best match in terms of gross payment flows happens to be misleading with regards to the potential savings that could be achieved in a correspondent-banking relationship.

As Figure 13 shows the match between the client and the correspondent institution is favorable in scenario 119 as they had strong mutual payment flows with one another before tiering, with the correspondent institution being a net payer (liquidity savings due to internalization). Furthermore, the tiered participant had strong net inflows with other DPs before tiering which adds to the liquidity savings due to pooling. Yet, the match in scenario 119 brings higher liquidity needs to the correspondent institution, in spite of the savings from internalization, as it inherits the net payment outflows of its client with other DPs.

The results suggest that basing tiering simulations on gross payment flows (either based on total or discretionary criteria) among participants only, as done in the current literature, may not be the best way to approach potential liquidity savings and credit exposures of tiering. Matching of potential client-correspondent pairs should look more carefully at the adjacent matrix of payment flows among participants in terms of connectivity and net payment flows. Indeed, those DPs with net outgoing payments to a potential correspondent institution, when tiered, would not longer provide this liquidity necessarily in real time and therefore would be a drain on the correspondent's available liquidity. In such case, the internalization of payments makes liquidity needs more stringent for the correspondent financial institution. On the contrary, if gross payment flows are mostly going from the correspondent DP to its potential client the former may have liquidity savings from internalization of payments. Pooling liquidity savings, on the other hand, would depend on whether a client receives significant payment inflows from DPs, other than their correspondent financial institution, in the benchmark scenario. But, correspondent DPs may experience tighter liquidity needs if their clients have many payment outflows to other DPs. These considerations suggest the need to evaluate tiering trade-offs in LVPS under different assumptions about payment flows and business relationships using alternative modeling techniques such as those in Adams et al. (2010) and Galbiati and Soramäki (2012).

6 Conclusion

Fair and open access to systemically important financial infrastructures is one of the pillars of the principles set of by the BIS-CPSS committee (BIS-CPSS, 2012). In the case of large value payment systems (LVPS), it is more so as it allows different financial infrastructures to be interoperable by having a common set of financial market participants being able to clear and settle a variety of obligations coming from various trading environments. However, in the case of LVPS that operate on a real-time gross settlement basis, there are clear tradeoffs between credit risk and liquidity efficiency as broader access is allowed into the system.

A prospective participant has to weigh the extra fees and technological demands of becoming a member of a LVPS and the extra liquidity needs associated with settling payments on a real-time gross settlement basis, with the competitive advantage that being a DP brings in terms of the enhanced services DPs can provide to their clientele, coverage against the default of a correspondent financial institution and direct access to different facilities, such as central bank liquidity and direct participation in other critical financial infrastructures. This tradeoffs will determin the level of access to a LVPS. System-wise, there is also a well known tradeoff between risk and efficiency. Lower costs from tiering come at the expense of higher risks

Figure 11: Intraday balances of correspondent and client banks between the 118th and 119th tiering scenarios: saving case

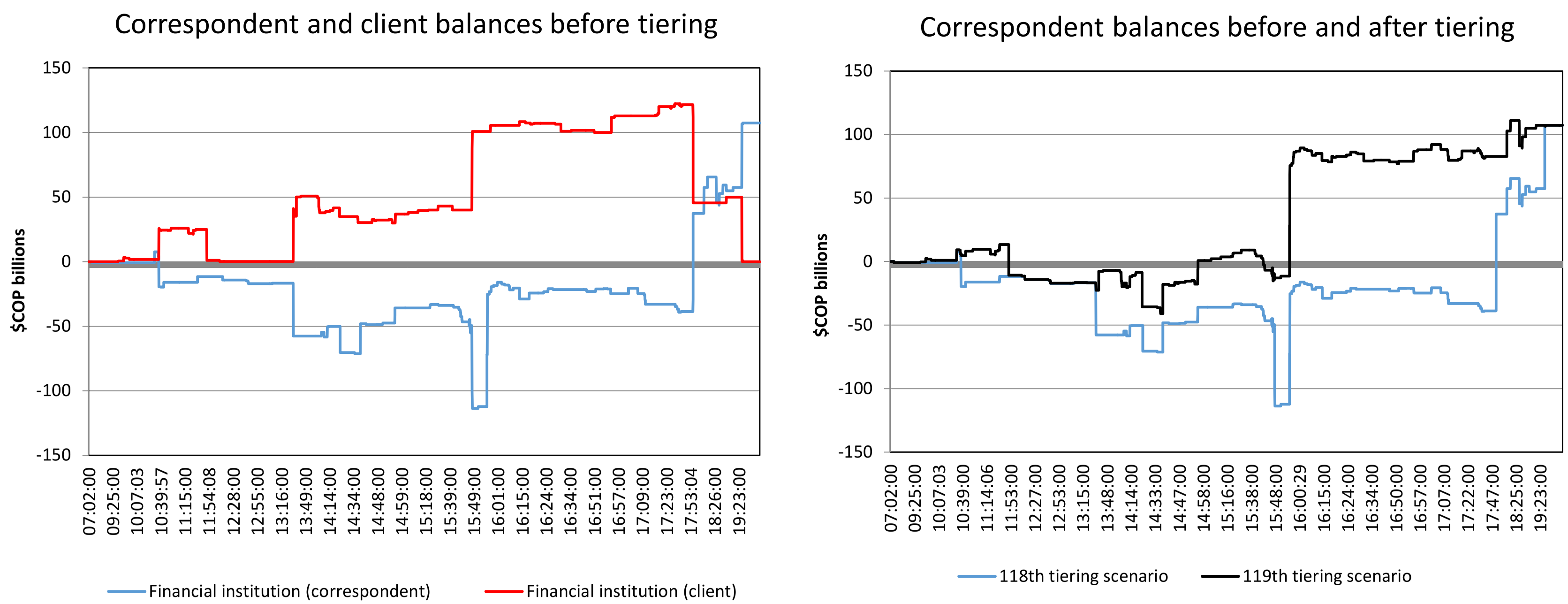


Figure 12: Intraday balances of correspondent and client banks between the 125th and 126th tiering scenarios: Stringent case

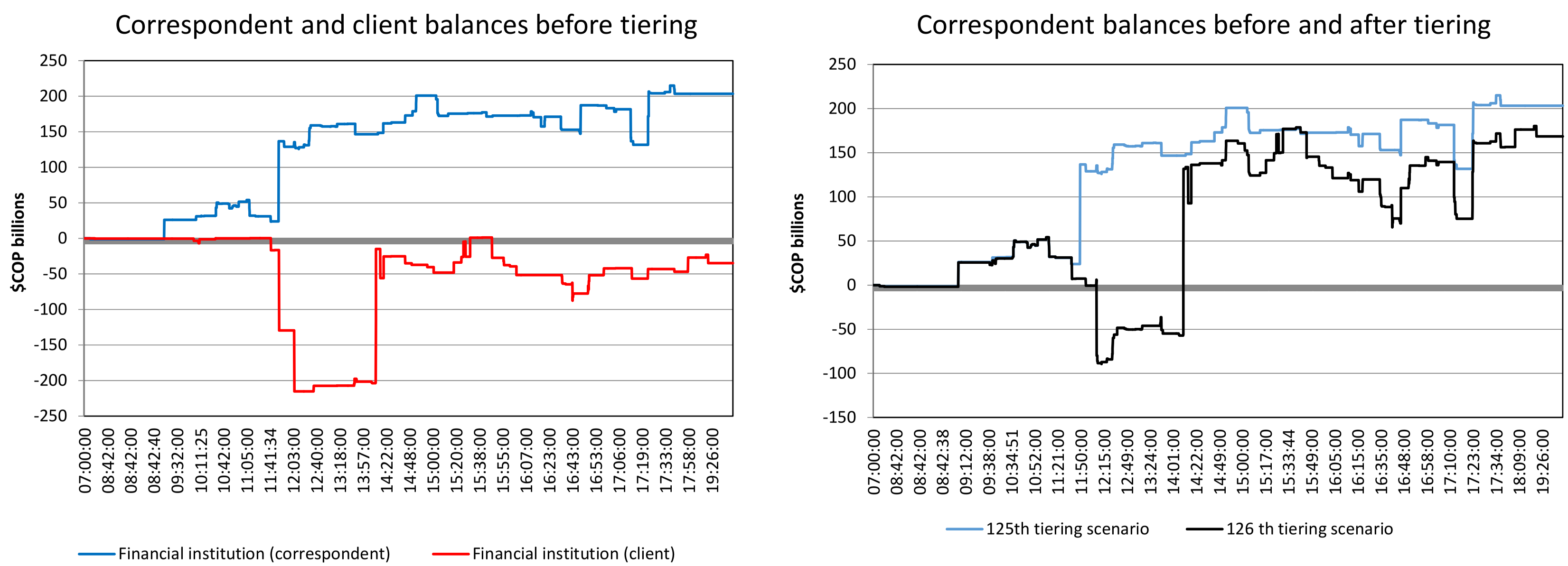
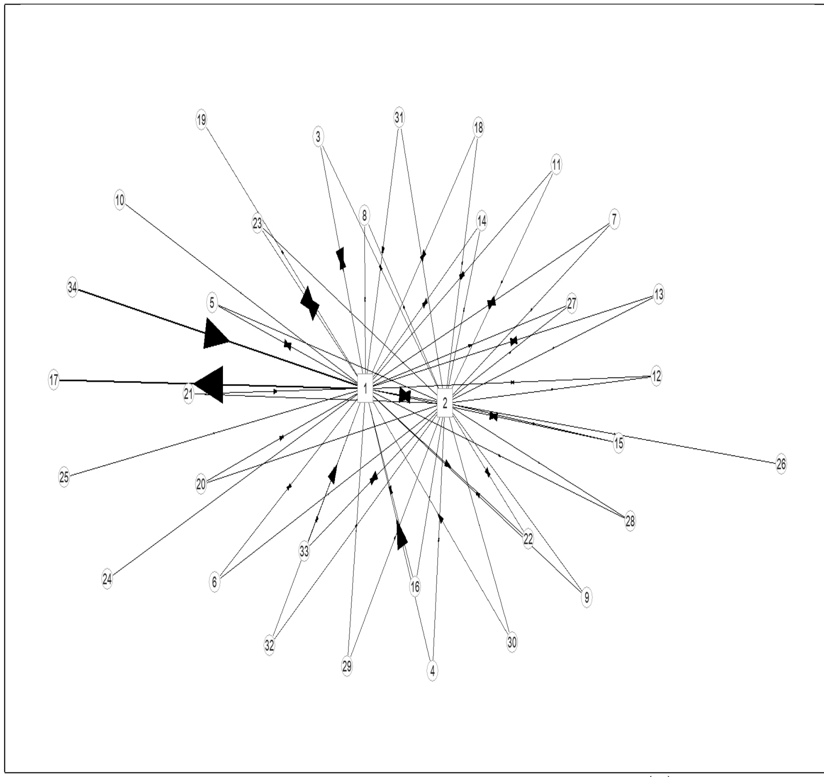
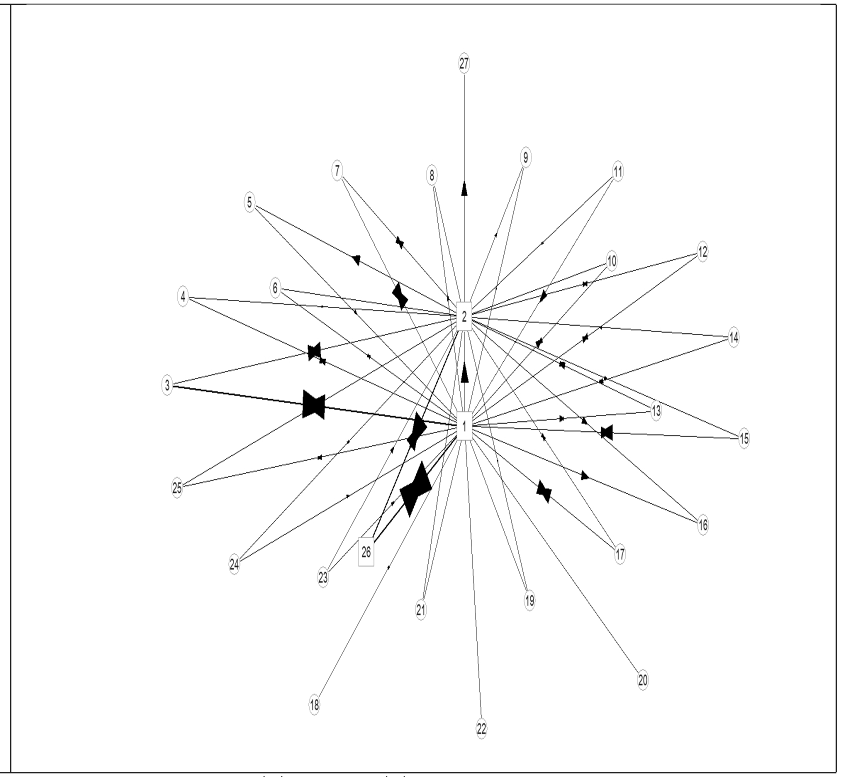


Figure 13: Client's and correspondent institution's payment flows before tiering scenarios 119 and 125





Note: The left graph plots the payment flows of client (2) and its correspondent institution (1) before (2) is tiered in scenario 119. The arrows in the middle of each edge represent the inflows and outflows of payments of (1) and (2) with the other DPs in the network. Their thickness represents the size of the payment inflows and outflows. Note the outflows from (1) to (2), which provide strong savings from internalization to (1). The right graph plots the case of client (2) and its correspondent institution (1) before scenario 125 is run. Note that in spite of the strong savings from internalization due to the outflows from (1) to (2), (1) inherits the strong net outflow position that (2) has with (26) inducing more stringent liquidity needs to (1) after tiering.

to the system. Therefore, LVPS operators face the question of what should be the optimal level of access compatible with social preferences.

Assuming that second-tier participants use correspondent services with direct participants with whom they had the strongest incoming and outgoing payment flows, our results show that increasing the tiering level in the Colombian payment system could be done without mayor changes in credit risk exposures for a large number of relatively small financial institutions (about 60 in terms of their gross payment flows in the system), yet without system-wide major liquidity savings. Credit risk, however builds up quickly as we tier larger direct participants reaching as much as 9.9% of Tier I capital own by core DPs in normal times.

In contrast, we show that the savings are not monotonically increasing with tiering, raising rapidly after tiering the 90th DP (out of 132 tierable institutions) and reaching a peak at the 120th tiering scenario when liquidity savings are about 13.4% of the liquidity needs currently used by the system. After that, additional tiering shrinks liquidity savings to only 3.6%.

Yet, in the Colombian case, the increase in liquidity needs as a result of a less tiered system may not be as stringent to participants due to the high levels of reserve requirements maintained by banking institutions at the central bank. Indeed, current maximum liquidity needs in CUD are about 54.7% of core participants' overnight balances at the central bank.

A more careful look at these results unveil the critical role that network relationships play in determining winners and losers in tiering. Our results warn against the belief that tiering would always bring liquidity savings and points at the importance of the nature of business relationships (e.g. financial groups, intraday liquidity usage and repo obligations) and payment flows between those being tiered and potential correspondent financial institutions. Indeed, future work should look into ways of finding optimal matches that would provide a liquidity savings frontier at different levels of tiering. One possible avenue could be extending the work of Adams et al. (2010) for tiering analysis, exploiting their optimal-matching algorithms based on potential pooling and internalization-liquidity-savings among participant pairs.

Our work also points at the need to look into possible network effects derived from tiering. As expected, concentration and connectivity in the LVPS increases with tiering as fewer direct participants remain in the system. However, the nature of the network in the LVPS may change significantly. Even thought, payment systems tent to be hierarchical with central participants working as liquidity hubs,[24] this array could be significantly changed through tiering an correspondent banking policies. Tiering also forces us to look into the entire payment network of second-tier and direct participants. As tiering increases, our simulations show the emerging of a low-connected, low-centralized total payment network (in the jargon of network theory), but with a cluster composed of several TBTF-TCTF direct participants and a hierarchical structure with a potentially highly vulnerable second-tier network of participants in a world of strong correspondent banking. These are some of the concerns that have moved the Bank of England to consider policies directed to increase participation in their LVPS, CHAPS (Finan et al. 2013).

References

Adams, M., Galbiati, M., and Giansante, S. (2010). Liquidity costs and tiering in large-value payment systems. Bank of England working papers 399, Bank of England.

[24]See León et al. (2014) who outline the hierarchical structure of the Colombian CUD.

Allen, L. and Saunders, A. (1986). The large-small bank dichotomy in the federal funds market. *Journal of Banking & Finance*, 10(2):219–230.

Arculus, R., Hancock, J., and Moran, G. (2012). The impact of payment system design on tiering incentives. RBA Research Discussion Papers rdp2012-06, Reserve Bank of Australia.

Assenza, S., Gutiérrez, R., Gómez-Gardeñes, J., Latora, V., and Boccaletti, S. (2011). Emergence of structural patterns out of synchronization in networks with competitive interactions. *Sci. Rep.*, 1.

Babus, A. and Hu, T.-W. (2015). Endogenous Intermediation in Over-the-Counter Markets. CEPR Discussion Papers 10708, C.E.P.R. Discussion Papers.

BdR (2015). Reporte de los sistema de pago. Technical report, Banco de la República de Colombia.

Bech, M. L. and Atalay, E. (2010). The topology of the federal funds market. *Physica A: Statistical Mechanics and its Applications*, 389(22):5223–5246.

Bernal-Ramirez, J. (2012). Access to central bank accounts and services for payment systems. In Summers, B., editor, *Payment Systems: Design, Governance and Oversight*, chapter 13. Central Banking Publications.

Bernal-Ramirez, J., Cepeda, F., and Ortega, F. (2012). Estimating the contribution of liquidity sources in the Colombian large-value real-time gross settlement payment system: A preliminary approach. *Journal of Payments Strategy & Systems*, 6(2).

BIS-CPSS (2005). New developments in large-value payment systems. Technical report, Bank for International Settlements.

BIS-CPSS (2012). Principles for financial market infrastructures. Technical report, Bank for International Settlements.

BIS-CPSS (2013). The role of central bank money in payment systems. Technical report, Bank for International Settlements.

Cepeda, F. and Ortega, F. (2015). A dynamic approach to intraday liquidity needs. *Journal of Financial Market Infrastructures*, 3(4).

Chapman, J., Chiu, J., and Molico, M. (2013). A Model of Tiered Settlement Networks. *Journal of Money, Credit and Banking*, 45(2-3):327–347.

Craig, B. and von Peter, G. (2014). Interbank tiering and money center banks. *Journal of Financial Intermediation*, 23(3):322–347.

Finan, K., Lasaosa, A., and Sunderland, J. (2013). Tiering in CHAPS. *Bank of England Quarterly Bulletin*, 53(4):371–378.

Flannery, M. J. (1996). Financial crises, payment system problems, and discount window lending. *Journal of Money, Credit and Banking*, 28(4):804–24.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, page 215.

Galbiati, M. and Soramäki, K. (2012). Clearing networks. *Journal of Economic Behavior & Organization*, 83(3):609–626.

Haldane, A. G. (2009). Rethinking the Financial Network.

Haldane, A. G. and May, R. M. (2011). Systemic risk in banking ecosystems. *Nature*, 469(7330):351–355.

Harrison, S., Lasaosa, A., and Merxe, T. (2005). Tiering in UK payment systems: credit risk implications. *Bank of England Financial Stability Review*, Bank of England.

Ho, T. S. Y. and Saunders, A. (1985). A Micro Model of the Federal Funds Market. *Journal of Finance*, 40(3):977–88.

Jackson, J. P. and Manning, M. J. (2007). Central Bank intraday collateral policy and implications for tiering in RTGS payment systems. DNB Working Papers 129, Netherlands Central Bank, Research Department.

Kahn, C., Quinn, S., and Roberds, W. (2016). Central banks and payment systems : the evolving trade-off between cost and risk. In Bordo, M. D., Øyvind Eitrheim, Flandreau, M., and Qvigstad, J. F., editors, *Central Banks at a Crossroads: what Can We Learn from History?* Cambridge University Press.

Kahn, C. M. and Roberds, W. (2009). Payments settlement: Tiering in private and public systems. *Journal of Money, Credit and Banking*, 41(5):855–884.

Lai, A., Chande, N., and O'Connor, S. (2006). Credit in a Tiered Payments System. Technical report.

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 papers 337, Bank of England.

Leinonen, H. and Soramäki, K. (1999). Optimizing Liquidity Usage and Settlement Speed in Payment Systems. Research Discussion Papers 16/1999, Bank of Finland.

León, C. and Berndsen, R. J. (2014). Rethinking financial stability: Challenges arising from financial networks' modular scale-free architecture. *Journal of Financial Stability*, 15(C):241–256.

León, C., Machado, C., and Murcia, A. (2013). Macro-prudential assessment of Colombian financial institutions' systemic importance. Borradores de Economia 800, Banco de la Republica de Colombia.

León, C., Machado, C., and Sarmiento, M. (2014). Identifying central bank liquidity super-spreaders in interbank funds networks. BORRADORES DE ECONOMIA 011187, Banco de la República.

Martin, A. and McAndrews, J. (2008). Liquidity-saving mechanisms. *Journal of Monetary Economics*, 55(3):554–567.

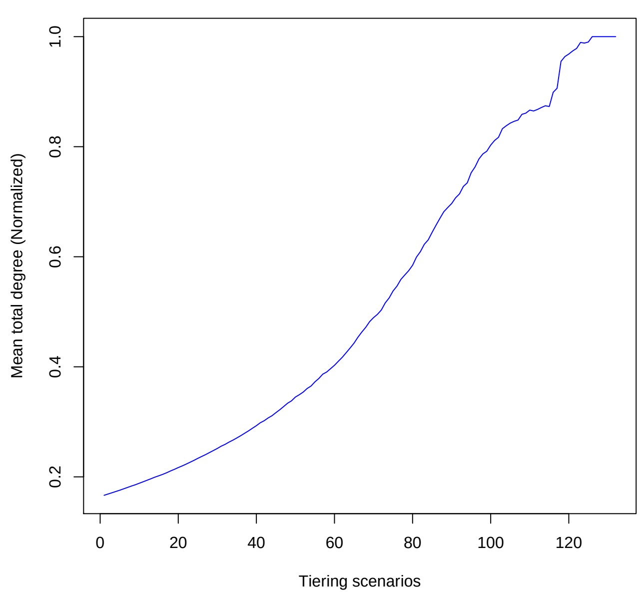
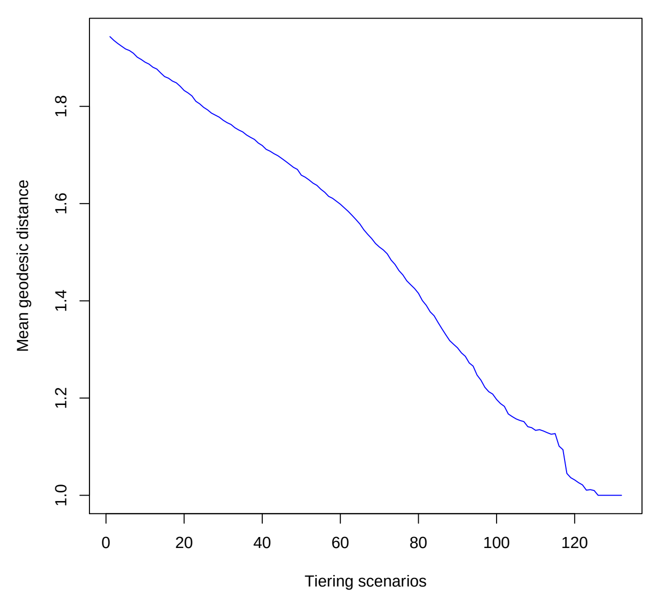
Newman, M. (2010). *Networks: an introduction.* Oxford University Press.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge University Press.

A Appendix

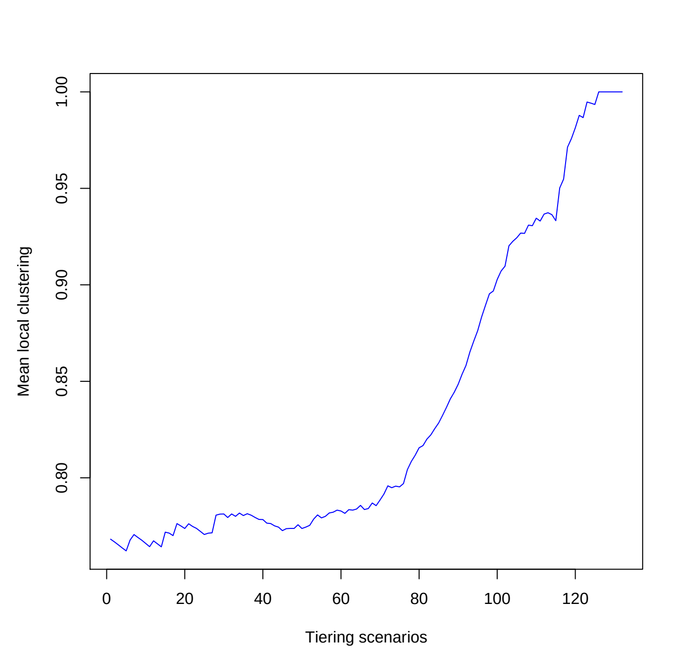
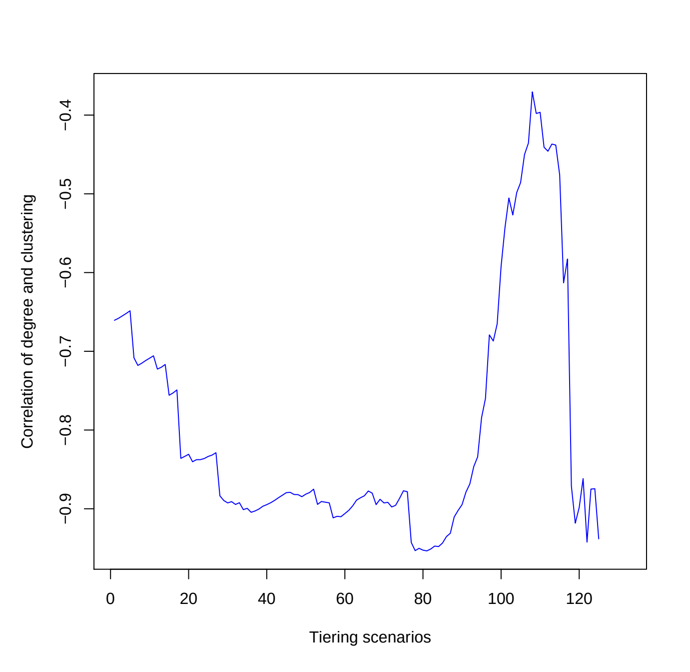
This appendix illustrates the evolution of some topological properties of networks when the level of tiering increases. Networks of direct participants (DPN) and total networks (TN), which includes both direct participants (DPs) and indirect participants (IPs), are analyzed. These payment networks assembled for each tiering scenario consist of a set of nodes (participants) and a set of edges between nodes (payment flows) that are the basis for different topological measures.

Figure A.1: DPN connectivity and tiering in terms of degree and geodicity

Normalized degree	Geodicity
	

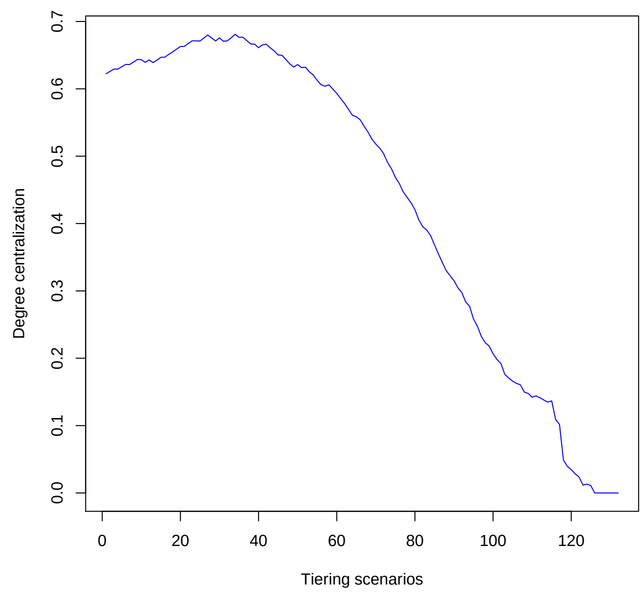
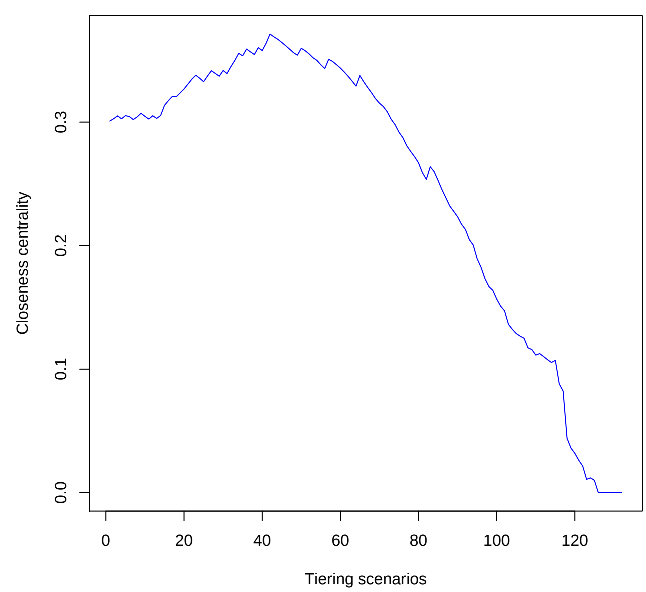
Note: The normalized degree measures how well connected are the nodes of a network. This topological metric is calculated as the average of the number of adjacent edges nodes in the system divided by the total number of possible edges. When the level of tiering increases, the DPN becomes more connected until reaching a “complete network” of core financial institutions fully connected. The geodesic distance of a network measures how far are nodes from each other in terms of the number of edges involved in a shortest path connecting them. Closeness between remaining direct participants grows, as the level of tiering increases.

Figure A.2: DPN local clustering, hierarchy and tiering

Local clustering	Hierarchy
	

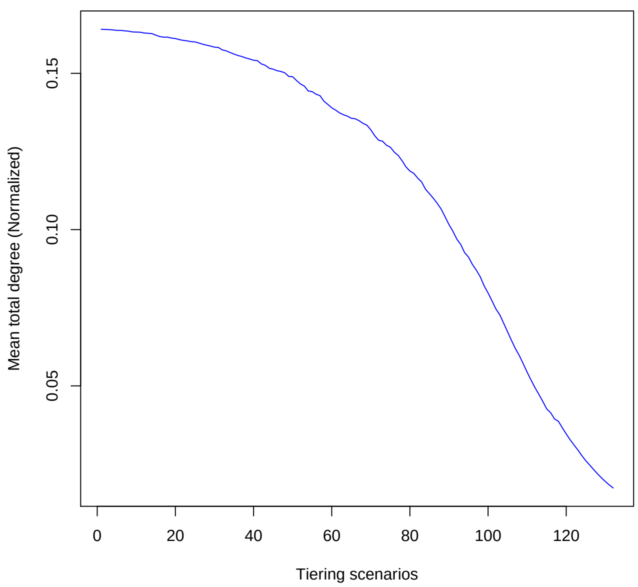
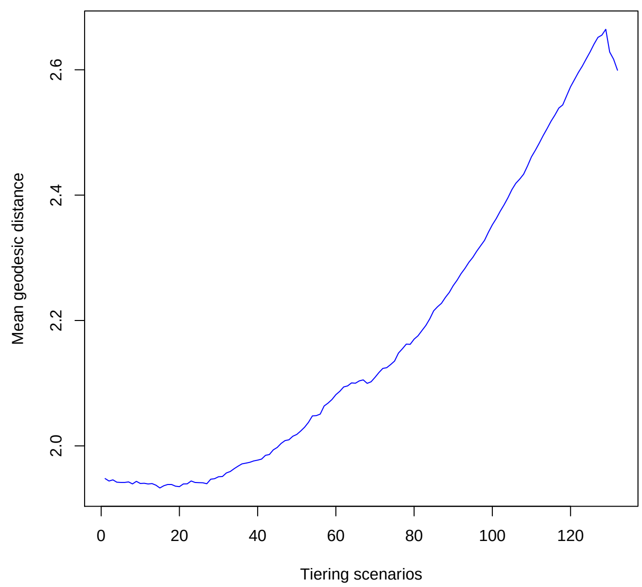
Note: The clustering coefficient reflects the transitivity of relations between nodes in a network. This metrics measures the average probability that two neighbors of a node are themselves neighbors. In the DPN, the increases in the level of tiering rises the frequency with which loops of length three between direct participants can appear in the network. The correlation between degree and clustering helps to identify the hierarchical structure across nodes in networks. A positive correlation indicates that high-degree (low-degree) vertices are connected to other high-degree (low-degree) nodes. A negative correlation points toward an inverse relation of high-degree (low-degree) vertices connected to other low-degree (high-degree) nodes. In the DNP, as tiering increases, the correlation describes a non-monotonic negative trend, with a window of scenarios where tiering may actually reduce the hierarchical level of the network.

Figure A.3: DPN centralization and tiering

Degree centralization	Closeness centralization
	

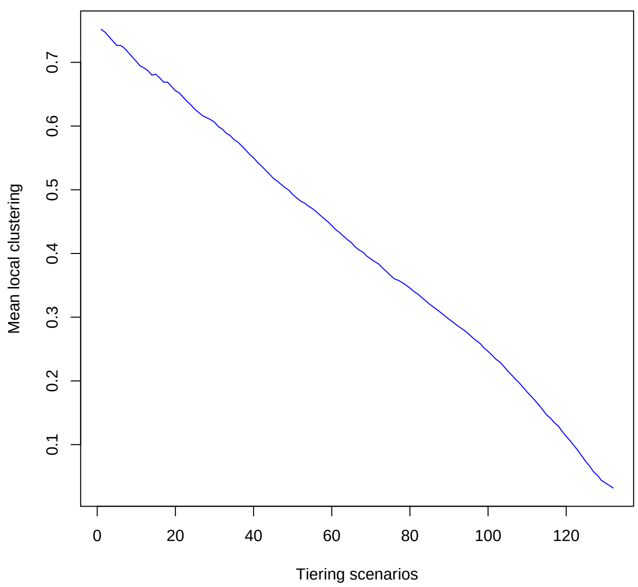
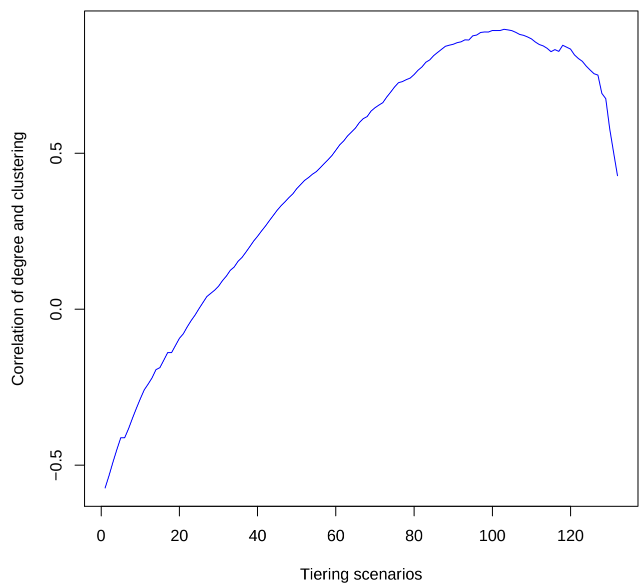
Note: Centralization is a concept that characterize an entire network. In a highly centralized network, there is a clear boundary between the center and the periphery. There are several ways of measuring centralization, among others, by degree and closeness. Degree centralization corresponds to the variation in the degrees of nodes divided by the maximum degree variation that is possible in a network of the same size. Closeness centralization is the variation in the closeness centrality of vertices divided by the maximum variation in closeness centrality scores possible in a network of the same size. Centralization, measured in terms of degree and closeness, reveals a DPN with lower levels of network centralization as tiering increases but with a hump shape.

Figure A.4: TN connectivity and tiering in terms of degree and geodicity

Normalized degree	Geodicity
	

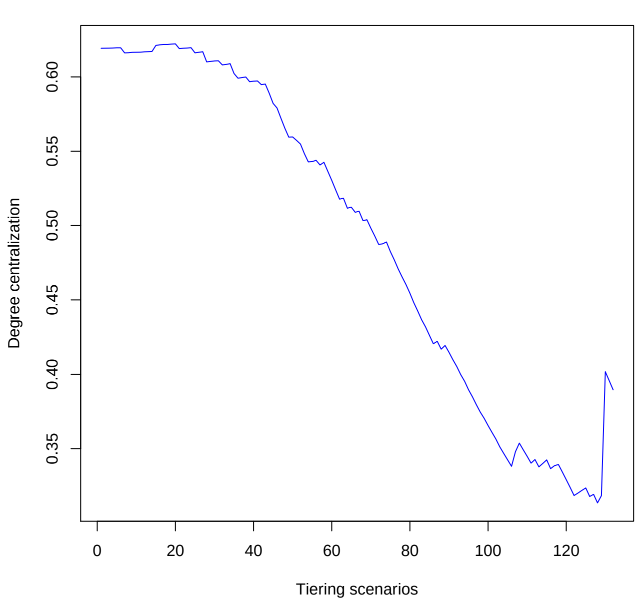
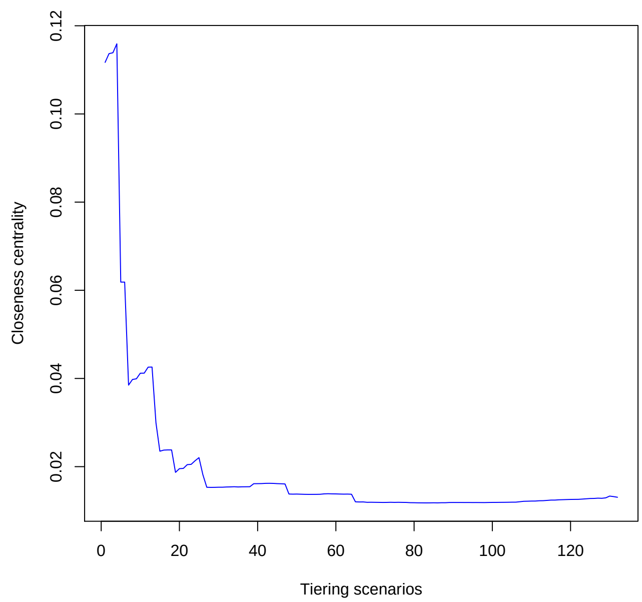
Note: The measures here are the same as in Figure A.1 but calculated for the total network (TN).

Figure A.5: TN local clustering, hierarchy and tiering

Local clustering	Hierarchy
	

Note: The measures here are the same as in Figure A.2 but calculated for the TN.

Figure A.6: TN centralization and tiering

Degree centralization	Closeness centralization
	

Note: The measures here are the same as in Figure A.3 but calculated for the TN.


BANCO DE LA REPÚBLICA
COLOMBIA