6-2023

Catching the fast payments trend: Optimal designs and leadership strategies of retail payment and settlement systems

Zhiling GUO
Singapore Management University, ZHILINGGUO@smu.edu.sg

Dan MA
Singapore Management University, madan@smu.edu.sg

Citation

GUO, Zhiling and Dan MA. Catching the fast payments trend: Optimal designs and leadership strategies of retail payment and settlement systems. (2023). *MIS Quarterly*. 47, (2), 669-704.

Available at: https://ink.library.smu.edu.sg/sis_research/7891

MIS Quarterly

RESEARCH ARTICLE

CATCHING THE FAST PAYMENTS TREND: OPTIMAL DESIGNS AND LEADERSHIP STRATEGIES OF RETAIL PAYMENT AND SETTLEMENT SYSTEMS[1]

Zhiling Guo and Dan Ma
School of Computing and Information Systems, Singapore Management University,
SINGAPORE {zhilingguo@smu.edu.sg} {madan@smu.edu.sg}

Recent financial technologies have enabled fast payments and are reshaping retail payment and settlement systems globally. We developed an analytical model to study the optimal design of a new retail payment system in terms of settlement speed and system capability under both bank and fintech firm heterogeneous participation incentives. We found that three types of payment systems emerge as equilibrium outcomes: batch retail (BR), expedited retail (ER), and real-time retail (RR) payment systems. Although the base value of the payment service positively affects both settlement speed and system capability, the expected liquidity cost negatively impacts settlement speed, and total transaction volume and technological effectiveness positively impact system capability. We identified three leadership strategies to maximize social welfare: the government mandate (GM) strategy, the fintech-inclusive (FI) strategy, and the fintech-exclusive and bank-exclusive (FE+BE) strategy. When the base value of the payment service is either low or high, the GM strategy leads to a socially optimal unified system. When the base value of the payment service is in the intermediate range, and if fintech firms have a significant technological advantage over banks, then both GM and FI strategies result in a socially optimal unified system; otherwise, the FE+BE strategy results in a socially optimal fragmented system. Further, we proposed a Shapley-value-based cost-sharing rule under the GM strategy to fairly allocate social welfare among the system participants. Our findings offer important policy insights into the optimal system designs, leadership strategies, and the government regulator's role in shaping future innovations in the payments industry.

Keywords: Fast payments, fintech, retail payment and settlement, analytical modeling, mechanism design

Introduction

The rapid growth of the internet and mobile commerce and the adoption of digital payment services across all market segments have led to a boom in the volume of retail payments. The *World Payments Report* documents an unprecedented eight years of double-digit growth for global noncash payments transaction volume, which increased to 725 billion in 2020, of which 605 billion was from retail payment transactions (Capgemini Research Institute, 2021). Several factors are driving this tremendous growth trend—the growing adoption of mobile payments, uptake in contactless technology, and digital innovation from giant technology players and card companies.

In addition to payment innovations, such as Square, Apple Pay, AliPay, and PayPal, which make transactions convenient and easy, some new financial technology (fintech) solutions are quickly gaining market momentum. For example, Ripple uses blockchain to process and secure its RippleNet payment network, making funds transfer faster, simpler, and more secure. The rapidly changing landscape in the payments industry has led consumers and businesses to expect fast payments to meet their needs (Bech et al., 2017).

Fast payment systems should process payments in real time, 24 hours a day, 365 days a year, and should perform an immediate clearing function between the payment service

[1] H. Raghav Rao was the accepting senior editor for this paper. Hong Guo served as the associate editor.

providers (PSPs) of the payer and payee (CPMI, 2016). The clearing function is the first leg of payment processing, which can be handled using the payment networks of customer-facing PSPs, such as financial institutions (e.g., bank card transactions) or technology companies (e.g., Apple Pay). The second leg of payment processing is the settlement of funds between PSPs that use their correspondent banks to settle payments in central bank money on a centralized infrastructure. However, such a settlement function does not necessarily occur immediately. Most countries today still rely on automated clearinghouses to settle retail payments. These legacy systems are called *batch retail* (BR) payment systems and usually adopt end-of-day settlement, resulting in significant payment delays. The disparity between immediate clearing and delayed settlement raises some system design concerns. In addition, rising retail volume, the entry of nonbank competitors, and the emergence of alternative fintech payment solutions present significant challenges to traditional retail payment settlement practices.

To defend their dominant market position and remain competitive in the changing payments game, banks are actively seeking to implement new technologies or collaborate with fintech firms to expedite payment processing. New types of systems, such as *expedited retail* (ER) payment systems, which settle payments multiple times per day, or *real-time retail* (RR) payment systems, which settle payments continuously and instantly, have emerged in response to fast payment initiatives. PSPs utilize card networks (e.g., Visa or Mastercard) or mobile platforms (e.g., Alipay or Apple Pay) for fast retail payment clearing and then connect to the country's backend core payment infrastructure for the final settlement of funds.[2] An early example of an ER system is the U.K.'s Fast Payment Service (FPS), which was launched in 2008 in an effort to reduce payment times between the customer accounts of different banks. A recent example of an RR system is Australia's New Payment Platform (NPP), completed in 2018, for real-time, low-value retail payments. (Please refer to Table A1 in Appendix A for a comparison of the BR, ER, and RR systems and their respective features.)

To date, different jurisdictions have taken different approaches to renewing their core payment systems and payment infrastructure. (Please refer to Table A2 in Appendix A for a summary of representative systems in different countries.) These systems represent brand-new retail payment infrastructures designed to enable faster and more efficient retail payment processing and settlement and to support state-of-the-art payment features, such as mobile payments (Tompkins & Olivares, 2016). In some cases, policy makers (e.g., central banks) take charge, regulate system development, and coordinate PSPs to guide new payment and settlement innovation. For example, in China, the Internet Banking Payments System (IBPS), which offers daily periodic retail payment settlements (an ER system design), is regulated by the People's Bank of China and operated by the China National Clearing Centre. In some countries, private PSPs, either major bank associations or fintech firms' consortiums, form their own partnerships to respond to the need for change. For example, in Sweden, Bankgirot, which is jointly owned by several major Swedish banks, is the only clearinghouse for retail payments. In 2012, under competitive pressure from nonbank payment providers in the industry, Bankgirot developed a mobile payment solution, Swish, and launched a new payment system, Payments in Real Time.[3] In Singapore, the Monetary Authority of Singapore (MAS) granted digital full bank licenses to the Grab-Singtel fintech consortium, which began providing retail customer banking services in August 2022.[4] M-Pesa, a mobile payment system operated by the two largest mobile network operators in Kenya, has achieved high levels of usage and successfully expanded in Africa. If multiple private PSPs (bank associations or fintech consortiums) build new systems independently and adopt an *exclusive strategy* for their own use only, this leads to a fragmented-system outcome. If one PSP leads innovation by building the new system and adopting an *inclusive strategy,* thereby attracting the use of others, it results in a unified-system outcome. As such, examining under what conditions these outcomes might arise would be interesting.

As previously explained, although fast payments require immediate payment clearing, the settlement function can either occur in real time or be deferred. Thus, a key infrastructure design feature of the new payment system is settlement speed. Historically, only time-critical, large-value interbank payments have been settled in RR systems using the *real-time gross settlement* (RTGS) mode, which imposed the highest liquidity requirements (Leinonen & Soramäki, 1999; Manning et al., 2009). In contrast, low-value retail payments have traditionally been processed by BR payment systems using the end-of-day *deferred net settlement* (DNS) to conserve liquidity. Moving from historical BR processing to a modern ER or RR retail payment system means that

[2] A core payment infrastructure refers to one that (1) includes at least the clearing and settlement of funds, for which the settlement occurs in central bank money, and (2) is central to the efficiency and stability of the financial system and economy. This paper focuses on the national retail payment system's core payment infrastructure's design and innovation.

[3] http://www.autogiro.se/en/about-bankgirot/about-us/bankgirots-history/

[4] https://www.reuters.com/business/finance/grab-led-gxs-rolls-out-singapores-first-digital-bank-2022-08-31/

PSPs need to have more liquidity available to support the fast movement of funds, which imposes high liquidity pressure and increases the operating costs for PSPs. Thus, the optimal fast payment system design is a trade-off between the benefit of accelerating payment settlements and the cost of handling increased liquidity needs.

In addition to settlement speed, system capabilities, such as 24x7x365 availability, adoption of the ISO20022 standard, support for state-of-the-art payment solutions (e.g., B2B, P2P, and mobile), and interaction with other databases and systems, represent another key design factor that captures the salient features and functionality of a new payment system. Table A3 in Appendix A lists major capabilities and attributes that should be considered in an advanced retail payment system.

In this research, we study how PSPs, including major banks and fintech firms, should design their core retail payment and settlement systems in response to the fast payments trend. To capture the heterogeneity among PSPs, we consider banks and fintech firms that have different transaction volumes and expected liquidity costs, as well as different levels of technological efficiency in building the new infrastructure. We aim to answer the following research questions: (1) What are the optimal system design configurations in terms of settlement speed and system capability to support fast payments? (2) What are banks' and fintech firms' participation incentives, leadership strategies, and market equilibrium outcomes? (3) From a social planner's perspective, what constitutes the socially optimal design, what roles should the government play, and what are the policy implications for next-generation retail payment system innovation?

We identify three payment system configurations in equilibrium: a BR system with traditional DNS settlement, an ER system with expedited settlement, and an RR system with RTGS settlement. In general, when the base value of a payment service increases, a faster payment settlement system with richer features and functionality is preferred by all PSPs. Moreover, the expected liquidity cost negatively impacts the choice of settlement speed, and both total transaction volume and technological effectiveness are crucial to enable high system capabilities.

Further, we show three socially optimal equilibrium leadership strategies: the *government mandate* (GM) strategy, the *fintech-inclusive* (FI) strategy, and the *fintech-exclusive and bank-exclusive* (EE+BE) strategy. When the base value of a payment service is either very small or very large, we recommend the GM strategy, under which governments should play a leading role in orchestrating new payment system development. In such a scenario, banks' and fintech firms' heterogeneous preferences for settlement speed have only marginal impacts on the optimal system design choice, whereas the benefits of a unified system are highly appreciated. Both PSPs have the incentive to build the new system for the other party to use, which would lead to inefficient double investments in technology and possibly an inferior equilibrium to emerge. Therefore, government mandates would dominate private initiatives to reach a socially optimal outcome and to ensure fair cost and benefit distribution.

When the base value of the payment service is in the intermediate range and fintech firms have a significant technological advantage over banks, the FI strategy under which fintech firms build a unified system to be used by all PSPs is socially optimal. It yields the same system design configurations as a central planner's system. However, as innovation leaders, fintech firms gain all innovation benefits, whereas banks obtain only their reservation value. Thus, we propose that, in this case, governments should play a coordinating role and mandate Shapley-value-based cost sharing among the PSPs to ensure a fair allocation of social welfare, promoting a transparent environment, and opening up the opportunity for future collaboration in new payment system innovation.

Finally, when the base value of the payment service is in the intermediate range and fintech firms do not possess a significant technological advantage, the FE+BE strategy under which fintech firms and banks independently build their own systems is recommended. In this situation, the disparity in the desired settlement speed between the two types of PSPs is too large to justify a unified system. Thus, we suggest that governments not intervene in private initiatives but allow PSPs to develop their systems and satisfy their respective unique needs. The resulting fragmented market with the coexistence of multiple payment systems is shown to be socially optimal. This finding is consistent with our observation that in current practice, multiple retail payment systems coexist in some countries.

We organize the rest of the paper as follows. The next section provides the relevant literature review. Then, we describe our model setup. Thereafter, we formulate different PSP decision-making problems and analyze the optimal system designs under both fintech firms' and banks' leadership strategies. Next, we present the market equilibria and propose government coordination to achieve socially optimal solutions. Finally, we examine key factors driving different equilibrium outcomes and discuss policy implications and then conclude the paper with directions for future research.

Related Literature

Several streams of literature are closely related to our work: the design of payment and settlement systems, the economic value and risk of fintech innovations, and the technology adoption research. We briefly review the related literature in each of these areas.

Core interbank payment systems are usually classified as either "large-value" or "retail" payment systems, depending on the primary types of transactions (CPSS, 2011, 2012). Because large-value payment systems play a key role in the economy, many central banks have adopted RTGS systems (e.g., Fedwire), which eliminates the credit risk between participants by allowing for the final and irrevocable settlement of each payment (Kahn & Roberds, 2001). Because each payment must be settled on a gross basis, RTGS systems are very demanding in terms of liquidity. To conserve liquidity, small-value retail payments are traditionally processed using DNS systems (e.g., the automated clearinghouse system), in which payments are settled periodically on a net basis (Johnson et al., 2004). A survey by the Boston Consulting Group (2012) shows that financial market participants often list slow settlement time as a key concern. Recently, the exponential growth in global retail volume and value has called for a transformation in the design of retail payment and settlement systems (Tompkins & Olivares, 2016).

Several recent studies have provided theoretical guidance for practice and have addressed payment and settlement system design issues. Khapko and Zoican (2020) claimed that the current intermediated security settlement process features several days of delay and is thus inconsistent with the fast-paced market, whereas immediate settlement overemphasizes counterparty risk and leads to suboptimal liquidity requirements for banks. They propose a "smart settlement" design, which allows counterparties to determine flexible time-to-settlement on a trade-by-trade basis. Guo et al. (2015) recognized the drawbacks of both the DNS and RTGS designs and proposed a hybrid faster payment settlement system that relies on a centrally managed priority queue to trade off the benefits of fast settlements against the increased liquidity pressure associated with increased settlement speed.

In addition to bank-dominated payment and settlement infrastructure, emerging fintech innovations such as blockchain demonstrate the potential of transforming the entire payment, clearing, and settlement process (Mills et al., 2016). Chiu and Keoppl (2019) studied the feasibility and optimal design of a permissionless blockchain-based settlement system, showing that the main advantage of such a decentralized payment settlement system is to accelerate settlement. Moreover, Wall and Malm (2016) found that the use of smart contracts can mitigate settlement risk and that an appropriate incentive mechanism design can help avoid dishonest mining behavior and ensure transaction safety.

Recent fintech studies have focused on understanding the economic value and risk of innovative technologies and their disruptive impact on traditional financial institutions (Ancri, 2016; Goldstein et al., 2019; Cong & He, 2019). Allison (2016) evaluated banks' adoption of blockchain technology and its technological effectiveness in facilitating the trade of debt instruments. Bohme et al. (2015), Evans (2016), and Maloumby-Baka and Kingombe (2016) examined the adoption of Bitcoin in remittance services from a cost-benefit perspective and evaluated its competitive effects on banks in this market. This stream of research suggests that the benefits of blockchain technologies could outweigh the associated costs and hence potentially threaten traditional bank services. Our research contributes to this literature by improving the understanding of both fintech firms' and banks' strategic reactions to the transformation of future payment and settlement infrastructure design.

Finally, our paper relates to the broader literature on technology adoption. Katz and Shapiro (1986), Choi and Thum (1998), and Farzin et al. (1998) investigated the optimal timing of technology adoption, showing that multiple factors, such as uncertainties about the speed of innovation arrival, the organization's initial technological attributes, the value of the new technology, strategic interactions in the market, and network externalities, affect the adoption strategy and timing decision. Furthermore, Kerr and Newell (2003) suggested that government regulations, together with some special economic instruments, such as taxes, tradable permits, or subsidies, should be used to provide incentives for technology adoption. Given the rapid advancement in information technology, recent studies in the information systems field have investigated technology adoption in various new contexts, such as open source software development (Peng & Dey, 2013), digital health records (Ozdemir et al., 2011), electronic payment systems (Plouffe et al., 2001; Bapna et al., 2011), mobile internet (Kim et al., 2007), mobile apps (Jung et al., 2019), and cloud IT services (Retana et al., 2018).

In the banking industry, Hannan and McDowell (1984) and Saloner and Shepard (1995) examined banks' technology adoption behavior. Clemons and Weber (1996) showed that the regulatory environment shapes banks' decisions to adopt alternative security trading systems. Liu et al. (2015) examined recent changes in the development of mobile payments, showing that competition and cooperation coexist among financial institutions and that regulatory roles are important in driving or delaying such innovation. Our work complements the stream of IT adoption research by studying next-generation retail payment and settlement system design and adoption in the wake of both traditional banks' revamped core payment infrastructure and fintech firms' new payment innovations.

Model Setup

New infrastructure designs of the core retail payment and settlement system are needed to catch the emerging fast payments trend. A number of issues need to be considered when building such an innovative and fast payment and settlement system.

Participants (PSPs): We consider a financial ecosystem of unit measure, of which a $\lambda \in (0,1)$ proportion comprises banks (denoted as "*B*") and a $1-\lambda$ proportion comprises fintech firms (denoted as "*F*") licensed to provide mobile payments, e-wallets, and other innovative payment services. Both banks and fintech firms are PSPs, which we call participants in the payment system. We denote the transaction volume of a bank as d_B and of a fintech firm as d_F. The transaction volume of a fintech firm, such as Alipay or WeChat Pay, might be lower or higher than that of a traditional bank.

Settlement Speed ($\boldsymbol{q_1}$): A key feature of the new payment system design is the settlement speed, denoted as q_1, which can be understood as how frequently the system settles payment requests at a predefined settlement cycle. Two extremes exist: Traditional BR retail payment systems adopt DNS settlement, for which net settlement only occurs at the end of the day. We normalize $q_1 = 0$ for the DNS payment system, serving as a lower bound of the settlement speed. In contrast, the RR retail payment systems implement the fastest RTGS settlement, which continuously processes payment settlement on an individual transaction basis, enabling settlement to occur in real time. We normalize $q_1 = 1$ to represent the upper bound of the settlement speed. In the new system design, the system operator can decide how often to run the settlement algorithm, ranging from very low frequency (several times a day) to almost continuous (every few seconds). The more frequent the netting algorithm is run, the faster the settlement speed. Therefore, $q_1 \in (0,1)$ characterizes the ER retail payment systems.

System Capability ($\boldsymbol{q_2}$): In addition to the settlement speed, a number of other important payment system attributes need to be considered, including ease of access, functionality, interoperability, and risk management (Chapman et al., 2015). A system that provides rich functionality and supports more innovative payment features, particularly with tools to enhance transaction security and efficiency, is considered a high capability system. Some examples of these features are automatic sorting, validation, and routing of payment files; detection of individual transaction errors or potential fraud; and the generation of automatic messages, notifications, and reports. Other key indicators of system capability include support for multiple payment applications (e.g., B2B, P2P, and mobile), 24×7×365 system access, adoption of the ISO20022 standard, and so on. These system features enhance the service for payment system participants and enable value-added end-user services. We denote q_2 as the system capability parameter, where a higher q_2 implies more innovative and supportive system features.

Value of fast payment service: A system with a faster settlement speed (i.e., a larger q_1) and/or higher capability (i.e., a larger q_2) better serves end users and generates higher value. For example, a faster settlement releases funds more quickly, enabling end users to obtain cash earlier for other opportunities. A high-capability settlement system supporting more innovative payment methods, such as mobile payments, has a high perceived value because of the convenience it offers to end users. Moreover, a system enabling tools for fraud detection and notification can improve payment security, creating value for consumers as well. High-quality end-user services enhance the relationship between PSPs and their customers and ultimately create higher business value.

Because both the settlement speed q_1 and system capability q_2 contribute to the overall valuation of the payment service, we model the valuation of the new system as the base value parameter v scaled by a measure of service quality, which is a weighted sum of these two system attributes. Denoting $w \in (0,1)$ as the weight parameter reflecting the relative importance between q_1 and q_2, we have:

$$V(q_1, q_2) = v[w(q_1 - cq_1^2) + (1-w)q_2]. \tag{1}$$

The first quadratic term $(q_1 - cq_1^2)$ captures the diminishing marginal return of utility as the settlement speed increases. We impose a mild condition $c \in [0, \frac{1}{2}]$ to ensure a strictly decreasing slope in the range of $q_1 \in [0,1]$. The second linear term states that the system value linearly increases as the system capability q_2 is enhanced; that is, as the number of innovative payment features and functionality enabled by the system increases, the valuation of the payment service increases.[5]

[5] Theoretically, we might normalize q_2 to be bounded by 1 to eliminate the scaling effect when compared with q_1 in the value function. However, practically, knowing the upper limit of innovations is impossible, and benchmarking a system design against the maximum number of innovative features defining the best future retail payment system would therefore be

Network effects: Positive network effects exist in almost all financial services. When more PSPs are in a payment and settlement system, more end customers gain access to fast payment services and are able to transact with other PSP customers, leading to a larger network effect. We denote the base network value as βn, where β is the network intensity parameter and n is the number of PSPs of the new fast payment and settlement service. Because a PSP that handles more payment requests obtains a higher network value from payment settlement, we assume PSP i's derived network value is proportional to its payment transaction volume, $d_i \beta n$, where $i = \{B, F\}$.

Infrastructure cost: The payment and settlement system requires a central infrastructure connecting participating PSPs. To respond to the fast payments trend, many countries have chosen to either renovate an existing BR system or build a brand-new ER or RR system. Either approach involves significant initial investment to take full advantage of the newest technological development. We assume that the investment cost takes a quadratic form, $\frac{k_i}{2} q_2^2$, where $i = \{F, B\}$. The quadratic functional form suggests that building a sophisticated, high-capability system with more functionality and more innovative features becomes increasingly difficult. We assume $k_F \leq k_B$ to reflect the fact that fintech firms have a technological advantage over traditional banks. Because fintech firms use modern technologies, such as machine learning, artificial intelligence, big data, and cloud computing, to provide improved financial services, they are able to efficiently and swiftly innovate, iterate, and improve systems. In contrast, banks' legacy systems restrict their ability to leverage new technologies, making them less efficient in taking advantage of emerging technologies. We further define $\Delta k = k_B - k_F$ and interpret $\frac{\Delta k}{k_B}$ and $\frac{\Delta k}{k_F}$ as the relative technological advantage of fintech firms over banks.

Liquidity cost: Liquidity management is critical for payment system efficiency. Although low liquidity and inadequate balances in settlement accounts slow down fund movement, the opposite case increases capital costs. Traditionally, participants pledge cash collateral into their settlement account held at the central bank. When the money in the settlement account is insufficient for settling payments, a liquidity shock occurs. To enable timely settlement, many central banks have in place automatic collateralized financing mechanisms or daylight overdraft facilities. For example, the People's Bank of China (the central bank of China) provides high-interest loans to participants for payment processing of insufficiently covered transactions, resulting in nonnegligible liquidity costs to them. When a liquidity shock occurs, the average liquidity cost per transaction L_i, where $i = \{F, B\}$, depends on the average size of the transaction (the amount of funds being transacted) handled by PSP i. Because the average monetary payment associated with each transaction for fintech firms is normally smaller than that for banks, we have $L_F \leq L_B$.[6]

A traditional DNS system accumulates payment requests throughout a day. At the end of the day, the net obligations between PSPs are calculated based on chosen netting algorithms. PSPs only need sufficient liquidity to cover net obligations, which significantly reduces the overall intraday liquidity needs. However, an expedited payment system settles payment requests in a timely manner by shortening the netting cycle. A PSP's liquidity cost increases with the frequency of the netting algorithm because it is less likely to find offsetting payments than would otherwise be possible with longer netting cycles. As a result, the PSP is more likely to encounter liquidity shocks. The expected liquidity costs are the highest in the RTGS system, in which each payment transaction is individually and immediately settled. We denote $\theta \in (0,1)$ as the base probability of a liquidity shock per payment transaction. To reflect the increasing likelihood of a liquidity shortfall as the payment speed increases, we assume that it is a linear function of the settlement speed θq_1. Accordingly, the expected liquidity cost per transaction is $\theta q_1 L_i$, where $i = \{F, B\}$. This modeling approach is supported by industry practice. First, because $q_1 = 0$ under DNS, both the probability of a liquidity shortfall and the expected liquidity cost per transaction are normalized to 0 under DNS, which is consistent with the current practice that PSPs normally maintain sufficient reserves at the central bank under the traditional DNS system. It establishes a benchmark to allow for a comparison of the additional risks introduced by the expedited settlement. Second, as the settlement speed increases, the likelihood of incurring a liquidity shortfall increases, and PSPs incur the highest expected liquidity cost under the RTGS system when $q_1 = 1$. Because payments are settled individually under RTGS, the expected liquidity cost per transaction should be the liquidity cost per transaction weighted by the likelihood of liquidity shock (i.e., θL_i), which makes intuitive sense. In fact, a high expected liquidity cost is well recognized as the major hurdle for banks to adopt (near) real-time payments (Leinonen & Soramäki, 1999; Kahn & Roberds, 2001; Johnson et al., 2004; Willison, 2005).

extremely challenging. This is different from q_1, where the maximum settlement speed (i.e., real-time settlement) is well defined and normalized to 1. For this reason, we decide not to normalize q_2. Instead, by appropriately choosing the other model parameters, we can ensure that neither q_1 nor q_2 dominates each other, and a clear trade-off exists between the two decisions, as shown in our subsequent numerical examples.

[6] Regulatory practices typically restrict the average amount involved in each transaction on fintech-based platforms to be lower than that of bank-managed transactions. For example, in China, QR-code payments for mobile payment providers such as WeChat and Alipay are limited to a maximum of 500 RMB per day. Apple Pay is not accepted for transaction amounts higher than HKD500 in Hong Kong and S$200 in Singapore.

Innovation initiative: Either fintech firms or banks can take the initiative to build and operate the new system, which we term *fintech-led* or *bank-led* innovation, respectively. When building the new system, the owner has two potential strategies: an *exclusive* strategy, for which the system is only optimally designed for their own use, and an *inclusive* strategy, for which the owner optimizes the design by considering other participants' incentives to adopt the system and charging them a fee p to use it.

Alternatively, governments can lead innovation. Under *government-led* innovation, the government calls for participation from both fintech firms and banks. The government requires fintech firms to build the new system because of its cost-effectiveness, mandates banks to share part of the infrastructure cost with fintech firms, and provides the new system free to all participants after it is built.[7] In this research, we propose a Shapley-value-based cost-sharing rule to allow both fintech firms and banks to fairly share the new infrastructure investment cost and benefit.

Putting all of these together, for PSP i, the net utility derived from using the new payment and settlement system characterized by (q_1, q_2) can be written as:

$$U_i(q_1, q_2) = d_i v[w(q_1 - cq_1^2) + (1-w)q_2] + d_i\beta n - d_i\theta q_1 L_i - p, \tag{2}$$

where $i = \{F, B\}$. For simplicity, we normalize $d_B = 1$ and denote $d_F \stackrel{\text{def}}{=} d$ throughout the paper. Appendix B provides a complete notation table.

Problem Formulation and Analysis

In practice, it is common to see a group of PSPs—a bank association or fintech consortium—form the hub of a payment system and build and operate the new retail payment system (Soramäki et al., 2007; Embree & Roberts, 2009). In the following subsection, we analyze optimal system designs when banks and fintech firms separately build and operate their own systems and charge a high fee to deter the other group's use. We call these the *fintech-exclusive* or *bank-exclusive* strategies (denoted FE and BE throughout the paper). In the "Fintech-Inclusive (FI) Strategy" and "Bank-Inclusive (BI) Strategy" subsections below, we analyze the case in which one group (the fintech consortium or the bank association) builds the system and charges the other an appropriate fee to use it. We call these the *fintech-inclusive* or *bank-inclusive* strategies (denoted as FI and BI throughout the paper).

Fintech-Exclusive (FE) and Bank-Exclusive (BE) Strategies

Under the exclusive strategy, major banks form an association, and fintech firms form a consortium to separately build their respective systems for their own use. The system owner (i.e., the association or consortium) determines the optimal design (q_1, q_2) to maximize the benefits of their members.

The fintech consortium and the bank association's optimization problems are, respectively:

$$\pi_F^r = max_{q_1, q_2}\ (1-\lambda)d_F\{v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta(1-\lambda) - \theta q_1 L_F\} - \frac{k_F}{2}q_2^2, \tag{3}$$

$$\pi_B^r = max_{q_1, q_2,}\ \lambda\{v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta\lambda - \theta q_1 L_B\} - \frac{k_B}{2}q_2^2. \tag{4}$$

Because fintech firms' new payment system is only for their exclusive use, the network size is $(1-\lambda)$. The objective function in Equation (3) maximizes the overall benefits of fintech firms, including the total value that the $(1-\lambda)$ fintech firms derive from using the new system minus their total expected liquidity and infrastructure costs. Similarly, the objective function in Equation (4) maximizes the overall benefits of λ banks when they build an exclusive system for their own use. Because π_i^r, where $i = \{F, B\}$, represents the total benefits each group could gain if they build their own payment and settlement system, we interpret the optimal objective function values in Equations (3) and (4) as the *reservation values* of the fintech consortium or bank association.

Under the exclusive strategy, each group charges a fee high enough to exclude the other group's participation; that is, the fintech consortium charges p^{FE} such that banks' incentive compatibility (IC) condition cannot be satisfied: $\lambda\{v[w(q_1^{FE} - c(q_1^{FE})^2) + (1-w)q_2^{FE}] + \beta - \theta q_1^{FE} L_B - p^{FE}\} < \pi_B^r$. If banks choose to join the fintech consortium's new system, the network size of the system is 1. In such a case, a bank derives the system value $v[w(q_1^{FE} - c(q_1^{FE})^2) + (1-w)q_2^{FE}]$, enjoys the network value β, and incurs the liquidity cost $\theta q_1^{FE} L_B$. In addition, the bank needs to pay a fixed access fee p^{FE}. The IC condition states that under the optimal system features and the fee chosen by fintech firms, banks' total benefits of joining the system are lower than their reservation value π_B^r. Similarly, under the bank-exclusive strategy, the bank association charges a high

[7] The shared infrastructure paradigm is common in practice. For example, the Bank of Mexico charges a fixed fee based on the annual cost of providing the Sistema de Pagos Electrónicos Interbancarios (SPEI) payment service. The annual cost, which is distributed among system participants, covers the overall infrastructure costs.

fee p^{BE} to exclude fintech firms' participation: $(1-\lambda)d_F\{v[w(q_1^{BE}-c(q_1^{BE})^2)+(1-w)q_2^{BE}]+\beta-\theta q_1^{BE}L_F-p^{BE}\}<\pi_F^r$.

Lemmas 1 and 2 summarize the optimal system design by fintech firms and banks, respectively. All proofs are provided in Appendix C.

Lemma 1: *Under the FE strategy, fintech firms collectively build their own new payment and settlement system and use it exclusively among themselves. The optimal system design is:*

$$q_1^{FE}=\begin{cases}0 & \text{if } v\le \underline{v}_F\\ \frac{1}{2c}\left(1-\frac{\theta L_F}{vw}\right) & \text{if } \underline{v}_F<v<\overline{v}_F\\ 1 & \text{if } v\ge \overline{v}_F\end{cases} \text{ and } q_2^{FE}=\frac{d_F(1-\lambda)v(1-w)}{k_F},$$

where $\underline{v}_F=\frac{\theta L_F}{w}$ *and* $\overline{v}_F=\frac{\theta L_F}{(1-2c)w}$.

The settlement speed q_1 is mainly determined by two factors. It decreases in the average liquidity cost per transaction, which is L_F in this case because only fintech firms use their own new system. Given high liquidity costs, the incentive to expedite settlements is reduced. On the other hand, q_1 increases in the payment value parameter v but not at a constant rate. As shown in Figure 1 (left side), two critical values $\underline{v}_F$ and $\overline{v}_F$ categorize the three regions. We call region $v\le\underline{v}_F$ the BR region; in this region, fintech firms still adopt the slowest end-of-day DNS arrangement with no intraday settlement of funds ($q_1^{FE}=0$). In the middle ER region $\underline{v}_F<v<\overline{v}_F$, settlement is expedited but not yet to the maximum of its potential, $0<q_1^{FE}<1$. The settlement speed increases in v in a concave form, approaching real time at the upper bound $\overline{v}_F$. Finally, region $v\ge\overline{v}_F$ is the RR region in which fintech firms provide the fastest real-time settlement services, $q_1^{FE}=1$.

In contrast, the system capability q_2 is independent of the liquidity cost. It linearly increases in the base value parameter v and decreases in the infrastructure cost parameter k_F. We further note that, although the total transaction volume of the system does not affect the optimal settlement speed q_1, it does positively affect system capability q_2. When the system is expected to handle a larger number of transactions, the owner has a stronger incentive to build more innovative features. Because $(1-\lambda)$ fintech firms exist in this new payment network, each with d_F transactions, the system capability increases in the total volume of transactions $d_F(1-\lambda)$.

Lemma 2: *Under the BE strategy, banks collectively build their own new payment and settlement system and use it exclusively among themselves. The optimal system design is:*

$$q_1^{BE}=\begin{cases}0 & \text{if } v\le \underline{v}_B\\ \frac{1}{2c}(1-\frac{\theta L_B}{vw}) & \text{if } \underline{v}_B<v<\overline{v}_B\\ 1 & \text{if } v\ge \overline{v}_B\end{cases} \text{ and } q_2^{BE}=\frac{\lambda v(1-w)}{k_B},$$

where $\underline{v}_B=\frac{\theta L_B}{w}$ *and* $\overline{v}_B=\frac{\theta L_B}{(1-2c)w}$.

Banks' choice of their own exclusive system exhibits patterns similar to those of fintech firms' systems. Three regions exist for the optimal settlement speed: in the BR region when $v\le\underline{v}_B$, banks stick to the slowest end-of-day DNS arrangement; in the ER region when $\underline{v}_B<v<\overline{v}_B$, settlement is expedited but not yet to the maximum of its potential; and in the RR region when $v\ge\overline{v}_B$, banks adopt the fastest real-time settlement. The optimal system capability decreases in the infrastructure cost (k_B) and increases in the total transaction volume (λ) that it handles.

Comparing the system design solutions characterized by Lemmas 1 and 2, we obtain the market outcome when both fintech firms and banks build their own payment and settlement systems independently without cooperation, as described in Proposition 1 and Figure 1.

Proposition 1—Exclusive strategies: *In the absence of cooperation, fintech firms and banks invest in building respective payment and settlement systems for their own exclusive use:*

(i) $q_1^{FE}\ge q_1^{BE}$*; that is, fintech firms are always keener to speed up settlement than banks.*

(ii) $q_2^{FE}\ge q_2^{BE}$ *if* $\frac{k_F}{k_B}\le\frac{d_F(1-\lambda)}{\lambda}$*, and* $q_2^{FE}<q_2^{BE}$ *otherwise; that is, when their technology advantage* ($\frac{k_F}{k_B}$) *and/or relative transaction volume* ($\frac{d_F(1-\lambda)}{\lambda}$) *is significant, fintech firms build more innovative payment features in their exclusive system relative to that of banks.*

Note that we have $\underline{v}_F<\underline{v}_B$ and $\overline{v}_F<\overline{v}_B$. In terms of settlement speed, banks' exclusive payment and settlement system has a larger BR region with the slowest DNS settlement and a smaller RR region with the fastest RTGS settlement. Furthermore, in the middle ER region, its settlement speed is always slower than that in fintech firms' exclusive system.

In terms of innovative payment features, which system offers higher capability depends on two key ratios. The first ratio, $\frac{k_F}{k_B}$, measures fintech firms' technology advantages relative to those of banks. When fintech firms possess more technology advantages (i.e., smaller $\frac{k_F}{k_B}$) and, thus, can offer more cost-effective solutions than banks, they build a system with more innovative features and functionalities.

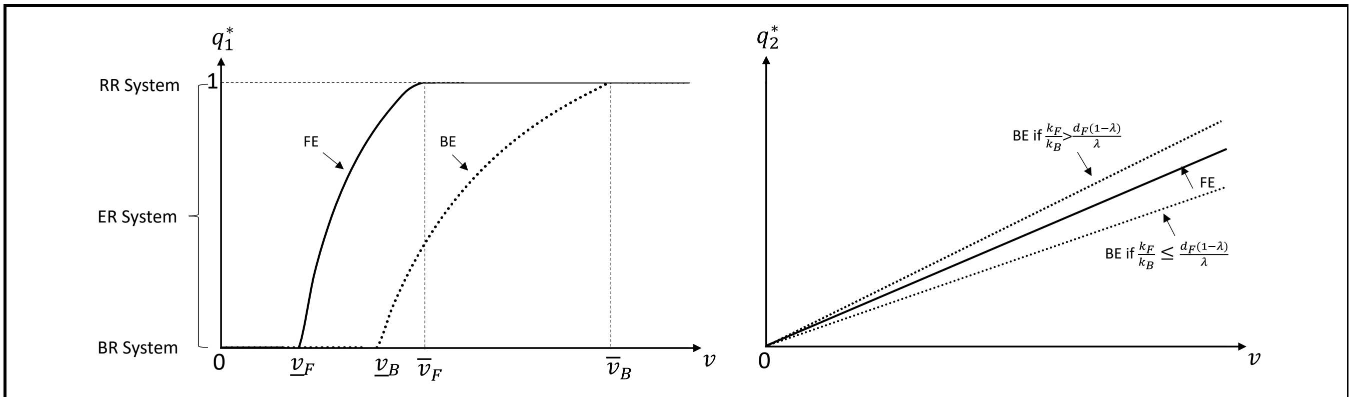

Figure 1. Optimal System Design: FE and BE Strategies

The second ratio, $\frac{d_F(1-\lambda)}{\lambda}$, denotes fintech firms' total transaction volume relative to that of the banks and, thus, is an indicator of their relative importance in the retail payments market. For many years, banks have been the dominant players in the payment industry. Fintech firms are only recent entrants, and their transaction volume is increasing with the use of mobile payments and other e-wallet solutions. According to Accenture, new entrants to the banking market, including nonbank payment institutions and tech companies, are amassing up to one-third of the new revenue, which is challenging the competitiveness of traditional banks.[8] In China, fintech firms Ant Financial and Tencent have dominated China's mobile payments market (Keyes & Magana, 2019).

Fintech-Inclusive (FI) Strategy

Under the FI strategy, fintech firms form a consortium to lead innovation. They build the system with the appropriate configuration (q_1, q_2) and set an access fee p to attract banks to use it. The fintech consortium's optimization problem is:

$$max_{q_1,q_2,p}\ (1-\lambda)d_F\{v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta - \theta q_1 L_F\} - \frac{k_F}{2}q_2^2 + \lambda p \tag{5}$$

$$s.t.\ \ \lambda\{v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta - \theta q_1 L_B - p\} \geq \pi_B^r \tag{6}$$

Objective Function (5) aims to maximize the total net profits of fintech firms. The terms in the curly brackets represent the system value, network value (given that the network size of the fintech-led unified system is 1), and the expected liquidity cost of $(1-\lambda)$ fintech firms, each with transaction volume d_F. The second term is fintech firms' infrastructure cost, and the last term is the total revenue generated from charging λ banks. Constraint (6) is banks' IC condition. A bank derives system value $v[w(q_1 - cq_1^2) + (1-w)q_2]$ from using the payment service, obtains network value β, incurs expected liquidity cost $\theta q_1 L_B$, and pays access fee p. The net payoff of λ banks, when using the fintech-led system, needs to be no less than their reservation value π_B^r, which is the payoff that banks can obtain if they choose to build their own system (as in our analysis in the "Fintech-Exclusive (FE) and Bank-Exclusive (BE) Strategies" subsection).

Lemma 3: *Under the FI strategy, the fintech consortium builds a unified payment and settlement system for all participants to use.*

(i) The optimal system design is:

$$q_1^{FI} = \begin{cases} 0 & if\ v \leq \underline{v}_{FI} \\ \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right) & if\ \underline{v}_{FI} < v < \overline{v}_{FI} \\ 1 & if\ v \geq \overline{v}_{FI} \end{cases} \ and\ q_2^{FI} = \frac{v(1-w)\Gamma}{k_F},$$

where $\tilde{L} \stackrel{\text{def}}{=} \frac{(1-\lambda)d_F L_F + \lambda L_B}{(1-\lambda)d_F + \lambda}$, $\Gamma \stackrel{\text{def}}{=} (1-\lambda)d_F + \lambda$, $\underline{v}_{FI} = \frac{\tilde{L}\theta}{w}$, *and* $\overline{v}_{FI} = \frac{\tilde{L}\theta}{(1-2c)w}$.

(ii) The optimal price is $p^{FI} = v[w(q_1^{FI} - c(q_1^{FI})^2) + (1-w)q_2^{FI}] + \beta - \theta q_1^{FI} L_B - \frac{\pi_B^r}{\lambda}$.

We define two variables: $\tilde{L}$ and Γ. The former is the volume-weighted average liquidity cost per transaction in the unified system, and the latter is the total transaction volume handled

[8] https://newsroom.accenture.com/news/banks-revenue-growth-at-risk-due-to-unprecedented-competitive-pressure-resulting-from-digital-disruption-accenture-study-finds.htm

in the system. Lemma 3(i) shows three regions of v that correspond to three different types of systems with different settlement speeds: the BR region when $v \leq \underline{v}_{FI}$, the RR region when $v \geq \overline{v}_{FI}$, and the middle ER region when $\underline{v}_{FI} < v < \overline{v}_{FI}$. Both $\underline{v}_{FI}$ and $\overline{v}_{FI}$ are functions of $\tilde{L}$. Figure 2 compares the optimal system design between the fintech-inclusive, fintech-exclusive, and bank-exclusive strategies.

We first compare this fintech-led unified system with the exclusive system that fintech firms build for their own use. First, $\underline{v}_{FI} > \underline{v}_F$, and $\overline{v}_{FI} > \overline{v}_F$, suggesting a wider BR region and a smaller RR region once banks are included in the system. In addition, in the middle ER region, $q_1^{FI} < q_1^{FE}$. Hence, when fintech firms build the system for all participants, they reduce the settlement speed to accommodate banks' increased liquidity pressure from fast payment settlement. On the other hand, $q_2^{FI} > q_2^{FE}$ because the total value of the new system increases as the network size increases. Therefore, fintech firms have a greater incentive to build more features in the unified settlement system.

Comparing this fintech-led unified system with the exclusive system that banks build for their own use, we find $\underline{v}_{FI} < \underline{v}_B$ and $\overline{v}_{FI} < \overline{v}_B$, showing a smaller BR region and a wider RR region; in addition, $q_1^{FI} > q_1^{BE}$ in the middle ER region. Banks have to adopt a higher settlement speed than their ideal level if joining the unified system. In addition, banks now enjoy more payment functionality, $q_2^{FI} > q_2^{BE}$, thereby economizing the benefits from the increased network size. However, although banks use an improved system if they adopt the fintech consortium's innovation rather than building a new system on their own, their total net benefits are the same. Banks earn their reservation value (π_B^r) because fintech firms charge an appropriate price to make banks' IC constraint (6) bind, as shown in Lemma 3(ii). A detailed expression of the optimal price is presented in Appendix C.

To summarize, when fintech firms lead payment innovation by building a unified system for all participants to use, compared with the case in which fintech firms and banks each build their own exclusive systems, we find the following. First, each party needs to deviate from their respective ideal settlement speed to accommodate the other's participation; therefore, the compromised settlement speed is set at the middle level: $q_1^{BE} \leq q_1^{FI} \leq q_1^{FE}$. Second, both parties enjoy more payment settlement supporting features: $q_2^{FI} > q_2^{FE}$ and $q_2^{FI} > q_2^{BE}$. Rich functionality is an essential element of the state-of-the-art payment settlement system design, allowing for other value-added services in addition to timely settlement (Chapman et al., 2015). Hence, a unified settlement system benefits all participants by offering more innovative payment features and functionalities.

Next, we examine when fintech firms have the incentive to build such a unified system.

Proposition 2—Fintech firms' optimal strategy: *There might exist $\underline{v}_{e1} > \underline{v}_F$ and $\overline{v}_{e1} < \overline{v}_B$ such that for fintech firms, the FE strategy is optimal if $v \in (\underline{v}_{e1}, \overline{v}_{e1})$, and the FI strategy is optimal in the remaining regions.*

(i) The FE-optimality region $(\underline{v}_{e1}, \overline{v}_{e1})$ might be empty.

(ii) When $\frac{\Delta k}{k_B}$ increases, the FE-optimality region is less likely to appear.

(iii) When $\frac{\lambda}{(1-\lambda)d_F}$ increases, the FE-optimality region is less likely to appear.

(iv) When β increases, the FE-optimality region is less likely to appear.

Proposition 2 shows that it is not always profit-improving for fintech firms to include banks in their new payment and settlement system. When deciding on whether to build the unified system for all, fintech firms face the following trade-offs. A unified system brings the benefits of enhanced system capability, which comes with higher infrastructure costs and utility loss from sacrificed speed because fintech firms must slow down settlements to accommodate banks' needs. In addition, under certain circumstances, considering the reservation value that banks must be given to motivate their participation, the price that fintech firms charge banks to use the system might be limited. Proposition 2 identifies the region $(\underline{v}_{e1}, \overline{v}_{e1})$ in which the overall gain from a unified system cannot compensate for fintech firms' additional costs; therefore, they opt for the exclusive strategy and build the system only for their own use.

The conditions under which the FE strategy is optimal depend on the interplay of a few key parameter values. Proposition 2 suggests that the FE-optimality region, if it exists, only appears in the intermediate range of v. Hence, as the base value parameter v increases, fintech firms' optimal strategy follows a pattern of "inclusive—exclusive—inclusive." When v is very large ($v \geq \overline{v}_B$) or relatively small ($v \leq \underline{v}_F$), the FI strategy is always optimal. The intuition is that if v is sufficiently large (small), fintech firms' and banks' preferences over settlement speed are similar and can be easily aligned. As a result, fintech firms neither need to deviate too much from their ideal settlement speed nor give significant compensation to banks when building and operating a unified system. Therefore, the value of the unified system outweighs the cost.

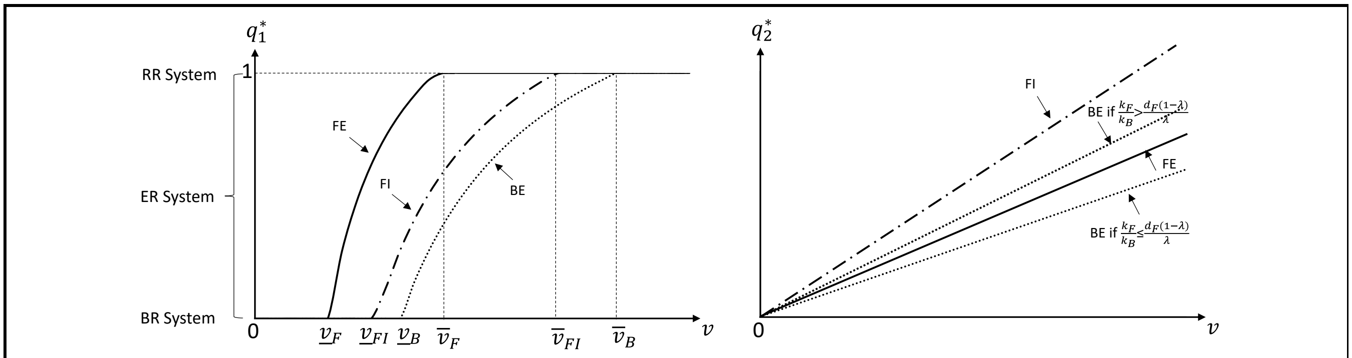

Figure 2. Optimal System Design Comparison: FI, FE, and BE Strategies

Propositions 2(*ii*), 2(*iii*), and 2(*iv*) identify three critical factors that affect the size of FE-optimality regions. $\frac{\Delta k}{k_B}$ measures the relative cost-efficiency between fintech firms and banks, which is an indication of fintech firms' technological advantage. When the ratio is large, fintech firms possess a significant technological advantage over banks (in terms of building innovative payment features). They are more willing to adopt the FI strategy because building the unified settlement system enables fintech firms to leverage their technology advancement to generate a higher total value, which can in turn attract banks to participate.

In addition, $\frac{\lambda}{(1-\lambda)d_F}$ is the ratio of banks' transaction volume to fintech firms, which measures their relative "market share." When banks take a dominant position, this ratio tends to be large, which is currently true for most countries. In this case, fintech firms are more willing to adopt the FI strategy because the gain from serving banks is expected to be significant.

Finally, we note that the FE-optimality region shrinks when the network intensity parameter β increases. This shrinkage is intuitive because the payment system builders are more likely to adopt the inclusive strategy for obtaining a large network value as the network externality becomes stronger. Figure 3 illustrates how the relative technological advantage ($\frac{\Delta k}{k_B}$), the ratio of transaction volume ($\frac{\lambda}{(1-\lambda)d_F}$), and the network effects (β) affect the fintech consortium's optimal strategy.[9]

Figure 3(a) shows that, in general, only in an intermediate range of v values might the fintech consortium prefer the FE strategy. As $\frac{\lambda}{(1-\lambda)d_F}$ increases, the boundaries that define the FE optimality region shrink inwards (the solid lines move toward the dashed lines). This shrinkage suggests that the fintech consortium is more likely to adopt the FI strategy when banks' transaction volume is high. In addition, the FI optimality region is larger at a higher level of $\frac{\Delta k}{k_B}$, implying that fintech firms also have a stronger incentive to include banks in a unified system when they have a greater technological advantage. Figure 3(b) further shows that the FE region shrinks and the FI region expands as the network effects become larger. The fintech consortium is more willing to include banks in its payment systems and adopt the FI strategy in the presence of stronger network effects.

Bank-Inclusive (BI) Strategy

Under the BI strategy, the bank association builds the infrastructure, owns the new system, and charges fintech firms a fee to use the payment services. Its optimization problem is:

$$max_{q_1,q_2,p}\ \lambda\{v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta - \theta q_1 L_B\} - \frac{k_B}{2}q_2^2 + (1-\lambda)p \quad (7)$$

$$s.t. \qquad (1-\lambda)\{d_F[v(w(q_1 - cq_1^2) + (1-w)q_2) + \beta - \theta q_1 L_F] - p\} \geq \pi_F^r \quad (8)$$

The interpretation of Objective Function (7) and Constraint (8) is similar to the fintech consortium's optimization problem under the FI strategy. When all PSPs join the new system, the network size of the unified system is 1. Banks maximize their total profits under a unified settlement system and ensure fintech firms' participation incentive by offering them no less than their reservation value. Lemma 4 presents the optimal system design under the BI strategy.

[9] In this numerical illustration, we set base parameter values $w = 0.75, \theta = 0.8$, $L_F = 1$, and $L_B = 2$; therefore, fintech firms' expected liquidity cost per transaction is lower than that of the banks. Furthermore, we set $c = 0.3$, $k_B = 10$ and $k_F \in (0,10]$ to represent fintech firms' technological advantage. We use base values $\beta = 0.02$ in Figure 3(a) and $\lambda = 0.5$ and $d_F = 1$ in Figure 3(b). To be consistent and for illustration purpose, this set of parameters is used throughout the paper, unless mentioned otherwise.

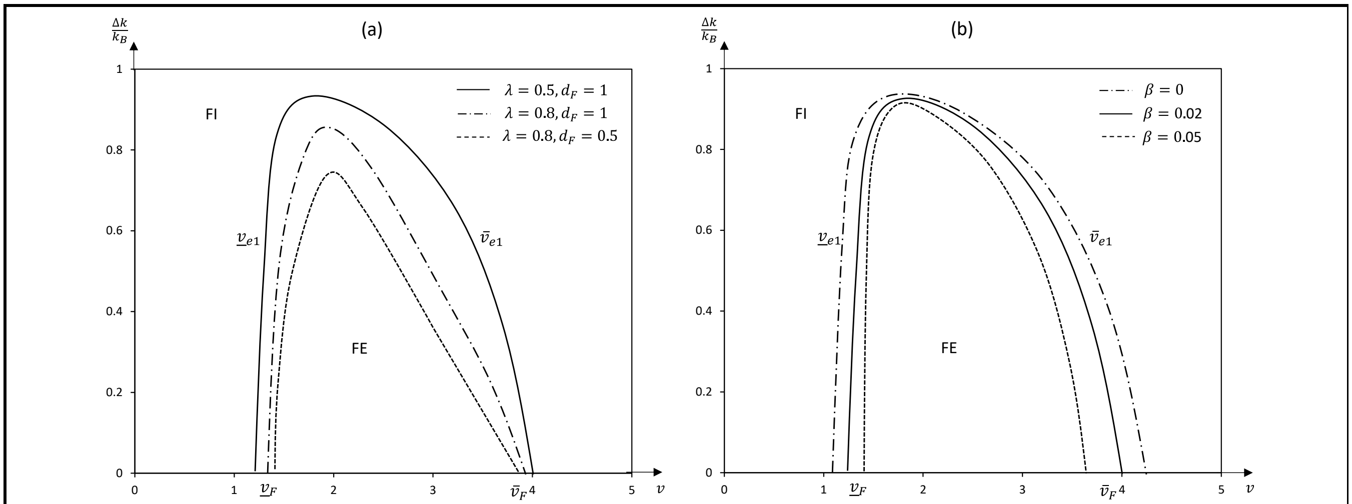

Figure 3. Effects of $\frac{\Delta k}{k_B}$, $\frac{\lambda}{(1-\lambda)d_F}$, and β on the Fintech Consortium's Optimal Strategy

Lemma 4: *Under the BI strategy, the bank association builds a unified payment and settlement system for all participants to use.*

(i) The optimal system design is:

$$q_1^{BI} = \begin{cases} 0 & \text{if } v \le \underline{v}_{BI} \\ \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right) & \text{if } \underline{v}_{BI} < v < \overline{v}_{BI} \\ 1 & \text{if } v \ge \overline{v}_{BI} \end{cases} \quad \text{and } q_2^{BI} = \frac{v(1-w)\Gamma}{k_B},$$

where $\underline{v}_{BI} = \frac{\tilde{L}\theta}{w}$ *and* $\overline{v}_{BI} = \frac{\tilde{L}\theta}{(1-2c)w}$.

(ii) The optimal price is $p^{BI} = d_F[v(w(q_1^{BI} - c(q_1^{BI})^2) + (1-w)q_2^{BI}) + \beta - \theta q_1^{BI} L_F] - \frac{\pi_F^r}{1-\lambda}$.

Similarly, banks charge fintech firms a price to use the new system. A detailed expression of the optimal price is presented in Appendix C. The two variables, $\tilde{L}$ and Γ, are defined in Lemma 3. Comparing the optimal system design in Lemma 4 with that in Lemma 3, we see that the border values of the three regions (BR, ER, and RR) are the same, $\underline{v}_{BI} = \underline{v}_{FI}$ and $\overline{v}_{BI} = \overline{v}_{FI}$, and the settlement speed in each region is also the same, $q_1^{BI} = q_1^{FI}$. This finding suggests that when a unified system is adopted by all participants, regardless of the system owner, the settlement speed is always the same. Both fintech firms and banks deviate from their respective most preferred settlement speed—banks need to use a system faster than their ideal choice, whereas fintech firms need to use a system slower than their ideal choice ($q_1^{BE} \le q_1^{BI} = q_1^{FI} \le q_1^{FE}$)—to accommodate the other party's participation.

Further, we find $q_2^{BE} < q_2^{BI} < q_2^{FI}$. Banks incorporate more innovative payment features when building a unified system for all participants than when building an exclusive system under the BE strategy ($q_2^{BE} < q_2^{BI}$) because the expected larger network size in the unified system enables economies of scale. However, the number of innovative features in the BI system is fewer than that in the FI system ($q_2^{BI} < q_2^{FI}$) because of banks' technological disadvantage.

Similarly, it is not always profit-improving for banks to build a unified system and attract fintech firms to use it. We further show that, compared with fintech firms, banks have less incentive to adopt the inclusive strategy. The following Proposition 3 summarizes our findings.

Proposition 3—Banks' optimal strategy: *When* $\frac{\Delta k}{k_F} \ge \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda)d_F{}^2}$, *the BE strategy is always optimal for banks; when* $\frac{\Delta k}{k_F} < \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda)d_F{}^2}$, *there might exist* $\underline{v}_{e2} > \underline{v}_F$ *and* $\overline{v}_{e2} < \overline{v}_B$, *such that the BE strategy is optimal if* $v \in (\underline{v}_{e2}, \overline{v}_{e2})$ *and the BI strategy is optimal in the remaining regions.*

(i) The BE-optimality region $(\underline{v}_{e2}, \overline{v}_{e2})$ *might be empty.*

(ii) $(\underline{v}_{e1}, \overline{v}_{e1}) \subset (\underline{v}_{e2}, \overline{v}_{e2})$, *where* $(\underline{v}_{e1}, \overline{v}_{e1})$ *is the FE-optimality region defined in Proposition 2.*

(iii) When $\frac{\Delta k}{k_F}$ *increases, the BE-optimality region is more likely to appear.*

(iv) When $\frac{\lambda}{(1-\lambda)d_F}$ *increases, the BE-optimality region is less likely to appear.*

(v) When β *increases, the BE-optimality region is less likely to appear.*

We have shown in Proposition 2 that, for fintech firms, the FI strategy is always optimal in some regions of v—for example, in the small or large v regions in which the discrepancy between fintech firms' and banks' ideal settlement speed is not too large. However, Proposition 3 suggests that, unlike fintech firms, banks may find that the BE strategy completely dominates the BI strategy in all regions of v if the condition $\frac{\Delta k}{k_F} \geq \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$ holds. This condition again highlights two important factors: $\frac{\Delta k}{k_F}$, an indication of fintech firms' relative technological advantage over banks, and $\frac{\lambda}{(1-\lambda)d_F}$, the two parties' relative market share in the payment and settlement system. In the scenario in which fintech firms are much more cost-effective than banks ($\frac{\Delta k}{k_F}$ is large), innovative system features q_2^{BI} chosen by banks in their settlement system are significantly fewer than what fintech firms would like to have. The cost of inducing fintech firms to use the unified system built by banks is too high. As a result, banks exclude fintech firms. In addition, in the scenario in which fintech firms take a relatively large market share of payment transactions ($\frac{\lambda}{(1-\lambda)d_F}$ is small), banks also prefer to exclude fintech firms because the large transaction volume from fintech firms results in a small volume-weighted average liquidity cost per transaction. Hence, the optimal settlement speed q_1^{BI} in a unified system increases accordingly. When this optimal speed is significantly faster than banks' ideal speed, they are not willing to compromise and opt to exclude fintech firms instead.

Only when $\frac{\Delta k}{k_F} < \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$, that is, when fintech firms' relative technology advantage is not too significant and their relative transaction volume is not too large, might banks have an incentive to include fintech firms in the unified bank-led system. Similar to the FI strategy, the BI strategy is optimal in the regions in which fintech firms' and banks' preferences for settlement speed are similar and incentives are thus well aligned. As the base value parameter v increases, banks' optimal strategy follows a similar pattern of "inclusive—exclusive—inclusive." In addition, Proposition 3(*ii*) shows that the FE-optimality region is always a subset of the BE-optimality region; that is, $(\underline{v}_{e1}, \overline{v}_{e1}) \subset (\underline{v}_{e2}, \overline{v}_{e2})$. This finding suggests that banks' incentive to provide a unified payment and settlement system for all participants is always lower than that of fintech firms.

Propositions 3(*iii*) and 3(*iv*) show the impact of $\frac{\Delta k}{k_F}$ and $\frac{\lambda}{(1-\lambda)d_F}$. When the technology advantage or the market share of fintech firms increases (i.e., $\frac{\Delta k}{k_F}$ increases or $\frac{\lambda}{(1-\lambda)d_F}$ decreases), the BE-optimality region expands. Until the condition $\frac{\Delta k}{k_F} < \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$ no longer holds, the BE strategy becomes optimal in all regions. Note that although both $\frac{\Delta k}{k_F}$ and $\frac{\lambda}{(1-\lambda)d_F}$ affect participants' incentives, their effects are different. As $\frac{\Delta k}{k_F}$ increases, the relative technological gap increases. As a result, banks have more incentive to build their own exclusive system, whereas fintech firms have more incentive to build a unified system and include banks. As $\frac{\lambda}{(1-\lambda)d_F}$ decreases, fintech firms take a more significant share in the payments market. Consequently, both fintech firms and banks are less incentivized to include each other in a unified system. This finding suggests that, over time, as the number of fintech firms increases (λ decreases) or their transaction volume increases (d_F increases), both fintech firms and banks are keen on independently developing their own exclusive systems, resulting in multiple payment systems coexisting in the market.

Proposition 3(v) summarizes the impact of the network effects. Similar to Proposition 2, when the network intensity parameter β increases, banks are more likely to choose the BI strategy and enjoy a higher network value. Hence, the BE-optimality region shrinks.

Market Equilibrium and Government Coordination

In this section, we first analyze the market equilibrium outcomes and their impacts on social welfare. We then identify scenarios in which government coordination is necessary and further propose a fair cost-benefit sharing rule among PSPs.

Market Equilibrium

A market equilibrium outcome is achieved when both fintech firms and banks have no incentive to deviate from their equilibrium strategies. We derive the following three possible equilibrium outcomes that lead to either a fragmented or unified future payment system.

Fragmented system: fintech firms and banks each adopt FE and BE strategies to separately build their own payment and settlement systems for exclusive use.

fintech-led unified system: fintech firms adopt the FI strategy to build the new system, and banks choose to use it.

Bank-led unified system: Banks adopt the BI strategy to build the new system, and fintech firms choose to use it.

Proposition 4 summarizes the conditions under which different equilibria might occur.

Proposition 4—Equilibrium systems without government coordination:

(i) If $\frac{\Delta k}{k_F} \geq \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$, *then there is a unique equilibrium outcome: a unified fintech-led system if* $v \in (0, \underline{v}_{e1}]$ *or* $v > \overline{v}_{e1}$, *and a fragmented system if* $v \in (\underline{v}_{e1}, \overline{v}_{e1}]$.

(ii) If $\frac{\Delta k}{k_F} < \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$, *then there are one or two equilibrium outcomes: either a unified fintech-led system or a unified bank-led system if* $v \in (0, \underline{v}_{e2}]$ *or* $v > \overline{v}_{e2}$; *a unified fintech-led system if* $v \in (\underline{v}_{e2}, \underline{v}_{e1}]$ *or* $v \in (\overline{v}_{e1}, \overline{v}_{e2}]$; *and a fragmented system if* $v \in (\underline{v}_{e1}, \overline{v}_{e1}]$.

Figure 4 illustrates the equilibrium systems and the strategies of banks/fintech firms accordingly.

The horizontal axis indicates the base value parameter v, and the vertical axis shows the relative technological advantage of the fintech firms over banks. The solid curves define the boundary regions between the optimal FE and FI strategies, and the dashed curves define the boundary regions between the optimal BE and BI strategies. Note that when $\Delta k = 0$, fintech firms and banks have the same investment cost function, and the fintech-led and bank-led unified systems result in the same optimal solution, which leads to $\underline{v}_{e1} = \underline{v}_{e2}$ and $\bar{v}_{e1} = \bar{v}_{e2}$ on the horizontal axis. As fintech firms' relative technological advantage over banks increases, fintech firms are less willing to adopt the FE strategy (the region defined by the solid $\underline{v}_{e1}$ and $\bar{v}_{e1}$ curves shrinks), whereas banks are more willing to adopt the BE strategy (the region defined by the dashed $\underline{v}_{e2}$ and $\bar{v}_{e2}$ curves expands). Because fintech firms' cost-effectiveness enables them to build a unified system with more innovative features for all PSPs to use (i.e., a larger q_2), the benefit from additional innovative features not only compensates for fintech firms' sacrifice in settlement speed but also accommodates banks' increased expected liquidity cost, making the FI strategy more attractive than the FE strategy. Therefore, the region of the fintech-led unified system expands (Regions I, III, and IV in Figure 4). Similarly, given banks' cost disadvantage, they tend to build a unified system with fewer innovative features (i.e., a smaller q_2). The lower benefit from innovative features not only leaves banks with less room to accommodate the increased expected liquidity cost from the higher settlement speed in a unified system but also makes it more difficult to compensate fintech firms to achieve the required reservation value. As a result, the BE strategy becomes more attractive than the BI strategy, and the region of the bank-led unified system shrinks (Regions I and III in Figure 4).

Figure 4 shows that either a unique equilibrium or multiple equilibria can emerge. In Region II, the fragmented system (FE+BE) emerges as a unique equilibrium. In Region IV, the fintech-led unified system (FI) emerges as a unique equilibrium. In contrast, in Regions I and III, two equilibria might emerge: the fintech-led or bank-led unified system. Note that the former is considered a superior equilibrium from the social perspective because it offers the same settlement speed but has more innovative payment features than the latter and results in higher total social welfare. To prevent the occurrence of multiple equilibria or the appearance of the inferior equilibrium, we next examine the role of government coordination in achieving the socially optimal outcome.

Government Coordination

The equilibrium outcomes that we have discussed thus far are based on the PSPs' voluntary participation without considering the role of the government. In reality, the regulator—the central bank or monetary authority in a nation—plays an important leading role in orchestrating the new retail payment system development. The government regulator is a social planner that aims to maximize the total social welfare in the entire economy. Social welfare is defined as the total social value minus the total social cost, where the social value comes from settling payment transactions, and the social cost consists of the infrastructure cost and the expected liquidity cost of payment settlement.

If the government attempts to prevent a socially inferior bank-led unified system or avoid a fragmented system for long-term strategic consideration, it might need to mandate the development of a government-led unified system. We next examine how the government can lead the innovation and help align the economic incentives of fintech firms and banks and provide not only a socially optimal but also a fair solution to all PSPs. We call this the *government mandate (GM)* strategy.

Under the GM strategy, the participation of both banks and fintech firms is a mandate to build the government-led nationwide new payment and settlement system. The government requires fintech firms to build the new system infrastructure because of its cost-efficiency, sets the cost-sharing rule $\{C_F, C_B\}$, where C_F and C_B are the system-building costs borne by fintech firms and banks and $C_F + C_B = \frac{k_F}{2}q_2^2$, and determines the optimal system design to maximize total social welfare.

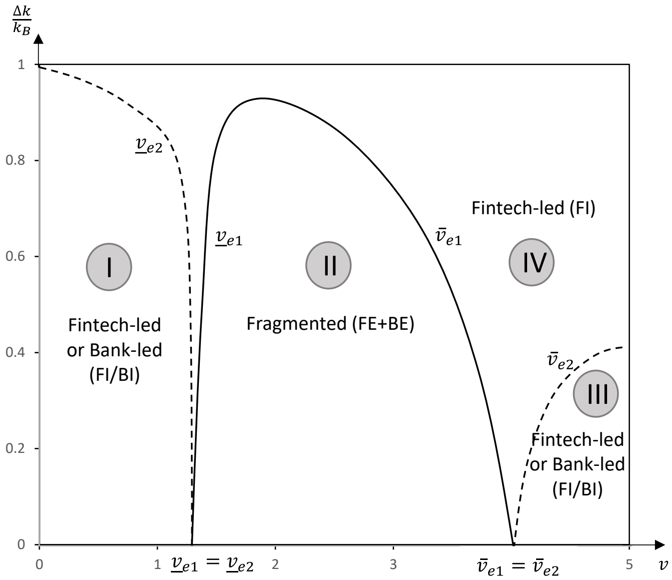

Figure 4. Equilibrium System Outcomes and Strategies

After the system has been built, it is provided as a national public service for all banks and fintech firms and is free to use. The government's optimization problem is formulated as:

$$max_{q_1,q_2}\ \Gamma v[w(q_1 - cq_1^2) + (1-w)q_2] + \Gamma\beta - \Gamma\theta q_1 \tilde{L} - \frac{k_F}{2}q_2^2 \tag{9}$$

$$s.t. \quad (1-\lambda)d_F\{v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta - \theta q_1 L_F\} - C_F \geq 0 \tag{10}$$

$$\lambda\{v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta - \theta q_1 L_B\} - C_B \geq 0 \tag{11}$$

$$C_F + C_B = \frac{k_F}{2}q_2^2 \tag{12}$$

Objective Function (9) is the total payment and settlement service value plus network value, net the total expected liquidity and system infrastructure costs. Although the regulator is able to mandate PSP participation, it needs to ensure that they do not incur negative payoffs to safeguard the stability of the entire financial system. Constraints (10) and (11) are the IC conditions for fintech firms and banks, respectively. Constraint (12) states that the new system's infrastructure cost is shared between fintech firms and banks. We next propose a fair Shapley-value-based cost-sharing rule.

The Shapley-value-based cost-sharing rule is widely recognized as a fair solution concept in cooperative game theory (Shapley, 1951; Roth, 1988). The Shapley values of banks (denoted as SP_B) and fintech firms (denoted as SP_F) define the distribution of the overall gain from cooperation: $SP_B + SP_F = SW^G$, where SW^G is the total surplus from cooperation (i.e., the social welfare of a government-led unified system). Under the Shapley-value-based cost-sharing rule, each participant needs to pay the difference between the value it obtains from using the new system and the marginal contribution of its participation (i.e., Shapley value). The detailed calculation of the Sharpley values for banks and fintech firms is presented in Appendix C (see proof of Proposition 5). We show that $SP_B = \frac{\pi_B^r + SW^G - \pi_F^r}{2}$ and that $SP_F = \frac{\pi_F^r + SW^G - \pi_B^r}{2}$. We can easily verify that $C_B + C_F = \frac{[v(1-w)\Gamma]^2}{2k_F}$; that is, the Shapley-value-based cost split between banks and fintech firms perfectly covers the total infrastructure cost. Proposition 5 (see below) summarizes the optimal system design of the government-led system $\{q_1^G, q_2^G\}$ and the optimal costs under the Shapely-value-based cost-sharing rule $\{C_B^*, C_F^*\}$.

Proposition 5 presents two important findings. First, the optimal system design under the GM strategy is the same as that under the fintech consortium's FI strategy: the three regions (BR, ER, and RR regions) are the same ($\underline{v}_G = \underline{v}_{FI}$ and $\overline{v}_G = \overline{v}_{FI}$), the settlement speed in each region is the same, and the optimal innovative payment features are the same. Second, each party's shared cost based on the Shapley-value rule varies in each region of v.

Proposition 5—Government mandate strategy: *When the government coordinates the new payment and settlement innovation, it employs the GM strategy to build a unified system for all participants.*

(i) The optimal system design is:

$$q_1^G = \begin{cases} 0 & if\ v \le \underline{v}_G \\ \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right) & if\ \underline{v}_G < v < \overline{v}_G \\ 1 & if\ v \ge \overline{v}_G \end{cases} \quad and\ q_2^G = \frac{v(1-w)\Gamma}{k_F},$$

where $\underline{v}_G = \frac{\tilde{L}\theta}{w}$ *and* $\overline{v}_G = \frac{\tilde{L}\theta}{(1-2c)w}$.

(ii) The optimal Shapely-value-based cost sharing is:

$$C_B^* = \begin{cases} C_B^{1*} = \frac{v^2(1-w)^2\lambda}{4}\left[\frac{3\lambda + 2(1-\lambda)d_F}{k_F} - \frac{\lambda}{k_B}\right] + \frac{1}{2}\lambda(1-\lambda)(1-d_F)\beta & if\ v \le \underline{v}_F \\ C_B^{2*} = C_B^{1*} + \frac{(1-\lambda)d_F(vw - \theta L_F)^2}{8cvw} & if\ \underline{v}_F \le v \le \underline{v}_{FI} \\ C_B^{3*} = C_B^{2*} + \frac{(vw - \theta\tilde{L})}{8cvw}\{[\lambda - (1-\lambda)d_F](vw + \theta\tilde{L}) - 2\theta[\lambda L_B - (1-\lambda)d_F L_F]\} & if\ \underline{v}_{FI} \le v \le \underline{v}_B \\ C_B^{4*} = C_B^{3*} - \frac{\lambda(vw - \theta L_B)^2}{8cvw} & if\ \underline{v}_B \le v \le \overline{v}_F \\ C_B^{5*} = C_B^{4*} - (1-\lambda)d_F\left[\frac{(vw - \theta L_F)^2}{8cvw} - \frac{vw(1-c) - \theta L_F}{2}\right] & if\ \overline{v}_F \le v \le \overline{v}_{FI} \\ C_B^{6*} = C_B^{5*} + \frac{[\lambda - (1-\lambda)d_F][v^2w^2(2c-1)^2 - \theta^2\tilde{L}^2]}{8cvw} + \frac{\theta[\lambda L_B - (1-\lambda)d_F L_F]}{2}\left(\frac{vw - \theta\tilde{L}}{2cvw} - 1\right) & if\ \overline{v}_{FI} \le v \le \overline{v}_B \\ C_B^{7*} = C_B^{6*} + \frac{\lambda(vw - \theta L_B)^2}{8cvw} - \frac{\lambda[vw(1-c) - \theta L_B]}{2} & if\ v > \overline{v}_B \end{cases}$$

and $C_F^* = \frac{[v(1-w)\Gamma]^2}{2k_F} - C_B^*$.

More importantly, we note that $SP_B = \frac{SW^G - \pi_F^r + \pi_B^r}{2} > \pi_B^r$ and $SP_F = \frac{\pi_F^r + SW^G - \pi_B^r}{2} > \pi_F^r$ when a unified system is socially optimal. Therefore, under Shapley-value cost sharing, all PSPs gain higher payoffs than their respective reservation values, suggesting that the Shapley-value cost-sharing rule enables a fair distribution of social value and prevents a "winner-takes-all" type of distribution of social gain.

We provide the following numerical example to illustrate the benefit and cost distribution between banks and fintech firms under the Shapley-value-based cost-sharing rule. We set $k_B = 10$ and $k_F = \{1, 4\}$; therefore, $\frac{\Delta k}{k_B} = \{0.9, 0.6\}$ represents different levels of fintech firms' technological advantage. We analyze $\lambda = \{0.5, 0.6\}$, where a larger value of $\lambda = 0.6$ represents a traditional payment system in which banks play a relatively larger role. We choose $d_F = 1$ to temporarily isolate the transaction size effect. We set $\theta = 0.8$, $L_F = 1$, and $L_B = 2$ such that fintech firms' expected liquidity cost per transaction is lower than that of the banks. We choose $w = 0.5$ for illustration. Under such parameter settings, both BI and FI can be the equilibrium if $v \le \underline{v}_{e2}$, and FI is the unique equilibrium otherwise. In the absence of government coordination, either banks or fintech firms reap all of the innovation benefits, whereas the other party only obtains their reservation value under the BI or FI strategy. When the government leads the innovation and sets the optimal Shapley-value-based cost-sharing rule, Figure 5 illustrates the share of social welfare (the top portion) and the percentage of total infrastructure cost borne by fintech firms (the bottom portion). The value of $\underline{v}_{e2}$ under each case is marked on the horizontal axis. The solid and dashed lines represent $\frac{\Delta k}{k_B} = 0.9$ and 0.6, respectively.

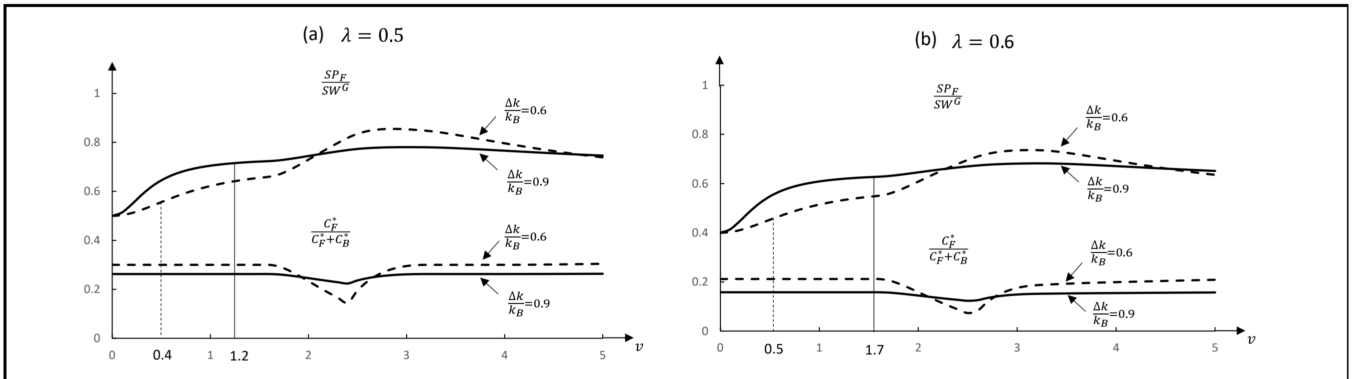

Figure 5. Illustration of Benefit and Cost Sharing under the Shapley-Value-Based Rule

We make several observations. First, as shown in Figure 5(a), because fintech firms contribute more to the unified system design, under the fair cost-sharing rule, they bear a smaller share of the total infrastructure cost (approximately 20-30%) and consequently enjoy a larger amount of the total welfare gain (approximately 50-85%). Fintech firms' higher relative technological efficiency $\frac{\Delta k}{k_B}$ generally results in them bearing a lower proportion of the total infrastructure cost, which could be viewed as a reward to fintech firms for contributing to a better system from which banks also benefit.

Second, fintech firms' share of the cost is the lowest when v falls in the intermediate range, which is when banks and fintech firms desire very different settlement speeds. To build a unified system, fintech firms must significantly accommodate banks; thus, the lower share of the infrastructure cost is a reward for their social contribution. Consequently, fintech firms' share of social benefits is also higher in this range than in other ranges.

Third, all else being equal, as the number of banks increases, fintech firms need to contribute more to building a unified system. As shown in Figure 5(b), the share of the infrastructure cost that fintech firms bear is further reduced in recognition of their efforts to accommodate the banks. In contrast, banks' share of the social benefit is computed as $\frac{SP_B}{SW^G} = 1 - \frac{SP_F}{SW^G}$. According to Figure 5, this value ranges from 15-50% when $\lambda = 0.5$ and increases to 25-60% when $\lambda = 0.6$, which contrasts sharply with the zero-gain outcome under the FI equilibrium, under which fintech firms take all of the surplus from banks. Moreover, when $v \leq \underline{v}_{e2}$, in addition to the FI equilibrium, BI can emerge as a socially inferior equilibrium (to the left of the vertical lines in the figure). Government coordination avoids the occurrence of this equilibrium. Overall, we find that the Shapley-value-based cost-sharing rule promotes the fair allocation of benefits and costs to all PSPs in the payment system—it not only appropriately rewards fintech firms for their contribution to system building but also safeguards banks' benefits and encourages their adoption of the new system.

Finally, the cost-sharing rule that governments can adopt is not unique. For example, transaction-based cost sharing is another rule that governments could consider. We calculate that fintech firms in the previous numerical example and under the transaction-based cost-sharing rule should bear 50% and 40% of the total cost when $\lambda = 0.5$ and $\lambda = 0.6$, respectively. Fintech firms apparently pay less under the Shapley-value-based cost-sharing rule in recognition of their technological efficiency and contribution to building the new system. In Appendix C, we provide further analysis for a general cost-sharing rule that the government might use to align each PSP's participation incentive and ensure the smooth operation of a unified system. We show that the regulator has a range of $\{C_F, C_B\}$ options and can thus control how the value of the payment and settlement innovation is distributed in the financial system, which to some extent exhibits the flexibility of governmental policy.

Socially Optimal Design

To identify the socially optimal design, we compare the social welfare in the government-led system (GM strategy) with equilibrium outcomes without government coordination, namely, the fragmented system (FE+BE strategy), the unified fintech-led system (FI strategy), and the unified bank-led system (BI strategy). Appendix C details the social welfare calculation under each equilibrium outcome. The following proposition presents our findings.

Proposition 6—social optimality and government coordination: *Three cases of socially optimal design exist, and the government plays different roles in each case:*

(i) If $v \in (\underline{v}_{e1}, \overline{v}_{e1}]$, then the fragmented system (FE+BE strategy) is socially optimal; government coordination is not needed because a government-led unified system leads to lower social welfare.

(ii) If $v \in (\underline{v}_{e2}, \underline{v}_{e1}]$ or $v \in (\overline{v}_{e1}, \overline{v}_{e2}]$, then both the unified fintech-led system (FI strategy) and the government-led system (GM strategy) are socially optimal; government coordination helps achieve a fair value distribution.

(iii) If $v \in (0, \underline{v}_{e2}]$ or $v > \overline{v}_{e2}$, the government-led unified system (GM strategy) is socially optimal; government coordination is necessary to prevent a socially inferior equilibrium.

When $v \in (\underline{v}_{e1}, \overline{v}_{e1}]$, the fragmented system emerges as the unique equilibrium under which both parties obtain their reservation value. In this region, fintech firms desire a much faster settlement speed than banks. Building two separate systems to satisfy the unique needs of fintech firms and banks is socially optimal, although independently developing two systems incurs dual infrastructure costs. Thus, in Region II, as shown in Figure 4, the fragmented system (FE+BE) is the unique equilibrium. This finding is consistent with our observation that multiple retail payment systems coexist in practice in many countries. The best strategy for governments is not to intervene in these private system innovations but to allow PSPs to develop and use their respective preferred systems.

When $v \in (\underline{v}_{e2}, \underline{v}_{e1}]$ or $v \in (\overline{v}_{e1}, \overline{v}_{e2}]$, referring to region IV in Figure 4, the fintech-led unified system emerges as an equilibrium without government coordination. This system has the same optimal design as the government-led unified system. Both the fee paid by banks in the fintech-led system and the side payment in the government-led system represent an internal value transfer between PSPs; thus, these fees will not affect the total social welfare. Hence, both the fintech-led and government-led unified systems are socially optimal. However, under the fintech-led unified system, banks only gain their reservation payoff, whereas fintech firms capture all of the gain from innovation. Under the government-led system, banks' side payments to fintech firms are based on the Shapley-value-based cost-sharing rule set by the government. Banks could gain more than their reservation value and, thus, enjoy the benefits of technological innovation. As a result, we conclude that government coordination in this region does not increase social welfare but plays the role of a "welfare redistributor."

When $v \in (0, \underline{v}_{e2}]$ or $v > \overline{v}_{e2}$, both the fintech-led and the bank-led unified systems can emerge as an equilibrium. Both systems result in the same settlement speed, but the fintech-led system has more innovative features than the bank-led one because of fintech firms' technological advantage. As a result, the fintech-led unified system yields higher social welfare than the bank-led unified system, as seen in Regions I and III in Figure 4. In these regions, two equilibria coexist, and the fintech-led system is socially superior. However, which equilibrium will emerge is unclear. Hence, we suggest that, in this case, the government needs to actively lead the innovation. Using the GM strategy not only avoids the appearance of the inferior (BI) equilibrium but also offers fair welfare sharing among PSPs.

In summary, our results lead to an important policy implication—in current payment system innovation, the major role of the government is not to mandate a different system than what can be developed by the most capable technological players. Instead, the government's intervention should focus on promoting and ensuring a fair distribution of the social value created by the new financial innovation and, thus, preventing a "winner-takes-all" outcome.

Discussion

In this section, we first focus on exploring the key drivers that influence the optimal system design. In line with our model prediction and the current global development of fast payment systems, we then discuss policy implications related to government oversight and regulatory considerations.

Key Drivers

Several key parameters jointly impact the equilibrium outcome: value (i.e., the base value parameter v), cost (i.e., the relative cost-efficiency $\frac{\Delta k}{k_F}$ or $\frac{\Delta k}{k_B}$), and payment volume (i.e., the ratio of banks' transaction volume to that of fintech firms $\frac{\lambda}{(1-\lambda)d_F}$).

A large λ suggests that there are increasingly more banks compared with fintech firms in the payments industry. A small d_F implies that fintech firms have not yet caught up with banks in terms of payment transaction demand. Therefore, a large $\frac{\lambda}{(1-\lambda)d_F}$ ratio indicates that the payment industry is still dominated by banks. In Figure 6, we illustrate how the regions of socially optimal systems shift as the payment system evolves from a traditional (largely dominated by banks) to an emerging (fintech firms playing increasingly important roles) ecosystem. Although the shapes of the boundary lines that define the equilibrium regions change under different parameter combinations, the qualitative insights remain the same.

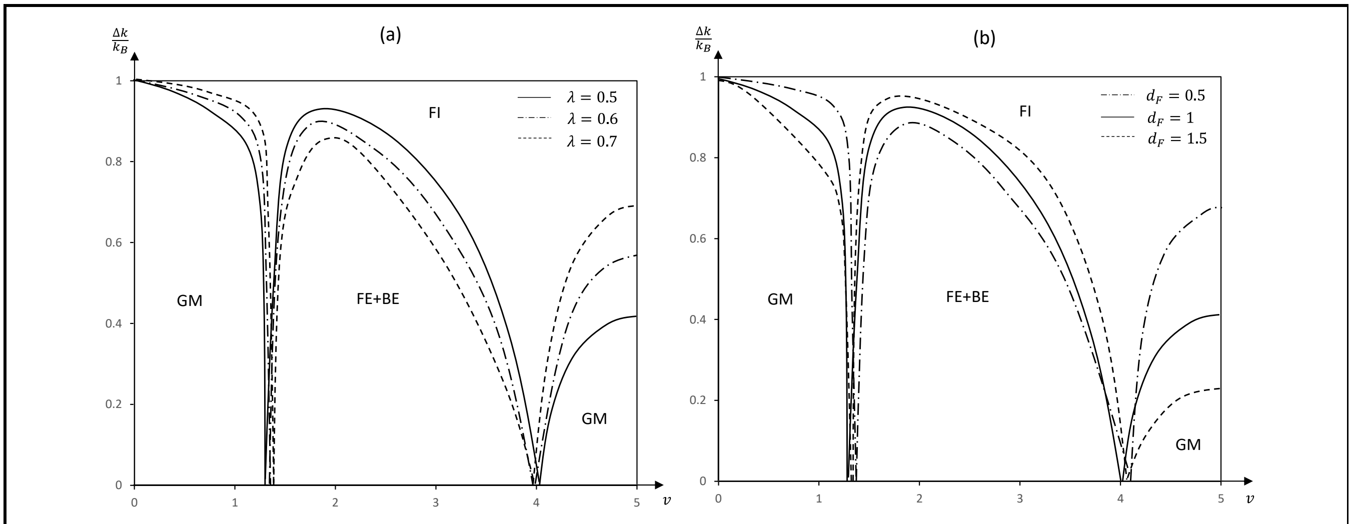

Figure 6. Effects of λ and d_F on Socially Optimal Leadership Strategies

The solid lines in Figure 6 show the equilibrium outcomes when $\lambda = 0.5$ and $d_F = 1$ (i.e., the transaction volume ratio $\frac{\lambda}{(1-\lambda)d_F} = 1$), which we use as a benchmark comparison. When the payment economy shifts from a traditional economy in which large banks maintain dominant positions to one in which fintech firms become an important player, namely, when either a large number of fintech firms exists (λ decreases) or the transaction volume of fintech firms picks up significantly (d_F increases), the region of the government-led system shrinks, whereas the region of the fintech-led system expands when v is small or large, and the region of fragmented systems expands when v is moderate.

When v is small, the equilibrium system design is a BR or ER system with a low settlement speed; when v is large, it is a high-speed ER or RR system. Under both cases, the views from all PSPs can be easily aligned, and a unified system is economically justified and thus welcomed by both banks and fintech firms. If fintech firms have a significant technological advantage (a large $\frac{\Delta k}{k_F}$), the government should allow these firms to voluntarily lead payment system innovation unless it would like to mandate a fair rule of welfare allocation. If fintech firms do not have a significant technological advantage (a small $\frac{\Delta k}{k_F}$), both FI and BI are likely to emerge as dual equilibria, and the government should lead innovation, play an active role in coordinating ecosystem development, and ensure a socially optimal equilibrium outcome.

When v is moderate, the preferences of banks and fintech firms cannot be easily aligned. Fintech firms desire a faster settlement speed that banks do not prefer to accommodate. Compared with a traditional economy in which large banks dominate, in an economy characterized by many fintech firms or a large number of transactions from fintech firms, such heterogeneous preferences are further amplified. Allowing banks and fintech firms to separately build their own preferred systems turns out to be socially desirable, particularly when fintech firms do not have a significant technological advantage (a small $\frac{\Delta k}{k_F}$). Therefore, in this region, the need is less for the government to coordinate the incentives among PSPs.

In addition, the value weight w dictates the trade-off between the settlement speed and innovative features. Figure 7 provides a sensitivity analysis on w to illustrate the shifts in the equilibrium regions of the socially optimal system design. We use $w = 0.5$ as the benchmark case (the solid line) in both Figures 7(a) and 7(b). As w decreases in Figure 7(a), the fragmented systems region (FE+BE) shrinks, and the unified system regions (FI or GM) expand. In particular, we see that when w is very small (e.g., $w = 0.1, 0.2$, or 0.3), the fragmented systems region (FE+BE) completely disappears in equilibrium. When the system design emphasizes speed improvement less and innovative features more, the misalignment of the ideal settlement speed between banks and fintech firms is less likely to become a hurdle in a unified system. Thus, a unified system is more likely to emerge in equilibrium. Following the same logic, we see that as w increases in Figure 7(b), the fragmented systems region (FE+BE) expands, and the unified system regions (FI or GM) shrink.

Similarly, as the liquidity cost parameters (L_F and L_B) or the probability of liquidity shocks (θ) decreases, the impact of accelerating payment processing becomes less of a critical concern. As a result, the incentives of all PSPs are more easily aligned, and the fragmented systems region shrinks.

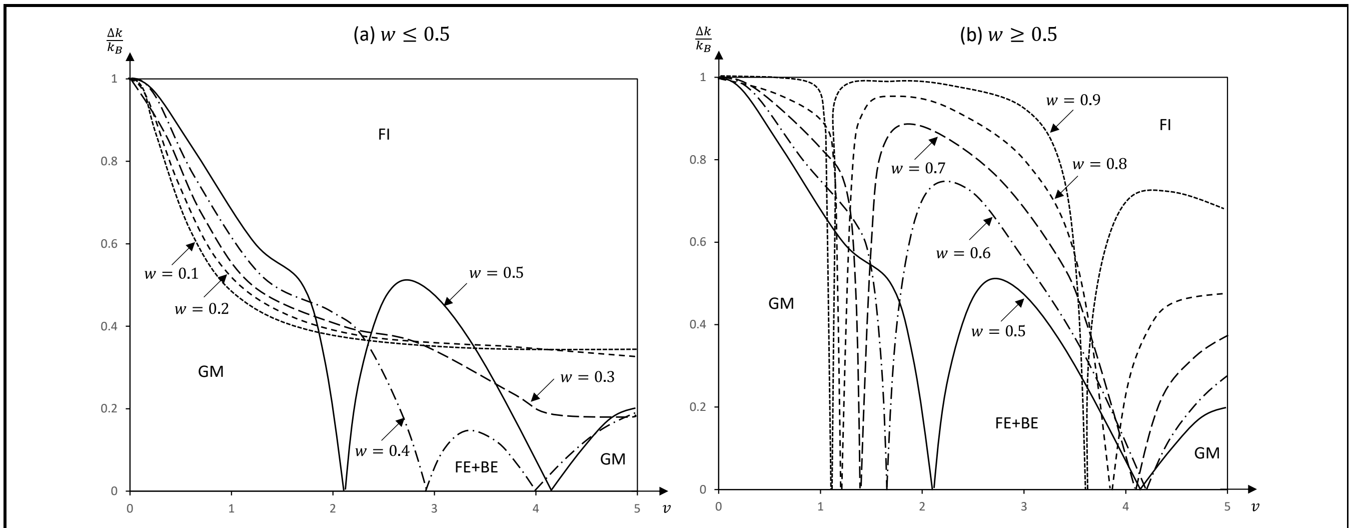

Figure 7. Effect of w on Socially Optimal System Design

Policy Implications

In this section, we discuss some emerging policy issues and provide regulatory suggestions for next-generation retail payment and settlement system design and implementation.

Although banks are still the dominant PSPs in many payment innovations, the rise of new entrants (nonfinancial institutions and tech firms) poses new policy challenges to regulators, especially in the wake of the proliferation of new technologies. For example, in the mobile payments market, two distinct models have emerged: the bank-led model and the fintech-led model. Sweden's mobile payment system Swish is an example of a bank-led model, as it was launched by six large Swedish banks, whereas China's mobile apps AliPay and WeChat Pay are examples of fintech-led models in which nonbank providers, typically technology companies or mobile network operators, can authorize transactions on their own platforms (Zhang, 2017).

In addition, blockchain shows its potential as an emerging payment technology. For example, IBM launched a real-time global blockchain-based payment network, IBM Blockchain World Wire, in which the majority of transaction volume will be retail remittances and consumers' e-commerce purchases.[10] However, many blockchain-based applications are initially structured as proprietary solutions, mainly because no agreed-upon standards yet exist in the market. If a new retail payment innovation aims to address the different needs of consumers and merchants and/or represent new payment instruments or payment channels, the new system would naturally have a low degree of interoperability. The lack of interoperability among different payment solutions, platforms, and networks may potentially lead to a fragmented payments market. Although having multiple exclusive systems to better serve different PSP heterogeneous needs could be socially optimal in certain cases, we note several disadvantages. Different groups of PSPs that independently develop multiple payment systems may lead to both technology investment waste and reduced network value (because of the smaller size of fragmented systems relative to a unified system). In addition, end users would have to manage multiple separate accounts because of the lack of interoperability among multiple exclusive systems and would thus incur inconvenience costs. Hence, over the long term, government regulators should facilitate the interoperability of multiple systems because payments play a crucial role in the functioning of the entire economy. Eurosystem TIPS[11] and Singapore's API payment gateway[12] are recent examples of regulators' efforts to promote a unified RR or ER system and achieve economies of scale.

As technological innovations radically reshape new forms of payment and new market developments alter traditional payment practices, changes cannot be left unguided. Thus, regulators' oversight and governance become necessary to orchestrate the dynamically evolving ecosystem and

[10] https://www.ibm.com/blockchain/solutions/world-wire

[11] TARGET Instant Payment Settlement (TIPS) is an infrastructure service launched by Eurosystem in November 2018. TIPS aims to minimize the risk of fragmentation into the European retail payments market by offering a service that any bank account holder in Europe can reach.

[12] The application programming interface (API) payment gateway was developed under the guidance of the Monetary Authority of Singapore. Starting in 2021, through the API payment gateway, nonbank financial institutions can directly connect to FAST and PayNow, which are fast e-payment services in Singapore.

harmonize payment system innovations. We recommend the following essential roles that government regulators could play in the rapidly evolving payments industry.

First, central banks need to be forward-looking and guide retail payment system innovation but can delegate system development to private sector players. For example, in Australia, a central bank strategic review identified the need for an RR system, NPP, to provide rich information, easy routing, and near-immediate funds availability on a 24×7 basis. NPP is mutually owned by 13 organizations as a result of unprecedented industry collaboration in response to the central bank's call for participation. The government acted as a catalyst and provided the outside impetus to empower and embrace innovations.

Second, government regulators could play a coordinating role to achieve public policy objectives striving for open and fair access, expanded participation, and a more efficient system. In practice, collaborative partnerships between banks and fintech firms are viewed as a useful complement to the existing payment service market. For example, IBM recently announced joining a blockchain financial network with 12 other banks to support cross-border and international transactions, forming a mutually beneficial collaboration partnership.[13] Banks can leverage the cutting-edge IT of nonfinancial institutions to offer diversified financial services. Tech firms, on the other hand, can leverage banks' existing network infrastructure to expand their reach. During this market transition, the government could introduce a regulatory sandbox to enable financial institutions and fintech players to experiment with innovative financial products or services within well-defined boundaries, comply with the relevant legal and regulatory requirements, and maintain the overall safety and soundness of the financial system.

Finally, we recognize that government mandates and oversight might have long-term strategic considerations that are beyond the economic benefit and cost analysis conducted in this study. For example, the Chinese government continues to clamp down on giant tech firms, including Alibaba and Tencent, implementing new laws designed to curtail private companies' dominance in the payments market. Moreover, the Central Bank of China is actively calling for banks' participation in developing the Digital Currency Electronic Payment (DCEP) system as the nationwide central-bank-based digital payment and processing network.[14] Such efforts aim to encourage banks to play a catch-up game in the Chinese mobile payments market and help achieve financial stability and balance market power to avoid a "winner-takes-all" outcome. In addition, given the long-run potential of the central bank's digital currency in subverting the power of the U.S. dollar in international trade, the development of DCEP has a high-level economic/political rationale to meet the country's strategic goal. Therefore, governments should effectively communicate with various market players, ensure their compliance with ongoing financial innovations, and build a robust ecosystem in the rapidly evolving payments market.

Conclusion

In response to dramatically increased e-commerce activities and retail transaction volume, the development of fast payment systems is one of the most important trends among recent financial innovations. In addition to faster settlement speeds, rich functionality allowing for value-added services is another essential element of advanced payment systems (Chapman et al., 2015). However, the new systems in different countries vary in terms of system operators, attributes, and architectures. Determining how fast settlements should be and how much functionality should be included in a new payment system is necessary and nontrivial. To date, no single best approach has been identified.

We developed an analytical model to study the optimal design for a fast payment and settlement system considering both settlement speed choice and the system capability supported by innovative payment functionality and features. We found that different parameters affect the optimal system design in different ways. As the base value of the payment service increases, both settlement speed and system capability increase. In contrast, while the expected liquidity cost negatively affects settlement speed, transaction volume and technological effectiveness positively affect system capabilities. These findings are consistent with our observations that higher payment values and lower liquidity costs lead to faster payment settlement and that new systems designed with richer payment features and functionality help attract and grow the demand for payment transactions.

We identified three distinct types of payment system designs in equilibrium: BR, ER, and RR. In regions with a small payment service value, the BR payment system supported by traditional DNS settlement is optimal, suggesting that accelerating the settlement of retail payment transactions is not necessary. In such a case, technological investment should focus on increasing the adoption of innovative payment features, such as adopting the new ISO 20022 messaging standard, supporting

[13] https://www.forbes.com/sites/robertanzalone/2020/05/21/ibm-doubles-down-on-blockchain-becomes-a-new-shareholder-in-wetrade-with-12-banks/?sh=7cbfd0b9724e

[14] https://www.afr.com/policy/economy/what-china-really-fears-about-its-big-tech-companies-20210818-p58jo9

mobile payments or peer-to-peer payments, and enhancing fraud detection abilities. In regions with intermediate payment service value, an ER payment system is optimal. The ER system operates at a higher settlement frequency than the traditional BR system and supports more innovative features and functions. In regions with a high payment service value, the RR payment system supported by RTGS settlement is optimal. In the RR system, the high value gained through fast transactions should outweigh liquidity concerns.

In terms of leadership strategy, from a social planner's perspective, government leadership is needed when the base value of payment services is very small or very large. Governments could adopt a mandate strategy to coordinate the co-development of the new system and ensure fair social welfare allocation. If the base value of payment services is in the intermediate range and fintech firms do not possess a significant technological advantage over banks, governmental coordination is unnecessary. Because different groups demand diverse system designs to satisfy their unique needs, banks and fintech firms should be given the opportunity to independently build their own preferred systems. If the base value of payment services is in the intermediate range and fintech firms exhibit a significant technological advantage, fintech-led innovation should be encouraged. In this case, if the government does not coordinate system development, fintech firms reap all the innovation benefits and banks only earn their reservation value. Alternatively, governments might coordinate innovation using a mandate strategy: designating fintech firms to develop and operate the system and requiring banks to pay a share of the infrastructure cost. We propose a Shapley-value-based cost-sharing rule to ensure a fair split of the social gains from cooperation.

As technologies offer faster and more effective ways to move funds, both fintech firms and banks are becoming actively involved in developing innovative payment systems. Central banks and other regulators worldwide need to establish rules and regulations to monitor compliance and achieve the public policy objective of promoting efficiency in the modernization and reform of retail payment systems. This research makes several policy recommendations regarding the socially optimal design of a nationwide fast payment infrastructure. First, although RTGS systems have the highest potential to be future-ready, not every country has chosen to pursue real-time settlement. Our results suggest that an RTGS system is optimal only when all PSPs in the financial ecosystem perceive that it has sufficiently high value. Second, although both fintech-led innovation and government mandates can yield the same socially optimal system design in a large range of our parameter space, to achieve fair value distribution, regulators should consider playing a leadership role by orchestrating the development and adoption of the new system. Third, heterogeneity among PSPs can give rise to coordination challenges. In a situation in which the payment service value is moderate and fintech firms do not possess a significant technological advantage over banks, we suggest that governments consider allowing the market to evolve itself and different PSPs to develop their own preferred systems. Blindly mandating cooperation among fintech firms and banks leads to suboptimal system designs that hurt social welfare. Overall, we identify both the opportunities and challenges that PSPs and regulators face in future payment system innovations. Our findings provide important policy insights into the design and implementation of the prevailing fast retail payment system development.

There are several directions for future research. First, our model is static and does not consider heterogeneous banks' adoption-timing decisions. In a dynamically evolving environment, banks might strategically choose to be innovation leaders or late adopters. Future research could build a two-period model to study banks' technological acceptance decisions and the overall adoption trend in the payments industry. Second, legacy systems and switching costs are other relevant factors affecting retail payment system innovation. For example, China's rapid development in the mobile payments market benefits from its immature credit card system, which enables Chinese consumers to leapfrog directly from cash to smartphones. In contrast, the United States has lagged behind in mobile and e-wallet payments markets because U.S. consumers are used to making credit card payments. U.S. regulatory inertia around promoting fast payments has also hindered payment innovation. Future research might consider the effects of legacy systems and switching costs on payment system innovation.

References

Allison, I. (2016). *R3 Connects 11 banks to distributed ledger using Ethereum and Microsoft Azure*. International Business Times. http://www.ibtimes.co.uk/r3-connects-11-banks-distributed-ledger-using-ethereum-microsoft-azure-1539044.

Ancri, C. (2016, October 19). *Fintech innovation: An overview.* Presented to the Board of Governors of the Federal Reserve System, Washington, DC. http://pubdocs.worldbank.org/en/767751477065124612/11-Fintech.pdf

Bapna, R., Goes, P., Wei, K. K., & Zhang, Z. (2011). A finite mixture logit model to segment and predict electronic payments system adoption. *Information Systems Research*, *22*(1), 118-133. https://doi.org/10.1287/isre.1090.0277

Bech, M. L., Shimizu, Y., & Wong, P. (2017). The quest for speed in payments. *BIS Quarterly Review*, *1*, 57-68. https://www.bis.org/publ/qtrpdf/r_qt1703g.htm

Bohme, R., Christin, N., Edelman, B., & Moore, T. (2015). Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives*, *29*(2), 213-238. http://dx.doi.org/10.1257/jep.29.2.213

Boston Consulting Group. (2012). *Cost benefit analysis of shortening the settlement cycle.* https://ccma-acmc.ca/en/wp-content/

uploads/Boston-Consulting-Gorup-Cost-Benefit-Analysis-on-Shortening-the-Settlement-Cycle-October-2012.pdf

Capgemini Research Institute. (2021). *World payments report 2021.* https://worldpaymentsreport.com/

Chapman, J., Chiu, J., Jafri, S., & Saiz, H. P. (2015). *Public policy objectives and the next generation of CPA systems: An analytical framework* (Bank of Canada discussion paper, No. 2015-6). https://www.bankofcanada.ca/2015/09/discussion-paper-2015-6/

Chiu, J., & Koeppl, T. V. (2019). Blockchain-based settlement for asset trading. *Review of Financial Studies*, *32*(5), 1716-1753. https://doi.org/10.1093/rfs/hhy122

Choi, J. P., & Thum, M. (1998). Market structure and the timing of technology adoption with network externalities. *European Economic Review*, *42*(2), 225-244. https://doi.org/10.1016/S0014-2921(97)00065-2

Clemons, E. K., & Weber, B.W. (1996). Alternative securities trading systems: Tests and regulatory implications of the adoption of technology. *Information Systems Research*, *7*(2), 163-188. http://dx.doi.org/10.1287/isre.7.2.163

Committee on Payments and market infrastructures (CPMI). (2016). *Fast payments: Enhancing the speed and availability of retail payments.* Bank for International Settlements. https://www.bis.org/cpmi/publ/d154.htm

Committee on Payment and Settlement Systems (CPSS). (2011). *Payment, clearing and settlement systems in the CPSS countries: Volume 1*, Bank for International Settlements. https://www.bis.org/cpmi/publ/d97.htm

Committee on Payment and Settlement Systems (CPSS). (2012). *Payment, clearing and settlement systems in the CPSS countries: Volume 2.* Bank for International Settlements. https://www.bis.org/cpmi/publ/d105.htm

Cong, L. W., & He, Z. (2019). Blockchain disruption and smart contracts. *Review of Financial Studies*, *32*(5), 1754-1797. https://doi.org/10.1093/rfs/hhz007

Embree, L., & Roberts, T. (2009). *Network analysis and Canada's large value transfer system.* (Bank of Canada discussion paper, No. 2009-13). https://www.bankofcanada.ca/2009/12/discussion-paper-2009-13/

Evans, D. S. (2016). *Economic aspects of Bitcoin and other decentralized public-ledger currency platforms* (University of Chicago working paper). https://chicagounbound.uchicago.edu/law_and_economics/680/

Farzin, Y. H., Huisman, K. J. M., & Kort, P. M. (1998). Optimal timing of technology adoption. *Journal of Economic Dynamics and Control*, *22*(5), 779-799. https://doi.org/10.1016/S0165-1889(97)00097-3

Goldstein, I., Jiang, W., & Karolyi, G. A. (2019). To fintech and beyond. *Review of Financial Studies*, *32*(5), 1647-1661. https://doi.org/10.1093/rfs/hhz025

Guo, Z. L., Kauffman, R., Lin, M., & Ma, D. (2015). Mechanism design for near real-time retail payment and settlement system. In *Proceedings of the 48th Annual Hawaii International Conference on System Science* (pp. 4824-4833). IEEE. https://ieeexplore.ieee.org/document/7070393

Hannan, T. H., & McDowell J. M. (1984). The determinants of technology adoption: The case of the banking firm. *Rand Journal of Economics*, *15*(3), 328-335. https://doi.org/10.2307/2555441

Johnson, K., McAndrews, J. J., & Soramäki, K. (2004). Economizing on liquidity with deferred settlement mechanisms. *Federal Reserve Bank of New York Economic Policy Review*, *10*(3), 56-72. https://www.newyorkfed.org/research/epr/04v10n3/0412mcan.html

Jung, J., Bapna, R., Ramaprasad, J., & Umyarov, A. (2019). Love unshackled: Identifying the effect of mobile app adoption in online dating. *MIS Quarterly*, *43*(1), 47-72. https://doi.org/10.25300/MISQ/2019/14289

Kahn, C., & Roberds, W. (2001). Real-time gross settlement and costs of immediacy. *Journal of Monetary Economics*, *47*(2), 299-319. https://doi.org/10.1016/S0304-3932(01)00047-2

Katz, M., & Shapiro, C. (1986). Technology adoption in the presence of network externalities. *Journal of Political Economy*, *94*(4), 822-841. http://www.jstor.org/stable/1833204

Kerr, S., & Newell, R. G. (2003). Policy-induced technology adoption: Evidence from the U.S. lead phasedown. *The Journal of Industrial Economics*, *51*(3), 317-343. https://doi.org/10.1111/1467-6451.00203

Keyes, D., & Magana, G. (2019). *REPORT: Chinese fintechs like Ant Financial's Alipay and Tencent's WeChat are rapidly growing their financial services ecosystems.* Business Insider. https://www.businessinsider.com/china-fintech-alipay-wechat.

Khapko, M., & Zoican, M. (2020). How fast should trades settle? *Management Science*, *66*(10), 4573-4593. https://doi.org/10.1287/mnsc.2019.3408

Kim, H., Chan, H., & Gupta, S. (2007). Value-based adoption of mobile internet: An empirical investigation. *Decision Support Systems*, *43*(1), 111-126. https://doi.org/10.1016/j.dss.2005.05.009

Leinonen, H., & Soramäki, K. (1999). *Optimizing liquidity usage and settlement speed in payment systems* (Bank of Finland research discussion paper, No. 16). Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3018083

Liu, J., Kauffman, R., & Ma, D. (2015). Competition, cooperation, and regulation: Understanding the evolution of the mobile payments technology ecosystem. *Electronic Commerce Research and Applications*, *14*(5), 372-291. https://doi.org/10.1016/j.elerap.2015.03.003

Maloumby-Bak, R. C., & Kingombe, C. (2016). *The quest to lower high remittance costs to Africa: A brief review of the use of mobile banking and Bitcoins* (Working paper). Available at http://ssrn.com/abstract=1769124.

Manning, M., Nier, E., & Schanz, J. (2009). *The economics of large-value payments and settlement: Theory and policy issues for central banks.* Oxford University Press.

Mills, D., Wang, K., Malone, B., Ravi, A., Marquardt, J., Chen, C., Badev, A., Brezinski, T., Fahy, L., Liao, K., Kargenian, V., Ellithorpe, M., Ng, W., & Baird, M. (2016). *Distributed ledger technology in payments, clearing, and settlement* (Washington: Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series 2016-095). https://doi.org/10.17016/FEDS.2016.095.

Ozdemir, Z., Barron, J., & Bandyopadhyay, S. (2011). An analysis of the adoption of digital health records under switching costs. *Information Systems Research*, *22*(3), 491-503. https://doi.org/10.1287/isre.1110.0349

Peng, G., & Dey, D. (2013). Research note—A dynamic view of the impact of network structure on technology adoption: The case of OSS development. *Information Systems Research*, *24*(4), 1087-1099. https://doi.org/10.1287/isre.2013.0494

Plouffe, C. R., Hulland, J. S., & Vandenbosch, M. (2001). Richness versus parsimony in modeling technology adoption decisions—Understanding merchant adoption of a smart card-based payment system. *Information Systems Research*, *12*(2), 208-222. https://doi.org/10.1287/isre.12.2.208.9697

Retana, G. F., Forman, C., Narasimhan, S., Niculescu, M. F., & Wu, D. J. (2018). Technology support and post-adoption IT service use: Evidence from the cloud. *MIS Quarterly*, *42*(3), 961-978. https://doi.org/10.25300/MISQ/2018/13064

Roth, A. E. (1988). *The Shapley value: Essays in honor of Lloyd S. Shapley*. Cambridge University Press. https://doi.org/10.1017/CBO9780511528446

Saloner, G., & Shepard, A. (1995). Adoption of technologies with network effects: An empirical examination of the adoption of automated teller machines. *RAND Journal of Economics*, *26*(3), 479-501. https://doi.org/10.2307/2555999

Shapley, L. S. (1951). *Notes on the n-person game—II: The value of an n-person game*. Research Memorandum RM-670. RAND Corporation. https://doi.org/10.7249/RM0670

Soramäki, K., Bech, M. L., Arnold, J., Glass, R. J. and Beyeler, W. E. (2007). The topology of interbank payment flows. *Physica A: Statistical Mechanics and its Applications*, *379*(1), 317-333. https://doi.org/10.1016/j.physa.2006.11.093

Tompkins, M., & Olivares, A. (2016). *Clearing and settlement systems from around the world: A qualitative analysis.* (Bank of Canada discussion paper, No. 14). https://www.bankofcanada.ca/2016/06/staff-discussion-paper-2016-14/

Wall, E., & Malm, G. (2016). *Using blockchain technology and smart contracts to create a distributed securities depository* (LUND University working paper). https://lup.lub.lu.se/student-papers/search/publication/8885750

Willison, M. (2005). *Real-time gross settlement and hybrid payment systems: A comparison* (Bank of England working paper, No. 252). https://www.bankofengland.co.uk/working-paper/2005/real-time-gross-settlement-and-hybrid-payment-systems

Zhang, M. (2017). China sets up clearing house for online payment services like AliPay and Tenpay. *South China Morning Post*. https://www.scmp.com/business/companies/article/2105825/china-sets-clearing-house-online-payment-services-alipay-and

About the Authors

Zhiling Guo (https://orcid.org/0000-0002-9058-4710) is an associate professor of information systems at Singapore Management University. She received her Ph.D. from The University of Texas at Austin. Her recent research focuses on platform business models, fintech innovations, digital payments, cloud computing, and data analytics. Her work appears in *Management Science*, *MIS Quarterly, Information Systems Research*, *INFORMS Journal on Computing*, *Journal of Management Information Systems*, *Production and Operations Management*, among other journals. She is currently an associate editor at *MIS Quarterly* and *INFORMS Journal on Computing* and a senior editor at *Production and Operations Management*.

Dan Ma (https://orcid.org/0000-0001-7206-7595) is an associate professor of information systems and management at Singapore Management University. She obtained her Ph.D. degree in computers and information systems from the University of Rochester. Her research interests focus on the strategic interactions of information technology (IT) and business—in particular, the economics of software-as-a-service business model analysis and the innovative technology adoption in the financial services market. Her works have been presented in international workshops and conferences including *International Conference on Information Systems (ICIS)*, *Workshop on Information Systems and Economics (WISE)*, *Hawaii International Conference on System Science (HICSS)*, and *Annual Workshop on Information Technologies and Systems (WITS)*; and appeared in top research journals such as *Information Systems Research, Management Information Systems Quarterly, Journal of Management Information Systems, and IEEE Transactions on Engineering Management.*

Appendix A

Worldwide Fast Payment System Design

Table A1. Three Types of Retail Payment System Design and Features

Types	Settlement model	Settlement frequency	Liquidity needs	Architecture attributes
Batch retail payment system (BR)	Deferred net settlement (DNS)	Once a day	Low	Slowest but cheapest design; cost efficiency of netting; enhanced functionality
Expedited retail payment system (ER)	More frequent netting	Multiple times a day	Medium	Timeliness of payments is linked to the system's settlement window; cost efficiency of netting largely remains; high functionality
Real-time retail payment system (RR)	Real-time gross settlement (RTGS)	Continuous	High	Fastest but most expensive payments; High liquidity cost; high functionality

Table A2. Representative Retail Payment and Settlement Systems in Selected Countries

Types	Country	Retail payment system	Leadership	Settlement frequency
BR	U.S.	FedACH	Central bank	Pending
BR	Canada	ACSS	Central bank	Pending
ER	Japan	Zengin	Central bank	4 times per day
ER	United Kingdom	FPS	Bank association	3 times per day
ER	Singapore	FAST	Bank association	2 times per day
ER	Korea	KFTC	Bank association	2 times per day
ER	China	IBPS	Central bank	Multiple times per day
RR	Sweden	BIR/Swish	Bank association	Real-time
RR	Australia	NPP	Bank association	Real-time
RR	Mexico	SPEI	Central bank	Real-time
RR	Eurosystem	TIPS	Central bank	Real-time

Table A3. Advanced Retail Payment System Capabilities and Attributes

System capabilities	Examples
System functionality	Batch file sorting and routing; item validation and data capture; automated messages and reports; standardized remittance information via ISO 20022; faster bilateral and multilateral clearing and settlement; 24x7 availability; interoperability
Design features	Types of entities involved (e.g., banks, nonfinancial institutions); types of accounts; types of transactions supported; interoperability; pricing and changes; clearing and settlement arrangements; security and fraud risk
Participant tools enabled	Convenient access channels (e.g., mobile banking); error and fraud detection; better authentication and validations; automated reconciliation and submission to settlement system; automated messages and reporting; real-time notification and alert services
Transaction types	Merchant payments; utility bill payments; person-to-person transfers; person-to-government payments; withdrawals; deposits; loan repayment; business-to-business payments; business-to-person payments; business-to-government payments; government-to-person payments; government-to-business payments; cross-border payments

Appendix B

Notation Table

Table B1. Notation Table

Notation	Definition
q_1	Settlement speed; $q_1 \in [0,1]$, where 0 represents the slowest DNS system and 1 the fastest RTGS system
q_2	System capability parameter indicating the number of innovative payment features of the new system
d_B, d_F	Volume of payment transactions for a bank and for a fintech firm; we normalize $d_B = 1$ and denote $d_F \overset{\text{def}}{=} d$
v	Base value parameter of the payment and settlement service
λ	Proportion of banks;$\lambda \in [0,1]$
w	Weight parameter of valuation between settlement speed and system capability; $w \in (0,1)$
c	Parameter capturing the diminishing-return of the settlement speed; $c \in [0,\frac{1}{2}]$
L_B, L_F	Average liquidity cost per payment transaction for banks and fintech firms; $L_B \geq L_F$
θ	Base probability of liquidity shock per payment transaction
β	Network intensity parameter
k_B, k_F	Cost parameter of infrastructure investment for banks and fintech; $k_B \geq k_F$
p	Price charged by a private system owner for other PSPs to use the payment and settlement service
π_B^r, π_F^r	Reservation value of the bank association and the fintech consortium; the respective payoff if they build private systems for their own exclusive use
π_i^j	Payoff of PSP i $(i = F, B)$ under the system outcome j, where $j =$ {FE, FI, BE, BI}
SW^j	Social welfare under the system outcome j, where $j =$ {FI, BI, BE+FE}
SW^G	Social welfare under the government-led unified system
C_B, C_F	System-building cost borne by banks and fintech firms under government-led innovation;$C_F + C_B = \frac{k_F}{2} q_2^2$
SP_B, SP_F	Shapley value of banks and fintech firms

Appendix C

Proofs of Lemmas and Propositions

Proof of Lemma 1

Under the FE strategy, fintech firms solve for the optimization problem characterized by Equation (3). Taking the first-order conditions with respect to q_1 and q_2 and solving the system of equations, we obtain the interior solution $q_1 = \frac{1}{2c}(1-\frac{\theta L_F}{vw}), q_2 = \frac{d_F(1-\lambda)v(1-w)}{k_F}$. Applying $0 \leq q_1 \leq 1$, we obtain two critical values, $\underline{v}_F = \frac{\theta L_F}{w}$ and $\overline{v}_F = \frac{\theta L_F}{(1-2c)w}$, such that:

(1) When $v < \underline{v}_F$, q_1 is bounded by the lower bound value 0. Thus, the optimal system design is $q_1^{FE} = 0, q_2^{FE} = \frac{d_F(1-\lambda)v(1-w)}{k_F}$, and $\pi_F^{r1} = \frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. This is a BR system.

(2) When $\underline{v}_F \leq v \leq \overline{v}_F$, we have the interior solution. Thus, the optimal system design is $q_1^{FE} = \frac{1}{2c}(1-\frac{\theta L_F}{vw}), q_2^{FE} = \frac{d_F(1-\lambda)v(1-w)}{k_F}$, and $\pi_F^{r2} = \frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw} + \frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. This is an ER system.

(3) When $v > \overline{v}_F$, q_1 is bounded by the upper bound value 1. Thus, the optimal system design is $q_1^{FE} = 1, q_2^{FE} = \frac{d_F(1-\lambda)v(1-w)}{k_F}$, and $\pi_F^{r3} = (1-\lambda)d_F[vw(1-c) - \theta L_F\,] + \frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. This is an RR system.

To ensure that the system is for fintech firms' own exclusive use, fintech firms set a high access price p^{FE} such that banks' participation incentive constraint is unsatisfied.

Proof of Lemma 2

Under the BE strategy, banks solve for the optimization problem characterized by Equation (4). Taking the first-order conditions with respect to q_1 and q_2 and solving the system of equations, we obtain the interior solution $q_1 = \frac{1}{2c}(1-\frac{\theta L_B}{vw}), q_2 = \frac{\lambda v(1-w)}{k_B}$. Applying the condition of $0 \leq q_1 \leq 1$, we obtain two critical values, $\underline{v}_B = \frac{\theta L_B}{w}$ and $\overline{v}_B = \frac{\theta L_B}{(1-2c)w}$, such that:

(1) When $v < \underline{v}_B$, q_1 is bounded by the lower bound value 0. Thus, the optimal system design is $q_1^{BE} = 0, q_2^{BE} = \frac{\lambda v(1-w)}{k_B}$, and $\pi_B^{r1} = \frac{\lambda^2v^2(1-w)^2}{2k_B} + \beta\lambda^2$. This is a BR system.

(2) When $\underline{v}_B \leq v \leq \overline{v}_B$, we have the interior solution. Thus, the optimal system design is $q_1^{BE} = \frac{1}{2c}(1-\frac{\theta L_B}{vw}), q_2^{BE} = \frac{\lambda v(1-w)}{k_B}$, and $\pi_B^{r2} = \frac{\lambda(vw-\theta L_B)^2}{4cvw} + \frac{\lambda^2v^2(1-w)^2}{2k_B} + \beta\lambda^2$. This is an ER system.

(3) When $v > \overline{v}_B$, q_1 is bounded by the upper bound value 1. Thus, the optimal system design is $q_1^{BE} = 1, q_2^{BE} = \frac{\lambda v(1-w)}{k_B}$, and $\pi_B^{r3} = \lambda[vw(1-c) - \theta L_B] + \frac{\lambda^2v^2(1-w)^2}{2k_B} + \beta\lambda^2$. This is an RR system.

To ensure that the system is for banks' own exclusive use, banks set a high access price p^{BE} such that fintech firms' participation incentive constraint is unsatisfied.

Proof of Proposition 1

(1) Because $L_F \leq L_B$, we obtain $\underline{v}_F \leq \underline{v}_B$ and $\overline{v}_F \leq \overline{v}_B$. In all intervals of v defined by these four boundary values, we can easily verify that $q_1^{FE} \geq q_1^{BE}$.

(2) Solving for $q_2^{FE} \geq q_2^{BE}$, we have $\frac{d_F(1-\lambda)}{k_F} \geq \frac{\lambda}{k_B}$, which is equivalent to $\frac{k_F}{k_B} \leq \frac{d_F(1-\lambda)}{\lambda}$.

Proof of Lemma 3

Under the FI strategy, the fintech consortium's optimization problem is characterized by Equations (5) and (6). Fintech firms always charge a price p to make constraint (6) binding. Taking the first-order conditions with respect to q_1 and q_2 and solving the system of equations, we obtain $q_1 = \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right)$ and $q_2 = \frac{v(1-w)\Gamma}{k_F}$, where $\tilde{L} \stackrel{\text{def}}{=} \frac{(1-\lambda)d_F L_F + \lambda L_B}{(1-\lambda)d_F + \lambda}$ and $\Gamma \stackrel{\text{def}}{=} (1-\lambda)d_F + \lambda$. Applying $0 \leq q_1 \leq 1$, we obtain two critical values, $\underline{v}_{FI} = \frac{\tilde{L}\theta}{w}$ and $\overline{v}_{FI} = \frac{\tilde{L}\theta}{(1-2c)w}$, such that

(1) When $v < \underline{v}_{FI}$, q_1 is bounded by the lower bound value 0. Thus, the optimal system design is $q_1^{FI} = 0, q_2^{FI} = \frac{v(1-w)[(1-\lambda)d_F+\lambda]}{k_F}$. This is a BR system.

(2) When $\underline{v}_{FI} \leq v \leq \overline{v}_{FI}$, we have the interior solution. Thus, the optimal system design is $q_1^{FI} = \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right), q_2^{FI} = \frac{v(1-w)[(1-\lambda)d_F+\lambda]}{k_F}$. This is an ER system.

(3) When $v > \overline{v}_{FI}$, q_1 is bounded by the upper bound value 1. Thus, the optimal system design is $q_1^{FI} = 1, q_2^{FI} = \frac{v(1-w)[(1-\lambda)d_F+\lambda]}{k_F}$. This is an RR system.

Next, we derive the price charged to banks for using the FI system. p is chosen by making constraint (6) binding: $p = v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta - \theta q_1 L_B - \pi_B^r/\lambda$. Note that the optimal system designs and banks' reservation values are different in different regions. We have already defined several border values: $\underline{v}_F < \overline{v}_F$ in Lemma 1, $\underline{v}_B < \overline{v}_B$ in Lemma 2, and $\underline{v}_{FI} < \overline{v}_{FI}$ in Lemma 3. We can show that $\underline{v}_F < \underline{v}_{FI} < \underline{v}_B$ and $\overline{v}_F < \overline{v}_{FI} < \overline{v}_B$, but the relative magnitudes of $\underline{v}_B$ and $\overline{v}_F$ might vary depending on the concrete parameter values. In the following, we present the detailed proof for the case of $\underline{v}_B \leq \overline{v}_F$, that is, when $\frac{L_F}{L_B} \geq 1 - 2c$. In this case, we have $\underline{v}_F < \underline{v}_{FI} < \underline{v}_B \leq \overline{v}_F < \overline{v}_{FI} < \overline{v}_B$. The other case of $\underline{v}_B > \overline{v}_F$, namely, when $\frac{L_F}{L_B} < 1 - 2c$, can be similarly analyzed.

We have the following five cases. In each case, we substitute into the optimal q_1, q_2 to obtain p:

$$p = \begin{cases} v^2(1-w)^2\left[\frac{\Gamma}{k_F} - \frac{\lambda}{2k_B}\right] + (1-\lambda)\beta & if\ v \leq \underline{v}_{FI}\ or\ v > \overline{v}_B \\ v^2(1-w)^2\left[\frac{\Gamma}{k_F} - \frac{\lambda}{2k_B}\right] + \left(1 - \frac{\tilde{L}\theta}{vw}\right)\frac{vw + \tilde{L}\theta - 2\theta L_B}{4c} + (1-\lambda)\beta & if\ \underline{v}_{FI} \leq v \leq \underline{v}_B \\ v^2(1-w)^2\left[\frac{\Gamma}{k_F} - \frac{\lambda}{2k_B}\right] + \left(1 - \frac{\tilde{L}\theta}{vw}\right)\frac{vw + \tilde{L}\theta - 2\theta L_B}{4c} - \frac{(vw - \theta L_B)^2}{4cvw} + (1-\lambda)\beta & if\ \underline{v}_B \leq v \leq \overline{v}_{FI} \\ v^2(1-w)^2\left[\frac{\Gamma}{k_F} - \frac{\lambda}{2k_B}\right] + vw(1-c) - \theta L_B - \frac{(vw - \theta L_B)^2}{4cvw} + (1-\lambda)\beta & if\ \overline{v}_{FI} \leq v \leq \overline{v}_B \end{cases}$$

Finally, we derive each PSP's payoffs and total social welfare. Fintech firms' profit can be written as $\pi_F^{FI} = \Gamma vw(q_1 - cq_1^2) + \Gamma\beta - \theta q_1[(1-\lambda)d_F L_F + \lambda L_B] + \Gamma v(1-w)q_2 - \frac{k_F}{2}q_2^2 - \pi_B^r$. We have the following five cases.

(1) When $v < \underline{v}_{FI}$: $\pi_B^{FI} = \pi_B^{r1} = \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \beta\lambda^2$, $\pi_F^{FI} = \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \Gamma\beta - \beta\lambda^2$, and $SW^{FI} = \frac{v^2\Gamma^2(1-w)^2}{2k_F} + \Gamma\beta$.

(2) When $v \in (\underline{v}_{FI}, \underline{v}_B)$: $\pi_B^{FI} = \pi_B^{r1} = \frac{\lambda^2 v^2 (1-w)^2}{2k_B} + \beta\lambda^2, \pi_F^{FI} = \frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv} + \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \Gamma\beta - \beta\lambda^2$, and $SW^{FI} = \frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv} + \frac{v^2\Gamma^2(1-w)^2}{2k_F} + \Gamma\beta$.

(3) When $v \in (\underline{v}_B, \overline{v}_{FI})$: $\pi_B^{FI} = \pi_B^{r2} = \frac{\lambda(vw-\theta L_B)^2}{4cvw} + \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \beta\lambda^2$, $\pi_F^{FI} = \frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv} + \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \left\{\frac{\lambda(vw-\theta L_B)^2}{4cvw} + \frac{\lambda^2 v^2(1-w)^2}{2k_B}\right\} + \Gamma\beta - \beta\lambda^2$, and $SW^{FI} = \frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv} + \frac{v^2\Gamma^2(1-w)^2}{2k_F} + \Gamma\beta$.

(4) When $v \in (\overline{v}_{FI}, \overline{v}_B)$: $\pi_B^{FI} = \pi_B^{r2} = \frac{\lambda(vw-\theta L_B)^2}{4cvw} + \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \beta\lambda^2, \pi_F^{FI} = \Gamma[vw(1-c) - \tilde{L}\theta] + \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \left\{\frac{\lambda(vw-\theta L_B)^2}{4cvw} + \frac{\lambda^2 v^2(1-w)^2}{2k_B}\right\} + \Gamma\beta - \beta\lambda^2$, and $SW^{FI} = \Gamma[vw(1-c) - \tilde{L}\theta] + \frac{v^2\Gamma^2(1-w)^2}{2k_F} + \Gamma\beta$.

(5) When $v > \overline{v}_B$: $\pi_B^{FI} = \pi_B^{r3} = \lambda[vw(1-c) - \theta L_B] + \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \beta\lambda^2$, $\pi_F^{FI} = \Gamma[vw(1-c) - \tilde{L}\theta] + \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \left\{\lambda[vw(1-c) - \theta L_B] + \frac{\lambda^2 v^2(1-w)^2}{2k_B}\right\} + \Gamma\beta - \beta\lambda^2$, and $SW^{FI} = \Gamma[vw(1-c) - \tilde{L}\theta] + \frac{v^2\Gamma^2(1-w)^2}{2k_F} + \Gamma\beta$.

Proof of Proposition 2

To find the optimal strategy, we need to compare fintech firms' profits under FE and FI. There are seven regions to compare, as subsequently shown for Cases 0-6.

Case 0. $\boldsymbol{v \in (0, \underline{v}_F)}$**:** Both FI and FE strategies yield the boundary solution $q_1^{FI} = q_1^{FE} = 0$, and banks' reservation value is their profit under the BE strategy when it yields a boundary solution $q_1^{BE} = 0$. Hence, we have $\pi_F^{FI} = \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \Gamma\beta - \beta\lambda^2$ and $\pi_F^{FE} = \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. Solving for $\pi_F^{FE} - \pi_F^{FI} \geq 0$, we have $\frac{\lambda(k_F - k_B) - 2d_F(1-\lambda)}{k_F k_B} - \frac{2(1-\lambda)(d_F+1)\beta}{v^2(1-w)^2} \geq 0$. This inequality never holds. Therefore, the FI strategy is always optimal in this region.

Case 1. $\boldsymbol{v \in (\underline{v}_F, \underline{v}_{FI})}$**:** The FI strategy yields a boundary solution $q_1^{FI} = 0$, the FE strategy results in an interior solution $q_1^{FE} > 0$, and banks' reservation value is their profit under the BE strategy when it yields a boundary solution $q_1^{BE} = 0$. Hence, we have $\pi_F^{FI} = \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \Gamma\beta - \beta\lambda^2$ and $\pi_F^{FE} = \frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw} + \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. Solving for $\pi_F^{FE} - \pi_F^{FI} \geq 0$, we have $\frac{(vw-\theta L_F)^2}{4cvw} \geq \frac{(1-w)^2 v^2 \lambda^2 \Delta k}{2k_F k_B (1-\lambda) d_F} + \frac{(1-w)^2 v^2 \lambda}{k_F} + \lambda(1-\lambda)(d_F+1)\beta$. When $v > \underline{v}_F$, $vw - \theta L_F > 0$. The inequality is equivalent to

$$vw - \theta L_F \geq \sqrt{\frac{4c(1-w)^2 v^3 \lambda w}{k_F}\left[\frac{1}{2k_B}\frac{\lambda\Delta k}{(1-\lambda)d_F} + 1\right] + \frac{4cvw\lambda(d_F+1)\beta}{d_F}} \tag{C1}$$

This is the condition for the FE strategy to be optimal in this region.

Note that the left-hand side (LHS) of condition (C1) is a linear function of v. The function inside the square root on the right-hand side (RHS) of condition (C1) is a cubic function of v. When $v = 0$, LHS < 0 and RHS = 0, violating condition (C1). When v increases, the RHS is always positive and strictly increases in v. When $v = \underline{v}_F$, LHS = 0 and RHS > 0, violating condition (C1). Therefore, the LHS and RHS functions at most intersect twice when $v > \underline{v}_F$. Let $v_2 > v_1$ be the two roots that solve equality (C1) if they exist. We have the following cases: (1) If $\underline{v}_{FI} < v_1$, then condition (C1) does not hold, and FI is always optimal when $\underline{v}_F < v \leq \underline{v}_{FI}$. (2) If $v_1 < \underline{v}_{FI} \leq v_2$, then condition (C1) does not hold, and FI is optimal when $v \in (\underline{v}_F, v_1]$, and condition (C1) holds and FE is optimal when $v \in (v_1, \underline{v}_{FI}]$. (3) If $v_2 < \underline{v}_{FI}$, then condition (C1) does not hold, and FI is optimal when $v \in (\underline{v}_F, v_1)$, condition (C1) holds, and FE is optimal when $v \in (v_1, v_2]$, and condition (C1) does not hold, and FI becomes optimal again when $v \in (v_2, \underline{v}_{FI}]$.

Overall, we summarize that in Case 1, FE is optimal in the range of $[\max(v_1, \underline{v}_F), \min(\underline{v}_{FI}, v_2)]$, and FI is optimal in the remaining feasible range.

Before moving to Case 2, we analyze the end point $v = \underline{v}_{FI}$. From Cases 1 to 2, the solution under FE remains the same (stay as an interior solution); the solution under FI changes from the boundary solution to the interior solution and is better off. Therefore, if FE is optimal in some subrange of Case 1 but at $v = \underline{v}_{FI}$, the optimal strategy has been switched to FI, then FI must remain optimal for the entire Case 2 range. In contrast, if at $v = \underline{v}_{FI}$ the optimal strategy is FE, then FE continues to dominate at least in some beginning subrange of Case 2.

Case 2. $\boldsymbol{v \in (\underline{v}_{FI}, \underline{v}_B)}$**:** Both the FI and FE strategies yield the interior solution $q_1^{FI} > 0, q_1^{FE} > 0$, and banks' reservation value is their profit under the BE strategy when it yields boundary solution $q_1^{BE} = 0$. Hence, we have $\pi_F^{FI} = \frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv} + \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \Gamma\beta - \beta\lambda^2$ and $\pi_F^{FE} = \frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw} + \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. Solving $\pi_F^{FE} - \pi_F^{FI} \geq 0$ yields

$$(1-\lambda)d_F\theta\big(2vw - \theta L_F - \theta\tilde{L}\big)\big(\tilde{L} - L_F\big) - \lambda\big(vw - \theta\tilde{L}\big)^2 \geq (1-\lambda)d_F\frac{4c(1-w)^2v^3\lambda w}{k_F}\left[\frac{1}{2k_B}\frac{\lambda\Delta k}{(1-\lambda)d_F} + 1\right] + 4cvw\lambda(1-\lambda)(d_F+1)\beta \quad \text{(C2)}$$

This is the condition for the FE strategy to be optimal in this region.

When $v = 0$, the LHS is negative and RHS is 0; thus, condition (C2) is unsatisfied. When $v > 0$, the LHS is a quadratic function of v, which opens downward, and the RHS is a cubic function of v, which strictly increases in v. Therefore, these two functions at most can intersect twice. Denote $v_4 > v_3$ as the two roots that satisfy the equality if they exist. By checking whether or not condition (C2) is satisfied, there are several cases: (1) If $\underline{v}_B < v_3$ or $v_4 < \underline{v}_{FI}$, then FI is always optimal when $\underline{v}_{FI} < v \leq \underline{v}_B$; (2) If $\underline{v}_{FI} < v_3 \leq \underline{v}_B \leq v_4$, then FI is optimal when $v \in (\underline{v}_{FI}, v_3]$ and FE is optimal when $v \in (v_3, \underline{v}_B]$; (3) If $\underline{v}_{FI} < v_3 \leq v_4 \leq \underline{v}_B$, then FI is optimal when $v \in (\underline{v}_{FI}, v_3]$, FE is optimal when $v \in (v_3, v_4]$, and FI becomes optimal again when $v \in (v_4, \underline{v}_B]$; (4) If $v_3 \leq \underline{v}_{FI} < \underline{v}_B \leq v_4$, then FE is always optimal over the entire range of $\underline{v}_{FI} < v \leq \underline{v}_B$; and (5) If $v_3 \leq \underline{v}_{FI} \leq v_4 \leq \underline{v}_B$, then FE is optimal in $v \in (\underline{v}_{FI}, v_4]$, and FI becomes optimal in $v \in (v_4, \underline{v}_B]$.

Overall, we summarize that in Case 2, the FE strategy is optimal in the range of $\left[\max(\underline{v}_{FI}, v_3), \min(v_4, \underline{v}_B)\right]$, and the FI strategy is optimal in the remaining feasible range.

Note that from Case 2 to 3, the solution under both FE and FI remains the same as the interior solution. Hence, if FE is optimal in some subrange of Case 2 but the optimal strategy switches to FI at $v = \underline{v}_B$, then FI must remain optimal for the entire Case 3 range. In contrast, if at $v = \underline{v}_B$ the optimal strategy is FE, then FE will continue to dominate at least in some beginning subrange of Case 3.

Case 3. $\boldsymbol{v \in (\underline{v}_B, \overline{v}_F)}$**:** Both FI and FE strategies yield the interior solution $q_1^{FI} > 0, q_1^{FE} > 0$, and banks' reservation value is their profit under the BE strategy when it yields interior solution $q_1^{BE} > 0$. Hence, we have $\pi_F^{FI} = \frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv} + \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \left\{\frac{\lambda(vw-\theta L_B)^2}{4cvw} + \frac{\lambda^2 v^2(1-w)^2}{2k_B}\right\} + \Gamma\beta - \beta\lambda^2$ and $\pi_F^{FE} = \frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw} + \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. Solving $\pi_F^{FE} - \pi_F^{FI} \geq 0$ yields

$$(1-\lambda)d_F\theta\big(2vw - \theta L_F - \theta\tilde{L}\big)\big(\tilde{L} - L_F\big) - \lambda\theta\big(2vw - \theta L_B - \theta\tilde{L}\big)\big(L_B - \tilde{L}\big) \geq$$

$$(1-\lambda)d_F\frac{4c(1-w)^2v^3\lambda w}{k_F}\left[\frac{1}{2k_B}\frac{\lambda\Delta k}{(1-\lambda)d_F} + 1\right] + 4cvw\lambda(1-\lambda)(d_F+1)\beta \quad \text{(C3)}$$

This is the condition for the FE strategy to be optimal in this region.

When $v > 0$, the LHS is a linear function of v, and the RHS is a cubic function of v, which strictly increases in v. Therefore, these two functions at most can intersect twice. Denote $v_6 > v_5$ as the two roots that satisfy the equality if they exist. By checking whether or not condition (C3) is satisfied, there are several cases: (1) If $\overline{v}_F < v_5$ or $v_6 < \underline{v}_B$, then FI is always optimal when $\underline{v}_B < v \leq \overline{v}_F$. (2) If $\underline{v}_B < v_5 \leq \overline{v}_F \leq v_6$, then FI is optimal when $v \in (\underline{v}_B, v_5]$ and FE is optimal when $v \in (v_5, \overline{v}_F]$. (3) If $\underline{v}_B < v_5 \leq v_6 \leq \overline{v}_F$, then FI is optimal when $v \in (\underline{v}_B, v_5]$, FE is optimal when $v \in (v_5, v_6]$, and FI becomes optimal again when $v \in (v_6, \overline{v}_F]$. (4) If $v_5 \leq \underline{v}_B < \overline{v}_F \leq v_6$, then FE is always optimal when $\underline{v}_B < v \leq \overline{v}_F$. (5) If $v_5 \leq \underline{v}_B \leq v_6 \leq \overline{v}_F$, then FE is optimal when $v \in (\underline{v}_B, v_6]$ and FI is optimal when $v \in (v_6, \overline{v}_F]$.

Overall, we summarize that in Case 3, the FE strategy is optimal in the range of $\left[\max(\underline{v}_B, v_5), \min(\overline{v}_F, v_6)\right]$, and the FI strategy is optimal in the remaining feasible range.

From Case 3 to 4, the solution under FI remains the same as the interior solution; the solution under FE switches from the interior solution to the boundary solution $q_1^{FE} = 1$ and thus worsens. Therefore, if FE is optimal in some subrange of Case 3 and at $v = \overline{v}_F$, the optimal strategy has already switched to FI, then FI must remain optimal for the entire Case 4 range. In contrast, if at $v = \overline{v}_F$ the optimal strategy is FE, then FE will continue to dominate at least in some beginning subrange of Case 4.

Case 4. $\boldsymbol{v \in (\overline{v}_F, \overline{v}_{FI})}$: The FI strategy yields the interior solution $q_1^{FI} > 1$, the FE strategy results in the boundary solution $q_1^{FE} = 1$, and banks' reservation value is their profit under the BE strategy when it yields the interior solution $q_1^{BE} > 0$. Hence, we have $\pi_F^{FI} = \frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv} +$

$\frac{v^2\Gamma^2(1-w)^2}{2k_F} - \left\{\frac{\lambda(vw-\theta L_B)^2}{4cvw} + \frac{\lambda^2 v^2(1-w)^2}{2k_B}\right\} + \Gamma\beta - \beta\lambda^2$ and $\pi_F^{FE} = (1-\lambda)d_F[vw(1-c) - \theta L_F] + \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F \beta$. Solving $\pi_F^{FE} - \pi_F^{FI} \geq 0$ yields:

$$-\lambda\theta(2vw - \theta L_B - \theta\tilde{L})(L_B - \tilde{L}) - (1-\lambda)d_F(vw - \theta\tilde{L})^2 \geq$$

$$(1-\lambda)d_F \frac{4c(1-w)^2 v^3 \lambda w}{k_F}\left[\frac{1}{2k_B}\frac{\lambda\Delta k}{(1-\lambda)d_F} + 1\right] - 4cvw(1-\lambda)d_F[vw(1-c) - \theta L_F] + 4cvw\lambda(1-\lambda)(d_F+1)\beta \tag{C4}$$

This condition makes the FE strategy optimal in this region.

When $v > 0$, the LHS is a quadratic function of v that opens downward, and the RHS is a cubic function of v that strictly increases in v. Therefore, these two functions at most can intersect twice. Denote $v_8 > v_7$ as the two roots that satisfy the equality if they exist. By checking whether or not condition (C4) is satisfied, we have the following cases: (1) If $\overline{v}_{FI} < v_7$ or $v_8 < \overline{v}_F$, then FI is always optimal when $\overline{v}_F < v \leq \overline{v}_{FI}$; (2) If $\overline{v}_F < v_7 \leq \overline{v}_{FI} \leq v_8$, then FI is optimal when $v \in (\overline{v}_F, v_7]$ and FE is optimal when $v \in (v_7, \overline{v}_{FI}]$; (3) If $\overline{v}_F < v_7 \leq v_8 \leq \overline{v}_{FI}$, then FI is optimal when $v \in (\overline{v}_F, v_7]$, FE is optimal when $v \in (v_7, v_8]$, and FI is optimal again when $v \in (v_8, \overline{v}_{FI}]$; (4) If $v_7 \leq \overline{v}_F \leq \overline{v}_{FI} \leq v_8$, then FE is optimal when $\overline{v}_F < v \leq \overline{v}_{FI}$; and (5) If $v_7 \leq \overline{v}_F \leq v_8 \leq \overline{v}_{FI}$, then FE is optimal when $v \in (\overline{v}_F, v_8]$ and FI is optimal when $v \in (v_8, \overline{v}_{FI}]$.

Overall, we summarize that in Case 4, FE is optimal in the range of $[\max(\overline{v}_F, v_7), \min(v_8, \overline{v}_{FI})]$, and FI is optimal in the remaining feasible range.

From Cases 4 to 5, the solution under FI switches from the interior solution to the boundary solution; the solution under FE remains the interior solution. The FI system stays at the highest speed, and banks' most desirable q_1^{BE} keeps increasing and approaching 1. It becomes easier for fintech firms to accommodate banks in a unified system. Therefore, if the optimal strategy is FI at the end point of Case 4, $v = \overline{v}_{FI}$, FI will always be better than FE in Case 5. In contrast, if at $v = \overline{v}_{FI}$ the optimal strategy is FE, then FE will continue to dominate at least in some beginning subrange of Case 5.

Case 5. $\boldsymbol{v \in (\overline{v}_{FI}, \overline{v}_B)}$**:** Both the FI and FE strategies yield the boundary solution $q_1^{FI} = 1, q_1^{FE} = 1$, and banks' reservation value is their profit under the BE strategy when it yields the interior solution $q_1^{BE} > 0$. Hence, we have $\pi_F^{FI} = \Gamma[vw(1-c) - \tilde{L}\theta] + \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \left\{\frac{\lambda(vw-\theta L_B)^2}{4cvw} + \frac{\lambda^2 v^2(1-w)^2}{2k_B}\right\} + \Gamma\beta - \beta\lambda^2$ and $\pi_F^{FE} = (1-\lambda)d_F[vw(1-c) - \theta L_F] + \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. Solving $\pi_F^{FE} - \pi_F^{FI} \geq 0$ yields

$$(vw - \theta L_B) \geq \sqrt{(1-\lambda)d_F \frac{4c(1-w)^2 v^3 w}{k_F}\left[\frac{1}{2k_B}\frac{\lambda\Delta k}{(1-\lambda)d_F} + 1\right] - \frac{4cvw}{\lambda}(1-\lambda)d_F\theta(\tilde{L} - L_F) + 4cvw[vw(1-c) - \theta\tilde{L}] + 4cvw(1-\lambda)(d_F+1)\beta} \tag{C5}$$

This is the condition for the FE strategy to be optimal in this region.

We first note that when $v = 0$, the LHS is negative and the RHS is 0. Therefore, this condition is unsatisfied. Note that the LHS of condition (C5) is a linear (increasing) function of v. Inside the square root on the RHS, it is a cubic function of v with a positive coefficient of v^3. Therefore, the LHS and RHS functions intersect twice at most. Let $v_{10} > v_9$ be the two roots that solve the binding condition (C5) if they exist. We have the following cases: (1) If $\overline{v}_B < v_9$ or $v_{10} < \overline{v}_{FI}$, then condition (C5) does not hold, and FI is always optimal when $\overline{v}_{FI} < v \leq \overline{v}_B$; (2) If $\overline{v}_{FI} < v_9 \leq \overline{v}_B \leq v_{10}$, then FI is optimal when $v \in (\overline{v}_{FI}, v_9]$ and FE is optimal when $v \in (v_9, \overline{v}_B]$; (3) If $\overline{v}_{FI} < v_9 \leq v_{10} \leq \overline{v}_B$, then FI is optimal when $v \in (\overline{v}_{FI}, v_9]$, FE is optimal when $v \in (v_9, v_{10}]$, and FI is optimal again when $v \in (v_{10}, \overline{v}_B]$; (4) If $v_9 \leq \overline{v}_{Fc} \leq \overline{v}_B \leq v_{10}$, then FE is always optimal when $\overline{v}_{FI} < v \leq \overline{v}_B$; and (5) If $v_9 \leq \overline{v}_{FI} \leq v_{10} \leq \overline{v}_B$, then FE is optimal when $v \in (\overline{v}_{FI}, v_{10}]$ and FI is optimal when $v \in (v_{10}, \overline{v}_B]$.

Overall, we summarize that in Case 5, FE is optimal in the range of $[\max(\overline{v}_{FI}, v_9), \min(v_{10}, \overline{v}_B)]$, and FI is optimal in the remaining feasible range.

Case 6. $\boldsymbol{v \geq \overline{v}_B}$**:** Both the FI and FE strategies yield the boundary solution $q_1^{FI} = 1, q_1^{FE} = 1$, and banks' reservation value is their profit under the BE strategy when it yields boundary solution $q_1^{BE} = 1$. Hence, we have $\pi_F^{FI} = \Gamma[vw(1-c) - \tilde{L}\theta] + \frac{v^2\Gamma^2(1-w)^2}{2k_F} - \left\{\lambda[vw(1-c) - \theta L_B] + \frac{\lambda^2 v^2(1-w)^2}{2k_B}\right\} + \Gamma\beta - \beta\lambda^2$ and $\pi_F^{FE} = (1-\lambda)d_F[vw(1-c) - \theta L_F] + \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$. Solving

$\pi_F^{FE} - \pi_F^{FI} \geq 0$ yields

$$(1-\lambda)d_F \frac{(1-w)^2 v^2 \lambda}{k_F}\left[\frac{1}{2k_B}\frac{\lambda\Delta k}{(1-\lambda)d_F} + 1\right] + (1-\lambda)\lambda(d_F+1)\beta \leq 0 \tag{C6}$$

This condition is never satisfied. Therefore, FI always dominates FE in the entire range of Case 6.

Putting all of the cases together, we conclude that the fintech consortium might adopt the exclusive strategy in a middle subrange of v. We denote this FE-optimality region as $(\underline{v}_{e1}, \overline{v}_{e1})$, where $\underline{v}_{e1}$ and $\overline{v}_{e1}$ might take different values (as described in the previous proof of Cases 1-5) depending on the concrete parameter values. In summary, the fintech consortium's optimal strategy, as v increases, will follow the pattern of "inclusive—exclusive—inclusive." Note that this FE-optimality region could be empty; if it appears, it will only appear once. Furthermore, the three terms, $\frac{\Delta k}{k_B}$, $\frac{\lambda}{(1-\lambda)d_F}$, and β, appear in all of the conditions. When they increase, the RHS of conditions (C1) - (C5) increases, thus making the conditions more difficult to hold. Therefore, the FE-optimality region is less likely to appear.

Proof of Lemma 4

Under the BI strategy, the bank association's optimization problem is characterized by Equations (7) and (8). Banks will always charge a price p to make constraint (8) binding. Taking the first-order conditions with respect to q_1 and q_2 and solving the system of equations, we obtain $q_1 = \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right), q_2 = \frac{v(1-w)\Gamma}{k_B}$, where $\Gamma = (1-\lambda)d_F + \lambda$ as defined in Lemma 3. Applying the condition of $0 \leq q_1 \leq 1$, we obtain two critical values, $\underline{v}_{BI} = \frac{\tilde{L}\theta}{w}$ and $\overline{v}_{BI} = \frac{\tilde{L}\theta}{(1-2c)w}$, such that:

(1) When $v < \underline{v}_{BI}$, q_1 is bounded by the lower bound value 0. Thus, the optimal system design is $q_1^{BI} = 0, q_2^{BI} = \frac{v(1-w)\Gamma}{k_B}$. This is a BR system.

(2) When $\underline{v}_{BI} \leq v \leq \overline{v}_{BI}$, we have the interior solution. Thus, the optimal system design is $q_1^{BI} = \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right), q_2^{BI} = \frac{v(1-w)\Gamma}{k_B}$. This is an ER system.

(3) When $v > \overline{v}_{BI}$, q_1 is bounded by the upper bound value 1. Thus, the optimal system design is $q_1^{BI} = 1, q_2^{BI} = \frac{v(1-w)\Gamma}{k_F}$. This is an RR system.

Next, we derive the price charged to fintech firms for using the BI system. p is given by setting constraint (8) binding: $p = d_F v[w(q_1 - cq_1^2) + (1-w)q_2] + d_F\beta - d_F\theta q_1 L_F - \pi_F^r/(1-\lambda)$.

Because the optimal system designs and fintech firms' reservation value are different in different regions, we have the following five cases. Define $p_0 = v^2(1-w)^2 d_F\left[\frac{\Gamma}{k_B} - \frac{d_F(1-\lambda)}{2k_F}\right] + \lambda d_F\beta$. In each case, we substitute into the optimal q_1, q_2 to obtain the optimal price p:

$$p = \begin{cases} p_0 & if\ v \leq \underline{v}_F\ or\ v > \overline{v}_{BI} \\ p_0 - \dfrac{d_F(vw - \theta L_F)^2}{4cvw} & if\ \underline{v}_F \leq v \leq \underline{v}_{BI} \\ p_0 - \dfrac{d_F(vw - \theta L_F)^2}{4cvw} + d_F\left(1 - \dfrac{\tilde{L}\theta}{vw}\right)\dfrac{vw + \tilde{L}\theta - 2\theta L_F}{4c} & if\ \underline{v}_{BI} \leq v \leq \overline{v}_F \\ p_0 - d_F[vw(1-c) - \theta L_F] + d_F\left(1 - \dfrac{\tilde{L}\theta}{vw}\right)\dfrac{vw + \tilde{L}\theta - 2\theta L_F}{4c} & if\ \overline{v}_F \leq v \leq \overline{v}_{BI} \end{cases}$$

Finally, we derive each PSP's payoffs and total social welfare. The profit of banks can be written as $\pi_B^{BI} = \Gamma vw(q_1 - cq_1^2) - \theta q_1[(1-\lambda)d_F L_F + \lambda L_B] + \Gamma v(1-w)q_2 + \Gamma\beta - \frac{k_B}{2}q_2^2 - \pi_F^r$. We have the following five cases.

(1) When $v < \underline{v}_F$: $\pi_F^{BI} = \pi_F^{r1} = \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$, $\pi_B^{BI} = \frac{v^2\Gamma^2(1-w)^2}{2k_B} - \frac{d_F^2(1-\lambda)^2 v^2(1-w)^2}{2k_F} + \Gamma\beta - (1-\lambda)^2 d_F\beta$, and $SW^{BI} =$

$\frac{v^2\Gamma^2(1-w)^2}{2k_B}+\Gamma\beta$.

(2) When $v \in (\underline{v}_F, \underline{v}_{BI})$: $\pi_F^{BI}=\pi_F^{r2}=\frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw}+\frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F}+(1-\lambda)^2d_F\beta$, $\pi_B^{BI}=\frac{v^2\Gamma^2(1-w)^2}{2k_B}-\left\{\frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw}+\frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F}\right\}+\Gamma\beta-(1-\lambda)^2d_F\beta$, and $SW^{BI}=\frac{v^2\Gamma^2(1-w)^2}{2k_B}+\Gamma\beta$.

(3) When $v \in (\underline{v}_{BI}, \overline{v}_F)$: $\pi_F^{BI}=\pi_F^{r2}=\frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw}+\frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F}+(1-\lambda)^2d_F\beta$, $\pi_B^{BI}=\frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv}+\frac{v^2\Gamma^2(1-w)^2}{2k_B}-\left\{\frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw}+\frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F}\right\}+\Gamma\beta-(1-\lambda)^2d_F\beta$, and $SW^{BI}=\frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv}+\frac{v^2\Gamma^2(1-w)^2}{2k_B}+\Gamma\beta$.

(4) When $v \in (\overline{v}_F, \overline{v}_{BI})$: $\pi_F^{BI}=\pi_F^{r3}=(1-\lambda)d_F[vw(1-c)-\theta L_F]+\frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F}+(1-\lambda)^2d_F\beta$, $\pi_B^{BI}=\frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv}+\frac{v^2\Gamma^2(1-w)^2}{2k_B}-\left\{(1-\lambda)d_F[vw(1-c)-\theta L_F]+\frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F}\right\}+\Gamma\beta-(1-\lambda)^2d_F\beta$, and $SW^{BI}=\frac{\Gamma(vw-\tilde{L}\theta)^2}{4cwv}+\frac{v^2\Gamma^2(1-w)^2}{2k_B}+\Gamma\beta$.

(5) When $v > \overline{v}_{BI}$: $\pi_F^{BI}=\pi_F^{r3}=(1-\lambda)d_F[vw(1-c)-\theta L_F]+\frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F}+(1-\lambda)^2d_F\beta$, $\pi_B^{BI}=\Gamma[vw(1-c)-\tilde{L}\theta]+\frac{v^2\Gamma^2(1-w)^2}{2k_B}-\left\{(1-\lambda)d_F[vw(1-c)-\theta L_F]+\frac{d_F^2(1-\lambda)^2v^2(1-w)^2}{2k_F}\right\}+\Gamma\beta-(1-\lambda)^2d_F\beta$, and $SW^{BI}=\Gamma[vw(1-c)-\tilde{L}\theta]+\frac{v^2\Gamma^2(1-w)^2}{2k_B}+\Gamma\beta$.

Proof of Proposition 3

To find the optimal strategy, we need to compare the bank association's profits under BE and BI. There are seven regions to compare, as subsequently shown for Cases 0-6. Note that $\underline{v}_{BI}=\underline{v}_{FI}$ and $\overline{v}_{BI}=\overline{v}_{FI}$; thus, the division of regions is the same as that in the proof of Proposition 2. The analysis of each case also follows the same approach as that in the proof of Proposition 2. Therefore, we omit the details and only present the following conditions, (C0') - (C6'), for the BE strategy to be optimal in the seven cases.

Case 0. $\boldsymbol{v \in (0, \underline{v}_F)}$**:**

$$\frac{\Delta k}{k_F}\geq\frac{2\lambda}{(1-\lambda)d_F}+\frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2} \tag{C0'}$$

Case 1. $\boldsymbol{v \in (\underline{v}_F, \underline{v}_{BI})}$**:**

$$vw-\theta L_F\geq\sqrt{\frac{4c(1-w)^2v^3\lambda w}{k_B}\left[1-\frac{1}{2k_F}\frac{(1-\lambda)d_F\Delta k}{\lambda}\right]+\frac{4cvw\lambda(d_F+1)\beta}{d_F}} \tag{C1'}$$

Case 2. $\boldsymbol{v \in (\underline{v}_{BI}, \underline{v}_B)}$**:**

$$(1-\lambda)d_F\theta\left(2vw-\theta L_F-\theta\tilde{L}\right)\left(\tilde{L}-L_F\right)-\lambda\left(vw-\theta\tilde{L}\right)^2\geq(1-\lambda)d_F\frac{4c(1-w)^2v^3\lambda w}{k_B}\left[1-\frac{1}{2k_F}\frac{(1-\lambda)d_F\Delta k}{\lambda}\right]+4cvw\lambda(1-\lambda)(d_F+1)\beta \tag{C2'}$$

Case 3. $\boldsymbol{v \in (\underline{v}_B, \overline{v}_F)}$**:**

$$(1-\lambda)d_F\theta\left(2vw-\theta L_F-\theta\tilde{L}\right)\left(\tilde{L}-L_F\right)-\lambda\theta\left(2vw-\theta L_B-\theta\tilde{L}\right)\left(L_B-\tilde{L}\right)\geq$$

$$(1-\lambda)d_F\frac{4c(1-w)^2v^3\lambda w}{k_B}\left[1-\frac{1}{2k_F}\frac{(1-\lambda)d_F\Delta k}{\lambda}\right]+4cvw\lambda(1-\lambda)(d_F+1)\beta \tag{C3'}$$

Case 4. $\boldsymbol{v \in (\overline{v}_F, \overline{v}_{BI})}$:

$$-\lambda\theta\left(2vw-\theta L_B-\theta\tilde{L}\right)\left(L_B-\tilde{L}\right)-(1-\lambda)d_F\left(vw-\theta\tilde{L}\right)^2\geq$$

$$(1-\lambda)d_F\frac{4c(1-w)^2v^3\lambda w}{k_B}\left[1-\frac{1}{2k_F}\frac{(1-\lambda)d_F\Delta k}{\lambda}\right]-4cvw(1-\lambda)d_F[vw(1-c)-\theta L_F]+4cvw\lambda(1-\lambda)(d_F+1)\beta \tag{C4'}$$

Case 5. $\boldsymbol{v \in (\overline{v}_{BI}, \overline{v}_B)}$**:**

$$(vw - \theta L_B) \geq \sqrt{(1-\lambda)d_F \frac{4c(1-w)^2 v^3 \lambda w}{k_B}\left[1 - \frac{1}{2k_F}\frac{(1-\lambda)d_F \Delta k}{\lambda}\right] - \frac{4cvw}{\lambda}(1-\lambda)d_F\theta(\tilde{L} - L_F) + 4cvw[vw(1-c) - \theta\tilde{L}] + 4cvw(1-\lambda)(d_F+1)\beta} \quad \text{(C5')}$$

Case 6. $\boldsymbol{v > \overline{v}_B}$**:**

$$(1-\lambda)d_F \frac{(1-w)^2 v^2 \lambda}{k_B}\left[1 - \frac{1}{2k_F}\frac{(1-\lambda)d_F \Delta k}{\lambda}\right] + (1-\lambda)\lambda(d_F+1)\beta \leq 0 \quad \text{(C6')}$$

If $\frac{\Delta k}{k_F} \geq \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$, then condition (C0') always holds, and the BE strategy is optimal in Case 0. Note that the LHS of (C6') monotonically decreases in $\frac{\Delta k}{k_F}$. If we plug in $\frac{\Delta k}{k_F} = \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$, we obtain that the LHS of (C6') is zero. Hence, for $\frac{\Delta k}{k_F} \geq \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$, condition (C6') always holds, meaning that the BE strategy is optimal in Case 6. Under these two cases (Case 0 and Case 6), banks can enjoy the greatest advantage of the BI strategy over the BE strategy for the following reasons: (1) Banks do not need to sacrifice their ideal settlement speed ($q_1^{BI} = q_1^{BE}$). (2) Banks obtain additional benefit from the innovative features ($q_2^{BI} > q_2^{BE}$). (3) Fintech firms' reservation values are constrained by the settlement speed ($q_1^{FE} = 0$ or 1). If banks cannot induce fintech firms' participation in the unified system under Cases 0 and 6, then it is impossible to induce their participation under Cases 1-5 because the advantage of the BI strategy over the BE strategy is further reduced in those cases—banks need to either sacrifice the settlement speed (i.e., setting the speed faster than banks' ideal level, $q_1^{BI} > q_1^{BE}$) for Cases 2-5 or reducing the rent when fintech firms' reservation values are unconstrained by the settlement speed ($0 < q_1^{FE} < 1$) for Cases 1-3. Therefore, for Cases 1-5, banks will certainly find BE to be better. Because fintech firms have a significant technological advantage in building their own systems (i.e., a large $\frac{\Delta k}{k_F}$), their outside option value is too high to be of banks' interest to induce fintech firms' participation. Hence, we conclude that when $\frac{\Delta k}{k_F} \geq \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$, banks' optimal strategy in all cases is BE.

If $\frac{\Delta k}{k_F} < \frac{2\lambda}{(1-\lambda)d_F} + \frac{2\lambda(d_F+1)k_B\beta}{v^2(1-w)^2(1-\lambda){d_F}^2}$, we note that the bank's optimal strategy in Case 0 is BI. For Cases 1-6, we conduct the analysis in exactly the same way as in Proposition 2. We also reach results similar to those in Proposition 2. The bank association might adopt the BE strategy in a middle subrange of v. We denote this BE-optimality region as $(\underline{v}_{e2}, \overline{v}_{e2})$, where $\underline{v}_{e2}$ and $\overline{v}_{e2}$ might take different values depending on the concrete parameter values. Therefore, as v increases, the optimal strategy follows the pattern of "inclusive—exclusive—inclusive." Following the same logic as in the proof of Proposition 2, we show that the BE-optimality region could be empty; if it appears, it will only appear once. Moreover, the FE-optimality region identified in Proposition 2 is a subset of the BE-optimality region, that is, $(\underline{v}_{e1}, \overline{v}_{e1}) \subset (\underline{v}_{e2}, \overline{v}_{e2})$. This can be easily seen by comparing the conditions (C1) - (C6) with (C1') - (C6')—whenever the former condition holds, the latter holds as well. Finally, the same three variables, $\frac{\Delta k}{k_F}$, $\frac{(1-\lambda)d_F}{\lambda}$, and β, appear in all of the conditions. When their values increase, the RHS of (C1') - (C5') and the LHS of (C6') decrease, thus making these conditions more likely to hold, and the BE-optimality region is more likely to appear.

Proof of Proposition 4

The results of Proposition 4 directly follow from the proofs of Propositions 2 and 3.

Proof of Proposition 5

Under the GM strategy, the government's optimization problem is characterized by Equations (9) - (11). Taking the first-order conditions with respect to q_1 and q_2 and solving the system of equations, we obtain $q_1 = \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right)$, $q_2 = \frac{v(1-w)\Gamma}{k_F}$. Note that the unconstrained system design solution is the same as that in the FI strategy. Applying the condition $0 \leq q_1 \leq 1$, we obtain two critical values, $\underline{v}_G = \frac{\tilde{L}\theta}{w}$ and $\overline{v}_G = \frac{\tilde{L}\theta}{(1-2c)w}$. Note that $\underline{v}_G = \underline{v}_{FI}$ and $\overline{v}_G = \overline{v}_{FI}$. Consequently, we have the same regions (BR, ER, and RR) and the same settlement speeds in each region as those under the FI strategy.

Next, we compute the Shapley value for banks and fintech firms. First, banks obtain the value of $V_B = \lambda\{v[w(q_1 - cq_1^2) + (1-w)q_2] + \beta - \theta q_1 L_B\}$ from using the government-led system. There are two cases. In the first case, banks participate first, and fintech firms participate next. When banks build the new system and use it alone, they create surplus π_B^r, which is their marginal contribution. In the second case, fintech firms participate first, and banks come next. If fintech firms build the new system alone, they obtain surplus π_F^r. When banks join, the total gain increases to SW^G. In this case, banks' marginal contribution is $SW^G - \pi_F^r$. Therefore, banks' marginal value of participation is the average of their marginal contributions under the two cases, which is $SP_B = \frac{\pi_B^r + SW^G - \pi_F^r}{2}$. Following the same logic and approach, we can compute $SP_F = \frac{\pi_F^r + SW^G - \pi_B^r}{2}$.

Finally, we compute the cost sharing of banks and fintech firms, C_B and C_F, under the Shapley-value-based sharing rule. Note that $C_B = V_B - SP_B = V_B - \frac{SW^G - \pi_F^r + \pi_B^r}{2} = \frac{V_B - V_F}{2} + \frac{\pi_F^r - \pi_B^r}{2} + \frac{k_F}{4}q_2^2$. We have the following seven cases:

(1) If $\leq \underline{v}_F$, then the settlement speeds q_1 in the government-led FE and BE systems are 0. Substituting into the optimal value $q_2 = \frac{v(1-w)\Gamma}{k_F}$, we have $V_B = \frac{\lambda v^2(1-w)^2\Gamma}{k_F} + \lambda\beta$, $V_F = \frac{(1-\lambda)d_F v^2(1-w)^2\Gamma}{k_F} + (1-\lambda)d_F\beta$. In addition, $\pi_F^r = \frac{(1-\lambda)^2 d_F^2 v^2(1-w)^2}{2k_F} + (1-\lambda)^2 d_F\beta$ and $\pi_B^r = \frac{\lambda^2 v^2(1-w)^2}{2k_B} + \beta\lambda^2$. Substituting into all terms and simplifying, we obtain $C_B^{1*} = \frac{v^2(1-w)^2\lambda}{4}\left[\frac{3\lambda + 2(1-\lambda)d_F}{k_F} - \frac{\lambda}{k_B}\right] + \frac{1}{2}\lambda(1-\lambda)(1-d_F)\beta$.

(2) If $\underline{v}_F \leq v \leq \underline{v}_{FI}$, the only change compared with Case (1) is that the settlement speed in the FE system becomes greater than 0; hence, the expression of π_F^r has an additional term $\frac{(1-\lambda)d_F(vw - \theta L_F)^2}{4cvw}$. Consequently, $C_B^{2*} = C_B^{1*} + \frac{(1-\lambda)d_F(vw-\theta L_F)^2}{8cvw}$.

(3) If $\underline{v}_{FI} \leq v \leq \underline{v}_B$, the only change compared with Case (2) is that the settlement speed in the government-led system becomes greater than 0; that is, $q_1^G = \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right)$. Substituting into q_1^G and recalculating $\frac{V_B - V_F}{2}$ yields a new term, $\frac{(vw-\theta\tilde{L})}{8cvw}\{[\lambda - (1-\lambda)d_F](vw + \theta\tilde{L}) - 2\theta[\lambda L_B - (1-\lambda)d_F L_F]\}$. Hence, $C_B^{3*} = C_B^{2*} + \frac{(vw-\theta\tilde{L})}{8cvw}\{[\lambda - (1-\lambda)d_F](vw+\theta\tilde{L}) - 2\theta[\lambda L_B - (1-\lambda)d_F L_F]\}$.

(4) If $\underline{v}_B \leq v \leq \overline{v}_F$, the only change compared with Case (3) is that the settlement speed in the BE system becomes greater than 0; thus, π_B^r contains one additional term $\frac{\lambda(vw-\theta L_B)^2}{4cvw}$. Consequently, $C_B^{4*} = C_B^{3*} - \frac{\lambda(vw-\theta L_B)^2}{8cvw}$.

(5) If $\overline{v}_F \leq v \leq \overline{v}_{FI}$, the only change compared with Case (4) is that the settlement speed in the FE system becomes 1; therefore, the term $\frac{(1-\lambda)d_F(vw-\theta L_F)^2}{4cvw}$ in π_F^r is replaced by $(1-\lambda)d_F[vw(1-c) - \theta L_F]$. Consequently, $C_B^{5*} = C_B^{4*} - \frac{(1-\lambda)d_F(vw-\theta L_F)^2}{8cvw} + \frac{(1-\lambda)d_F[vw(1-c)-\theta L_F]}{2}$.

(6) If $\overline{v}_{FI} \leq v \leq \overline{v}_B$, the only change compared with Case (5) is that the settlement speed in the government-led system becomes 1. Substituting into $q_1^G = 1$ and recalculating $\frac{V_B - V_F}{2}$, the term $\frac{(vw-\theta\tilde{L})}{8cvw}\{[\lambda - (1-\lambda)d_F](vw+\theta\tilde{L}) - 2\theta[\lambda L_B - (1-\lambda)d_F L_F]\}$ is replaced by $\frac{vw[\lambda-(1-\lambda)d_F](1-c) - \theta[\lambda L_B - (1-\lambda)d_F L_F]}{2}$. Further combining terms and simplifying, we have $C_B^{6*} = C_B^{5*} + \frac{[\lambda-(1-\lambda)d_F]}{2}\left[vw(1-c) - \frac{v^2w^2 - \theta^2\tilde{L}^2}{4cvw}\right] + \frac{\theta[\lambda L_B - (1-\lambda)d_F L_F]}{2}\left(\frac{vw - \theta\tilde{L}}{2cvw} - 1\right)$.

(7) If $v > \overline{v}_B$, the only change compared with Case (6) is that the settlement speed in the BE system becomes 1; therefore, the term $\frac{\lambda(vw-\theta L_B)^2}{4cvw}$ in π_B^r is replaced by $\lambda[vw(1-c) - \theta L_B]$. Consequently, $C_B^{7*} = C_B^{6*} + \frac{\lambda(vw-\theta L_B)^2}{8cvw} - \frac{\lambda[vw(1-c)-\theta L_B]}{2}$.

General Cost-Sharing Rule in Government-Led System

To derive the valid value range under a general cost-sharing rule, we check both IC conditions (10) and (11) in each region.

(1) When $v \leq \underline{v}_G$, substituting $q_1 = 0$ and $q_2 = \frac{v(1-w)\Gamma}{k_F}$ into conditions (10) and (11), we obtain $\frac{v^2(1-w)^2\Gamma}{k_F}\left(\frac{(1-\lambda)d_F - \lambda}{2}\right) \leq C_B \leq \frac{v^2(1-w)^2\Gamma}{k_F}\lambda$.

(2) When $\underline{v}_G < v \leq \overline{v}_G$, substituting the interior solution $q_1 = \frac{1}{2c}\left(1 - \frac{\tilde{L}\theta}{vw}\right)$ and $q_2 = \frac{v(1-w)\Gamma}{k_F}$ into conditions (10) and (11) and simplifying, we obtain $\frac{v^2(1-w)^2\Gamma}{k_F}\left(\frac{(1-\lambda)d_F - \lambda}{2}\right) - \frac{(1-\lambda)d_F}{4c}\left(1 - \frac{\tilde{L}\theta}{vw}\right)(vw + \tilde{L}\theta - 2\theta L_F) \leq C_B \leq \frac{v^2(1-w)^2\Gamma}{k_F}\lambda + \frac{\lambda}{4c}\left(1 - \frac{\tilde{L}\theta}{vw}\right)(vw + \tilde{L}\theta - 2\theta L_B)$.

(3) When $v > \overline{v}_G$, substituting $q_1 = 1$ and $q_2 = \frac{v(1-w)\Gamma}{k_F}$ into conditions (10) and (11), we obtain $\frac{v^2(1-w)^2\Gamma}{k_F}\left(\frac{(1-\lambda)d_F-\lambda}{2}\right) - (1-\lambda)d_F[vw(1-c)-\theta L_F] \leq C_B \leq \frac{v^2(1-w)^2\Gamma}{k_F}\lambda + \lambda[vw(1-c)-\theta L_B]$).

Hence, we obtain the following cost range that the government could adopt:

$$C_B = \begin{cases} \left[\dfrac{v^2(1-w)^2\Gamma\big((1-\lambda)d_F-\lambda\big)}{2k_F}, \dfrac{v^2(1-w)^2\Gamma\lambda}{k_F}\right] & if\ v \leq \underline{v}_G \\ \left[\begin{array}{c}\dfrac{v^2(1-w)^2\Gamma\big((1-\lambda)d_F-\lambda\big)}{2k_F} - \dfrac{(1-\lambda)d_F}{4c}\left(1-\dfrac{\tilde{L}\theta}{vw}\right)\left(vw+\tilde{L}\theta-2\theta L_F\right), \\ \dfrac{v^2(1-w)^2\Gamma\lambda}{k_F} + \dfrac{\lambda}{4c}\left(1-\dfrac{\tilde{L}\theta}{vw}\right)\left(vw+\tilde{L}\theta-2\theta L_B\right)\end{array}\right] & if\ \underline{v}_G < v < \overline{v}_G \\ \left[\begin{array}{c}\dfrac{v^2(1-w)^2\Gamma\big((1-\lambda)d_F-\lambda\big)}{2k_F} - (1-\lambda)d_F(vw(1-c)-\theta L_F), \\ \dfrac{v^2(1-w)^2\Gamma\lambda}{k_F} + \lambda(vw(1-c)-\theta L_B)\end{array}\right] & if\ v \geq \overline{v}_G \end{cases}$$

and $C_F = \frac{[v(1-w)\Gamma]^2}{2k_F} - C_B$ in each region.

Note that for these ranges to be nonempty, the sufficient condition is $\frac{(1-\lambda)d_F-\lambda}{2} \leq \lambda$, which is equivalent to $(1-\lambda)d_F \leq 3\lambda$. The condition states that the total transaction volume from fintech firms should be no more than triple the total transaction volume from banks. This is generally true in the current financial payments industry, and the condition will continue to hold in the foreseeable future.

Proof of Proposition 6

(*i*) When $v \in (\underline{v}_{e1}, \overline{v}_{e1}]$, Proposition 4 shows that the unique equilibrium is the fragmented system. The proof of Proposition 3 shows that $(\underline{v}_{e1}, \overline{v}_{e1}) \subset (\underline{v}_{e2}, \overline{v}_{e2})$. From Propositions 2 and 3, we know that the optimal strategy for fintech firms is FE and for banks is BE. Therefore, in this region, $\pi_F^{FE} = \pi_F^r > \pi_F^{FI}$ and $\pi_B^{BE} = \pi_B^r > \pi_B^{BI}$. Total social welfare under the fragmented systems is $SW^{FE+BE} = \pi_B^r + \pi_F^r$. Total social welfare under a unified system is $SW^{FI} = \pi_F^{FI} + \pi_B^{FI} = \pi_F^{FI} + \pi_B^r < \pi_F^r + \pi_B^r = SW^{FE+BE}$ in a fintech-led unified system and $SW^{BI} = \pi_B^{BI} + \pi_F^{BI} = \pi_B^{BI} + \pi_F^r < \pi_B^r + \pi_F^r = SW^{FE+BE}$ in a bank-led unified system. In addition, we have shown that the government-led system is the same as the fintech-led system. Therefore, $SW^G = SW^{FI} < SW^{FE+BE}$ and government coordination will not improve social welfare in this region.

(ii) When $v \in (\underline{v}_{e2}, \underline{v}_{e1}]$ or $v \in (\overline{v}_{e1}, \overline{v}_{e2}]$, Proposition 4 shows that the unique equilibrium is the fintech-led unified system. Propositions 2 and 3 further show that the optimal strategy for fintech firms is FI. Therefore, in this region, $\pi_F^{FE} = \pi_F^r < \pi_F^{FI}$ and $\pi_B^{FI} = \pi_B^r$. Hence, $SW^G = SW^{FI} = \pi_F^{FI} + \pi_B^{FI} > \pi_F^r + \pi_B^r = SW^{FE+BE}$. In this region, although the government-led system results in the same total social welfare as the fintech-led system, it can achieve fair cost and benefit sharing, as shown in the proof of Proposition 5.

(iii) When $v \in (0, \underline{v}_{e2}]$ or $v > \overline{v}_{e2}$, Proposition 4 shows that both the fintech-led and bank-led unified system can be the equilibrium outcome. From Propositions 2 and 3, we know that the optimal strategy for fintech firms is FI and for the banks is BI. In the proofs of Lemmas 3 and 4, we have derived total social welfare in different parameter ranges. Comparing social welfare under FI and BI in all the parameter ranges, we have $SW^G = SW^{FI} > SW^{BI}$. Therefore, to prevent the appearance of the inferior equilibrium outcome BI, government coordination is needed.