Springer Series in Supply Chain Management

Volodymyr Babich
John R. Birge
Gilles Hilary *Editors*

Innovative Technology at the Interface of Finance and Operations

Volume II

Springer

Springer Series in Supply Chain Management

Volume 13

Series Editor
Christopher S. Tang
University of California
Los Angeles, CA, USA

Supply Chain Management (SCM), long an integral part of Operations Management, focuses on all elements of creating a product or service, and delivering that product or service, at the optimal cost and within an optimal timeframe. It spans the movement and storage of raw materials, work-in-process inventory, and finished goods from point of origin to point of consumption. To facilitate physical flows in a time-efficient and cost-effective manner, the scope of SCM includes technology-enabled information flows and financial flows.

The Springer Series in Supply Chain Management, under the guidance of founding Series Editor Christopher S. Tang, covers research of either theoretical or empirical nature, in both authored and edited volumes from leading scholars and practitioners in the field – with a specific focus on topics within the scope of SCM.

Springer and the Series Editor welcome book ideas from authors. Potential authors who wish to submit a book proposal should contact Ms. Jialin Yan, Associate Editor, Springer (Germany), e-mail: jialin.yan@springernature.com

Volodymyr Babich • John R. Birge • Gilles Hilary
Editors

Innovative Technology at the Interface of Finance and Operations

Volume II

Springer

Editors
Volodymyr Babich
Washington, DC, USA

John R. Birge
Chicago, IL, USA

Gilles Hilary
Washington, DC, USA

ISSN 2365-6395 ISSN 2365-6409 (electronic)
Springer Series in Supply Chain Management
ISBN 978-3-030-81944-6 ISBN 978-3-030-81945-3 (eBook)
https://doi.org/10.1007/978-3-030-81945-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Blockchain Intra- and Interoperability

Alexander Lipton and Thomas Hardjono

1 Introduction

The publication of the seminal white paper by Satoshi Nakamoto in October of 2008 opened a new era of technological innovation—the era of distributed ledgers; see Nakamoto (2008). Nakamoto's paper introduced the public blockchain concept and explained how one could build such a blockchain in practice. It started a revolution, which shows no signs of abating.

At the basic level, blockchain is a shared, distributed ledger capable of supporting ownership of assets, tracking transactions between assets' owners, and changing ownership accordingly; see, e.g., Gupta (2017). In theory, blockchains can support tangible assets, such as money, shares, real estate, and intangible assets, such as intellectual property. Eventually, blockchain networks can support ownership of anything of value. Recording ownership of a blockchain reduces risks by increasing business interoperability, cuts costs for all involved, and eliminates intermediaries. By using blockchains, one could radically reorganize business *modus operandi*.

One can think about three complementary ways of organizing information: (a) centralized, (b) decentralized, (c) distributed; see Fig. 1.

Currently, the centralized hub and spoke model is standard for most industries, for example, banking in a single country. In this example, the hub is the central bank, and the spokes are individual commercial banks. A decentralized system with several hubs and spokes is also quite common. A typical example is cross-border banking. The hubs are central banks, and the spokes are individual commercial banks in their respective countries. Finally, a distributed system relies on direct or

A. Lipton (✉) · T. Hardjono
MIT Connection Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: alexlipt@mit.edu; hardjono@mit.edu

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_1

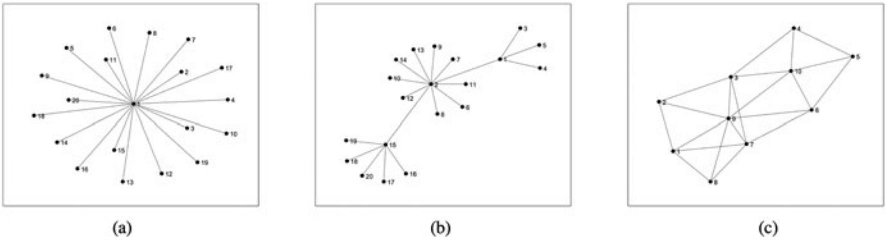

Fig. 1 Three possible ways of organizing networks. (**a**) Centralized. (**b**) Decentralized. (**c**) Distributed

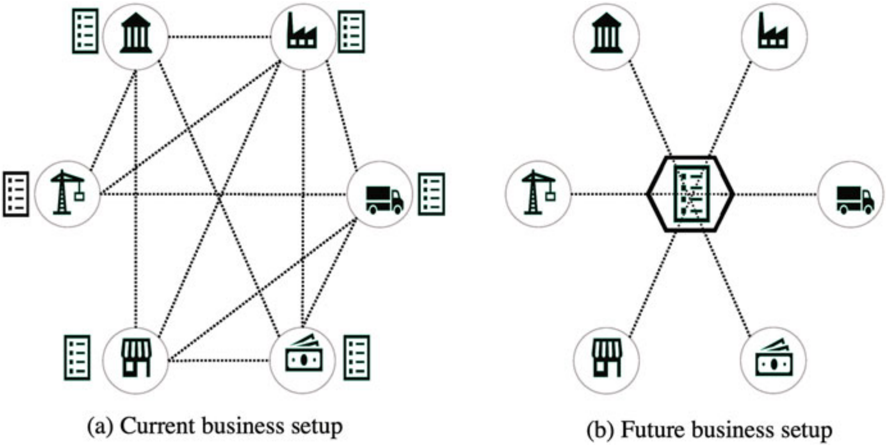

Fig. 2 (**a**) Current arrangement—Each participant maintains her ledger; participants periodically reconcile ledgers against each other. (**b**) Future arrangement—Each participant contributes to the shared distributed ledger; a suitable consensus algorithm maintains the ledger's integrity. The information has to be suitably encrypted to satisfy business requirements and data privacy laws

peer-to-peer (P2P) business organization (currently a rare instance) and potentially can be viewed as the most robust of the three. The great hope for the distributed ledger technology (DLT) is that it can serve as the backbone of distributed business models.

Using the recent developments in DLT, we can reorganize business activities in general. In Fig. 2, we compare two possibilities: (a) the current system, with each participant holding her own ledger, reconciled against other ledgers periodically, (b) a future system relies on all parties, maintaining a shared ledger as a group.

The current system enjoys several advantages, including each participant's ability to rely on her trusted systems to control the information released to other participants. However, such an approach results in redundancies, sporadical errors, and potentially fraudulent activities by its very nature. In the future system, each participant can write transactions on the ledger and access the relevant information

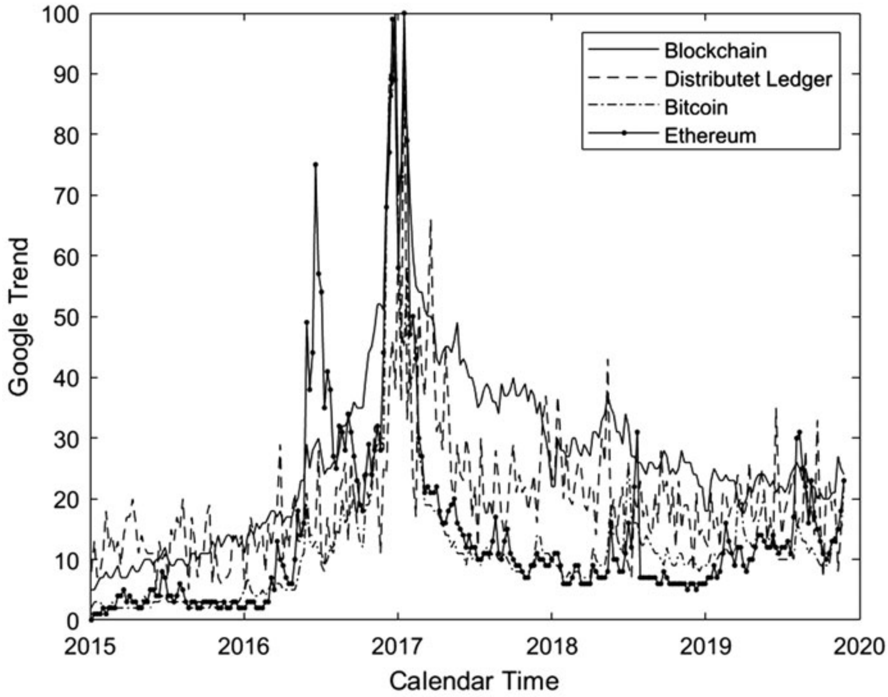

Fig. 3 Google trends reflecting the worldwide interest in blockchains, distributed ledgers, Bitcoin, and Ethereum. Source: https://trends.google.com/

she is entitled to know. Such a system reduces the business process's overall frictions and increases its robustness because of built-in (rather than accidental) redundancies. Clearly, streamlining the business process using DLT is not free because it requires maintaining consensus on the shared ledger and properly obfuscating private data.

Figure 3 convincingly illustrates the fact that Google trends showing interest in DLT, undeniably kindled by the meteoric price appreciation of Bitcoin and other cryptocurrencies, did not go away (although naturally diminished) after Bitcoin's bubble burst. It suggests that DLT has many more usages in addition to its applications to cryptocurrencies.

There is currently tremendous interest in using blockchain and smart contracts technology to re-implement many of the existing functions within the financial sector (e.g. automated market maker) in a decentralized fashion. Numerous "decentralized finance" (DeFi) projects or offerings have been reported in the media in the past year. A core proposition in decentralized finance is that anyone can be a market-maker, and that the replicated copies of smart contracts ensures the "decentralization" of functions.

However, we believe that decentralization should occur across distinct blockchain systems in a manner that provides the application developer with a single view regardless of how many distinct blockchain systems are involved. That is, true decentralization means that functions are spread across distinct blockchain systems, where nodes may be implemented using different software stacks and where each blockchain system may employ different consensus mechanisms and different ledger data structures (Hardjono et al., 2019). Indeed, this is how the TCP/IP Internet is architected and this is largely why the Internet has been able to grow in size and in traffic capacity to serve end-users at a global scale. The Internet is not a single contiguous IP network. It is in fact a collection of interconnected *Autonomous Systems* (AS) (Clark, 1988), where each AS has a well-defined physical boundary and each AS is operated by an ISP. In contrast, in the case of the recent DeFi offerings, most (if not all) of the DeFi efforts have been conducted on one blockchain platform, namely Ethereum.

Although much of DeFi activities are today on the Ethereum platform, the history of the Internet indicates that it is unlikely that there will be a single "world computer". Private (permissioned) blockchains and DLT systems today are used in closed communities (e.g. see Whittemore, 2020), and maintaining separate blockchains may be part of the business survivability of many of these communities. Other platforms are being developed (e.g. Dfinity Martin, 2020; Heaven, 2020), and geopolitical factors may prevent the market dominance by a single blockchain platform (Huang, 2020).

If blockchain technology is to become a foundation for the future decentralized finance then the issue of interoperability must be addressed. Standard technical protocols for the various aspects of interoperability must be created, tested and deployed—just as numerous standard protocols have been developed over the past three decades for the TCP/IP Internet.

The purpose of this paper is to introduce the concepts of blockchain intra- and interoperability and describe their potential practical applications. It is organized as follows. Section 2 describes several essential examples, including Bitcoin, and Ethereum. Section 3 introduces the concept of blockchain intraoperability and shows how to achieve intraoperability in an Ethereum-like blockchain by using automated market makers. Section 4 defines the concept of blockchain interoperability and develops several viable approaches to designing the corresponding mechanisms, including gateway asset transfer protocols. Conclusions are drawn in Sect. 5.

A recent paper by Hardjono et al. and a book by Lipton and Treccani cover this chapter's material in much greater detail and contains additional references; see Hardjono et al. (2019) and Lipton and Treccani (2021).

2 Blockchains in a Nutshell

2.1 Bitcoin

Nakamoto described her intentions as follows: "I've been working on a new electronic cash system that's fully peer-to-peer, with no trusted third party—The main properties: Double-spending is prevented with a peer-to-peer network. No mint or other trusted parties. Participants can be anonymous. New coins are made from Hashcash style proof-of-work. The proof-of-work for new coin generation also powers the network to prevent double-spending." The result of these efforts is the celebrated Bitcoin protocol for moving the corresponding token known as BTC from one of the protocol participants to the next. Nakamoto's designed blockchain relies on public-key or asymmetric cryptography, specifically elliptic-curve cryptography (ECC). It uses pairs of keys: a public key, known to all, and a secret key, known only to the owner. Public keys are the protocol participants' addresses, while secret keys are tools for unlocking BTCs held at the corresponding addresses.

All BTC transactions are cryptographically secured via the elliptic curve digital signature algorithm (ECDSA) and do not require further efforts to ensure that they are valid. However, in the absence of strict controls, nothing prevents the owner of a particular address from spending her money BTCs twice—the celebrated double-spend problem. To ensure that the Bitcoin protocol is self-consistent and prevents double-spending, it is necessary to have a distinctive class of participants in the Bitcoin protocol, called miners. Miners listen to the network for upcoming transactions, assemble these transactions into blocks, and participate in a competition to have their block added to the system. Other miners will accept a new block only if it contains valid transactions and is correctly stamped by the successful miner. To stamp a block, miners need to solve a computational hash puzzle, requiring spending electricity and other resources on a prodigious scale.[1] Thus, the name—a proof-of-work (PoW) consensus algorithm.

2.2 Ethereum

Ethereum addresses the limitations of Bitcoin scripts by introducing smart contracts (SCs), which are stateful and Turing-complete scripts, at least in theory. Ethereum considerably augments the capabilities of DLTs and their possible applications. While the Bitcoin protocol focuses on a single use case of consistently mining and moving BTC, which it views as electronic cash, the Ethereum protocol offers a decentralized, trusted computing platform for executing arbitrary code. Its native

[1] Last year (2019), the Bitcoin protocol confirmed 100 million transactions. For doing so, it used more electricity than Denmark.

cryptocurrency, ether (ETH), while undeniably very important, merely acts as the token financing the execution of SCs by thousands of machines. Since SCs regulate all kind of separate use cases, the best way of thinking about Ethereum is as consensus as a service (CaaS) provider for any application that requires a high level of trust and auditability.[2]

Ethereum was never shy about its grandiose ambitions. The Ethereum white paper states its purpose as follows: "Satoshi Nakamoto's development of Bitcoin in 2009 has often been hailed as a radical development in money and currency, being the first example of a digital asset which simultaneously has no backing or intrinsic value and no centralized issuer or controller. However, another—arguably more important - part of the Bitcoin experiment is the underlying blockchain technology as a tool of distributed consensus, and attention is rapidly starting to shift to this other aspect of Bitcoin."

The Bitcoin and Ethereum protocols have many superficial similarities since both are public distributed ledgers relying on ECDSA and PoW.[3] However, there are profound differences, too. Bitcoin is a protocol squarely aimed at supporting BTC as an alternative currency. Ethereum is a protocol designed for running a distributed computing platform capable of storing and executing arbitrary code on the Ethereum Virtual Machine (EVM). In addition to externally owned accounts (EOA), which operate just like Bitcoin accounts, Ethereum supports SCs, maintaining a ledger for external tokens and defining applications unrelated to financial instruments. For example, one can trivially implement an equivalent to BTC as a token contract on the Ethereum platform.

2.3 Other Blockchains

The Bitcoin and Ethereum protocols launched the blockchain revolution. By now, there are thousands of blockchain utilizing a wide variety of consensus mechanisms, from the classical PoW, to the proof-of-stake (PoS), to the practical Byzantine- fault-tolerant consensus. Some of these blockchains are public and open to all and sundry to join; others are private and can be joint only by preapproved entities.

What is important to us is that many of the blockchains can support smart contracts so that they are CaaS providers carrying all kinds of digital assets representing both tangible and intangible real-world assets. Accordingly, two exciting and practically relevant problems arise: (a) how to exchange assets defined on the same blockchain; (b) how to move assets from one blockchain to the next. We explicitly formulate and discuss these problems in detail in Sects. 3 and 4.

---

[2]Consensus as a service is an end-to-end process ensuring the overall consistency of operations run on distributed ledgers.

[3]Ethereum 2.0 is going to switch to a proof of stake (PoS) consensus algorithm.

3 Challenges in Blockchain Intraoperability

We define blockchain intraoperability as exchanging different assets defined on the same blockchain supporting smart contracts. Swapping of stable coins, described below, is particularly important. Exchanging Ethereum-based stable coins, such as True USD (TUSD) and USD Coin (USDC), is a typical example of intraoperability in action. Exchanging Tether (USDT), a Bitcoin-based stable coin, for USD Coin (USDC), an Ethereum-based stable coin, is an example of blockchain interoperability. Interoperability forms the backbone of the burgeoning field of decentralized finance (DeFi). We implement blockchain intraoperability by using Automatic Market Makers (AMMs).

For illustrative purposes, we show prices of USDT and TUSD in Fig. 4.

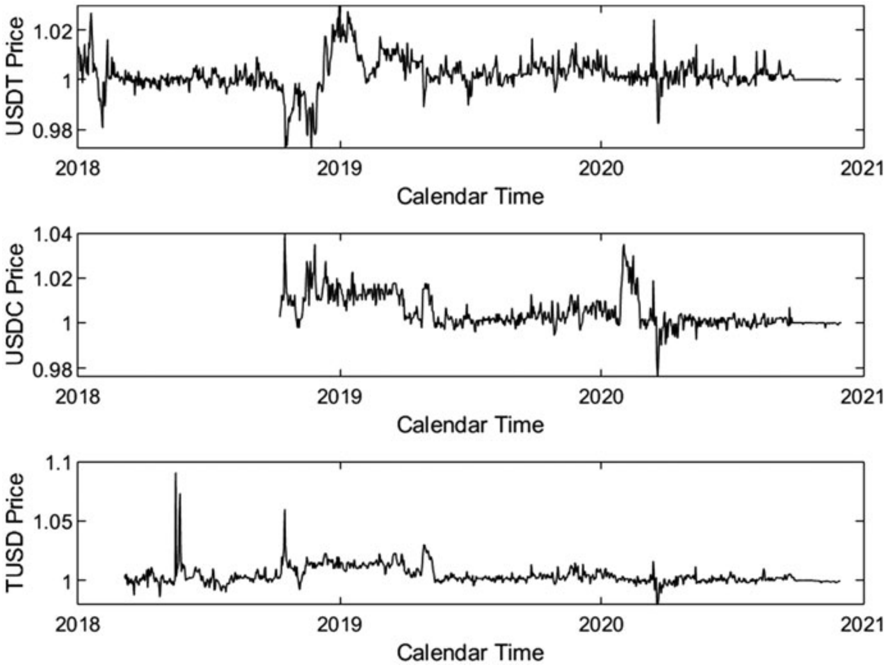

Fig. 4 Prices of USDT, USDC, and TUSD since inception. Their market capitalizations are 19, 3, and 0.3 billion USD, respectively. Source: https://coinmarketcap.com/

3.1 Stablecoins

Despite occasional claims to the contrary, conventional cryptocurrencies, such as BTC and ETH, are ill-suited to commerce's needs. There are several reasons why it is the case, the most apparent being that cryptocurrencies have colossal volatility. Hence, considerable efforts are directed towards building the so-called stablecoins, which live on blockchains, but have low volatility. The ultimate culmination of these efforts would be introducing central bank digital currencies (CBDCs), which are digital representations of the corresponding fiat.

The simplest way of building a stablecoin is to use the Ethereum protocol as a CaaS provider, designing such a coin as a token pegged to an asset or a basket of assets, viewed as stable in a conventional economic sense.

Because of its technical capabilities, the Ethereum protocol is often viewed as a natural tool for building stablecoins. In contrast, the Bitcoin protocol is considered ill-suited for anything other than recording BTC transactions. In reality, the biggest by far of all stablecoins is Tether, a Bitcoin Omni Layer token. Several other blockchains, such as Stellar or Algorand, can be used to support stablecoins. In many respects, they are better than Ethereum, simply because using Ethereum might be prohibitively expensive.

The development of robust and trustworthy stablecoins is of paramount importance for the future real-life applications of DLT. Here we mention that there are four viable types of stablecoins:

- fully collateralized by individual fiat currencies;
- fully collateralized by baskets of fiat currencies;
- overcollateralized with native token;
- asset-backed.

Other possibilities considered in the literature are partially collateralized and algorithmically stabilized stable coins. Such coins are not viable and are not worth discussing.

Although the prices of stable coins, even fully fiat-backed, deviate from their equilibrium values due to the varying supply-demand and other considerations, these deviations are kept relatively small by arbitrageurs.

3.2 Automated Market Makers

Let us design an SC, capable of making markets between two tokens TN_1, TN_2, whose relative price, i.e., the price of the second token expressed in terms of the first one, is denoted by P. We call such a contract an automated market maker (AMM). AMMs gained popularity over the last couple of years. Initially, anyone can become a market maker by delivering TN_1 and TN_2 simultaneously in the right proportion to the collateral pool. Subsequently, anyone can remove one of the tokens from the pool by simultaneously providing the other token to the pool according to the rule defining the underlying SC. The best use case for AMMs is swapping stablecoins. However, exchanging other tokens against each other, for instance, a stablecoin against ETH, is also possible.

The actual exchange rate relies on rules, which have to be agreed upon in advance. We consider three possible choices: the constant sum, constant product, and mixture rules. Several sources cover AMMs; see, e.g., Angeris et al. (2019), Egorov (2019), Lipton and Treccani (2021), Schär (2020), and Zhang et al. (2018).

We assume that the initial prices of TN_1, TN_2 tokens are equal to each other and consider an automated market maker defined by the constant sum rule:

$$X + Y = \Sigma_0, \quad X_0 = Y_0 = N, \quad \Sigma_0 = 2N. \tag{1}$$

Here X, Y are the numbers of TN_1, TN_2 in the pool. Equation (1) yields:

$$Y = \Sigma_0 - X, \quad \left|\frac{dY}{dX}\right| = 1. \tag{2}$$

According to Eq. (2), the pool becomes exhausted when $X = \Sigma_0$, since increasing X from N to $2N$ is a rational thing for an arbitrageur to do when TN_2 becomes more expensive than TN_1. The marginal price of TN_2 expressed in terms of TN_1, is given by the second equation (2). This price is constant and equal to one. A constant sum AMM makes sense when TN_1, TN_2 are stablecoins, with prices weakly fluctuating around their equilibrium values. If transaction fees are zero, it is rational to exhaust the pool even if the deviation from equilibrium is minuscule. However, in the more realistic situation when transaction fees are nonzero, the corresponding deviation has to be above a threshold to make arbitrage profitable.

The constant product rule defines more interesting (and practically important) AMMs:

$$XY = \Pi_0, \quad X_0 = Y_0 = N, \quad \Pi_0 = N^2. \tag{3}$$

It is clear that

$$Y = \frac{\Pi_0}{X}, \quad \left|\frac{dY}{dX}\right| = \frac{\Pi_0}{X^2}. \tag{4}$$

Thus, an arbitrageur cannot exhaust such a pool, so that it shall exist indefinitely. It is clear that the price of TN_2 expressed in terms of TN_1 is no longer constant and increases (decreases) when X decreases (increases).

If necessary, we can generalize the constant sum, and constant product rules. The constant sum and constant product rules, given by Eqs. (1) and (3), can be written as follows:

$$\left(\frac{\Sigma}{\Sigma_0}-1\right)=0, \quad X_0=Y_0=N, \quad \Sigma_0=2N,$$
$$\left(\frac{\Pi_0}{\Pi}-1\right)=0, \quad X_0=Y_0=N, \quad \Pi_0=N^2. \tag{5}$$

where $\Sigma = X + Y$, $\Pi = XY$ are the current sum and product, respectively. These rules can be combines as follows:

$$\left(\frac{\Pi_0}{\Pi}-1\right)-\alpha\left(\frac{\Sigma}{\Sigma_0}-1\right)=0,$$
$$X_0=Y_0=N, \quad \Sigma_0=2N, \quad \Pi_0=N^2. \tag{6}$$

where $\alpha > 0$ is an adaptive parameter that characterizes the transition from the constant product to the constant sum rule. The product Π appears in the denominator to avoid the possibility of exhausting the entire pool and ensuring that:

$$Y(X) \underset{X\to 0}{\to} \infty, \qquad X(Y) \underset{Y\to 0}{\to} \infty. \tag{7}$$

Of course, by opening herself to arbitrageurs' actions, an AMM will suffer a loss caused by a drop in the collateral value below its buy-and-hold level. In the language of mathematical finance, an AMM, who can be viewed as an option seller, suffers from negative convexity. To compensate for this loss, AMMs have to charge transaction fees, see the next section. The AMM's loss is called impermanent because, under the mean-revertion assumption, it tends to disappear. However, the mean-reversion assumption might or might not hold in real life. Introducing x, y, such that $X = Nx$, $Y = Ny$, we rewrite Eqs. (1)–(2) as follows:

$$x+y=2, \quad x_0=y_0=1, \tag{8}$$

$$y(x)=2-x, \quad \left|\frac{dy}{dx}\right|=1. \tag{9}$$

In terms of x, y, the constant product rule given by Eqs. (3)–(4) can be written as follows:

$$xy=1, \quad x_0=y_0=1, \tag{10}$$

$$y(x)=\frac{1}{x}, \quad \left|\frac{dy}{dx}\right|=\frac{1}{x^2}. \tag{11}$$

Finally, Eqs. (6) written in terms of x, y become:

$$\left(\frac{1}{xy}-1\right)-\alpha\left(\frac{x+y}{2}-1\right)=0, x_0=y_0=1. \tag{12}$$

A simple algebra yields:

$$\begin{aligned} y_\alpha &= \tfrac{1}{2\alpha}\left(-\left(2\left(1-\alpha\right)+\alpha x\right)+\sqrt{\left(2\left(1-\alpha\right)+\alpha x\right)^2+\tfrac{8\alpha}{x}}\right), \\ \tfrac{dy_\alpha}{dx} &= \tfrac{1}{2}\left(-1+\tfrac{2(1-\alpha)+\alpha x-4/x^2}{\sqrt{(2(1-\alpha)+\alpha x)^2+8\alpha/x}}\right). \end{aligned} \tag{13}$$

For brevity, below we suppress N, the initial number of tokens delivered to the pool.

Assume that P moves away from its equilibrium value $P_0 = 1$. Let $P > 1$. For the constant sum contract, an arbitrageur can choose a number x, $1 < x \le 2$, and deliver $(x-1)$ of TN_1 tokens to the pool in exchange for getting $(x-1)$ of TN_2 tokens. Her profit or loss is given by

$$\Omega\left(x\right) = \left(P-1\right)\left(x-1\right). \tag{14}$$

Since Ω is a linear function of x, it is rational to exhaust the entire pool, by choosing the following optimal values (x^*, y^*, Ω^*):

$$x^* = 2, \quad y^* = 0, \quad \Omega^* = \left(P-1\right). \tag{15}$$

Similarly, when $P < 1$:

$$x^* = 0, \quad y^* = 2, \quad \Omega^* = \left(1-P\right). \tag{16}$$

The arbitraged portfolio's value is $\pi^*\left(P\right)$, where

$$\pi^*\left(P\right) = \begin{cases} 2, & P \ge 1, \\ 2P, & P < 1. \end{cases}, \tag{17}$$

while the buy and hold portfolio's value is $(P+1)$. The difference ω has the form

$$\omega = \left(P+1\right) - \pi^*\left(P\right). \tag{18}$$

In DeFi, ω is called the impermanent loss. This terminology is misleading because the loss can quickly become permanent when P drifts away from its "equilibrium" value of one. The percentage loss of the realized portfolio compared to the buy and hold portfolio has the form:

$$\lambda = 1 - \frac{|P-1|}{P+1}. \tag{19}$$

We can repeat the above arguments for the constant product contract. Assuming that P deviates from one, an arbitrageur can choose a number $x > 1$ and deliver

$(x-1)\ TN_1$ tokens to the pool and take $(1-y)\ TN_2$ tokens from the pool, where $y = 1/x$. In this case, her profit or loss can be written as follows:

$$\Omega(x) = \left(P\left(1 - \frac{1}{x}\right) - (x-1)\right). \tag{20}$$

The optimality condition has the form

$$\Omega'(x) = \left(\frac{P}{x^2} - 1\right) = 0, \tag{21}$$

so that the corresponding optimal values (x^*, y^*, Ω^*) are

$$x^* = \sqrt{P}, \quad y^* = \frac{1}{\sqrt{P}}, \quad \Omega^* = \left(\sqrt{P} - 1\right)^2. \tag{22}$$

Thus, a constant product collateral pool can never be exhausted. At every stage, the optimal amounts of TN_1 and TN_2 held in the portfolio are equal to $\sqrt{P}$, each. Since the value of both tokens in the portfolio has to be equal, the implied optimal value of TN_2 expressed in terms of TN_1 is $P^* = x^*/y^* = P$. The arbitraged portfolio's value is $\pi^* = 2\sqrt{P}$, while the buy-and-hold portfolio's value is $(P+1)$. The difference ω is given by

$$\omega = (P+1) - 2\sqrt{P}. \tag{23}$$

The corresponding percentage loss is

$$\lambda = 1 - \frac{2\sqrt{P}}{(P+1)} = \frac{\left(\sqrt{P} - 1\right)^2}{(P+1)}. \tag{24}$$

For the mixed rule AMM, the arbitrageur's profit for $P > 1$ has the form:

$$\Omega(x) = \left(P\left(1 - y_\alpha(x)\right) - (x-1)\right), \tag{25}$$

with the optimum achieved at x^*_α, y^*_α, Ω^*_α of the form:

$$y'_\alpha\left(x^*_\alpha\right) = -\frac{1}{P}, \quad y^*_\alpha = y_\alpha\left(x^*_\alpha\right), \quad \Omega^*_\alpha = \left(P\left(1 - y^*_\alpha\right) - \left(x^*_\alpha - 1\right)\right). \tag{26}$$

We find the optimal x^*_α via the Newton-Raphson method starting with a suitable $x^{(0)}_\alpha$:

$$x^{(n+1)}_\alpha = x^{(n)}_\alpha - \frac{y'_\alpha\left(x^{(n)}_\alpha\right) + \frac{1}{P}}{y''_\alpha\left(x^{(n)}_\alpha\right)}. \tag{27}$$

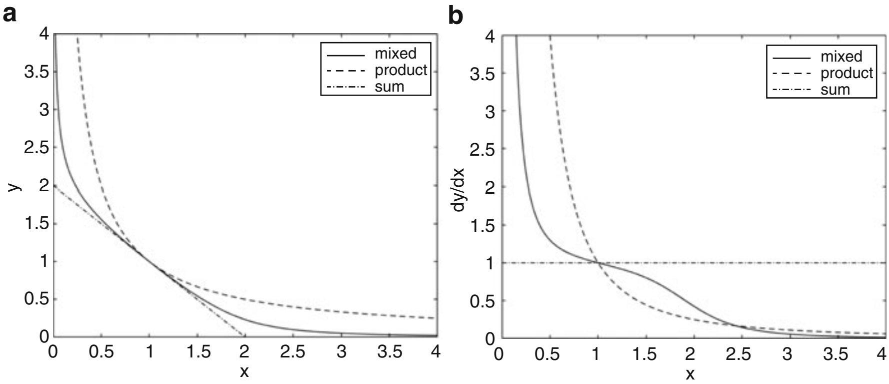

Fig. 5 (**a**) y as a function of x for three different types for constant sum, constant product, mixed rule AMMs. (**b**) The relative price of TN_2 expressed in terms of TN_1 equal to $|dy/dx|$ for constant sum, constant product, and mixed rule AMMs. For the mixed rule, $\alpha = 10$

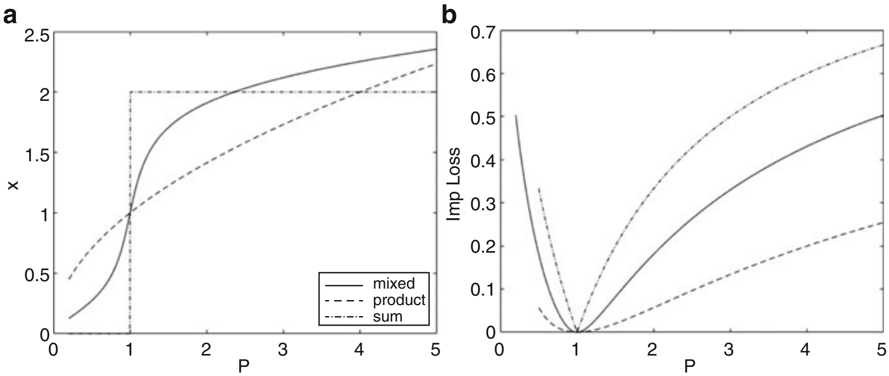

Fig. 6 (**a**) x as a function of P for three different types of AMM: constant sum, constant product, mixed rule; (**b**) impermanent loss as a function of P: constant sum, constant product, mixed rule. For the mixed rule, $\alpha = 10$

Since the Newton-Raphson method has quadratic convergence, ten iterations provide machine accuracy, so that we can set $x_\alpha^* = x_\alpha^{(10)}$. The value of the arbitraged portfolio is

$$\pi^* = x_\alpha^* + P y_\alpha \left(x_\alpha^* \right). \tag{28}$$

The constant sum, constant product, and mixed rule curves, as well as relative prices of TN_2, expressed in terms of TN_1, and impermanent losses are presented in Figs. 5 and 6.

These figures demonstrate that a market maker experiences a loss whenever the tokens' relative price deviates from its equilibrium value. The impermanent loss is relatively small for the constant product rule, moderate for the mixed rule, and very high for the constant sum rule. Even when the price P deviates by a factor of five from its equilibrium value, the impermanent loss for the constant product rule is tolerable, especially compared to the mixed rule.

3.3 P&L Modelling for AMMs

Of course, the objective of any market maker is profit. This profit comes from transaction fees charged by the pool, which have to exceed the impermanent loss caused by a drop in the collateral value below its buy-and-hold level. In this section, we model the profit and loss (P&L) distribution of an AMM. We consider an AMM charging proportional transaction fees every time an arbitrageur or a generic market participant removes tokens of one kind and adds tokens of the other kind. These fees have to cover the impermanent loss and then some. An arbitrageur has to add more tokens to the pool than is required by its constituent rule to cover transaction fees. Consider an AMM defined by the constant product rule. Let ε be a percentage fee. Denote by T_0 and T_1 two time slices. At time T_0 the price is P_0, the pool's composition is (x_0, y_0), and the product value is $\pi_0 = x_0 y_0$. At time T_1 the price is P_1 and the arbitraged pool's composition (x_1, y_1) has to be determined. First, we assume that $P_1 > P_0$, so that, with zero transaction costs, an arbitrageurs would withdraw a certain amount of TN_2 and add the corresponding amount of TN_1. With non-zero transaction costs her decision is more complicated. She can choose a number $x_1 > x_0$ and deliver $(1+\varepsilon)(x_1 - x_0)$ of TN_1 tokens to the pool in exchange for getting $(y_0 - y_1)$ of TN_2 tokens from the pool, where $y_1 = \pi_0/x_1$. The profit or loss is as follows:

$$\Omega(x) = \left(P_1\left(y_0 - \frac{\pi_0}{x}\right) - (1+\varepsilon)(x - x_0)\right). \tag{29}$$

The profit is maximized when

$$\Omega'(x) = \left(\frac{P_1\pi_0}{x^2} - (1+\varepsilon)\right) = 0, \tag{30}$$

so that

$$\begin{aligned} x_1^* &= \max\left(\sqrt{\tfrac{P_1\pi_0}{(1+\varepsilon)}}, x_0\right), \quad y_1^* = \tfrac{\pi_0}{x_1^*}, \\ \pi_1^* &= \tfrac{((1+\varepsilon)(x_1^* - x_0) + x_0)\pi_0}{x_1^*}. \end{aligned} \tag{31}$$

We have to take the maximum in Eq. (31) to ensure that $x_1^* \geq x_0$, so that TN_1 tokens are added to the pool, rather than withdrawn from it. Thus, for $P_1 > P_0$, the adjustment occurs only when

$$P_1 > \frac{(1+\varepsilon)\, x_0}{y_0}. \tag{32}$$

Similarly, for $P_1 < P_0$,

$$\begin{aligned} y_1^* &= \max\left(\sqrt{\tfrac{\pi_0}{(1+\varepsilon) P_1}}, y_0\right), \quad x_1^* = \tfrac{\pi_0}{y_1^*}, \\ \pi_1^* &= \tfrac{((1+\varepsilon)(y_1^* - y_0) + y_0)\pi_0}{y_1^*}, \end{aligned} \tag{33}$$

so that adjustment happens when

$$P_1 < \frac{x_0}{(1+\varepsilon)\, y_0}. \tag{34}$$

Finally, for

$$\frac{x_0}{(1+\varepsilon)\, y_0} \leq P_1 \leq \frac{(1+\varepsilon)\, x_0}{y_0}, \tag{35}$$

it is suboptimal to adjust the composition of the pool, so that $\left(x_1^*, y_1^*\right) = (x_0, y_0)$, $\pi_1^* = \pi_0$.

Equations (32), (34), and (35) show that in the presence of non-zero transaction costs the actual composition of the pool is not only time-dependent, as expected, but, more surprisingly, path-dependent.

Let us study the profitability of a constant product rule AMM. We shall assume that the relative price $P(t)$ is mean-reverting and is driven by an Ornstein-Uhlenbeck process:

$$\begin{aligned} P(t) &= \exp(p(t)), \\ dp(t) &= -\kappa p(t)\, dt + \sigma dW(t), \quad p(0) = 0. \end{aligned} \tag{36}$$

Here $W(t)$ is the Wiener process driving random variations of the log-price, κ is mean-reversion rate, and σ is volatility; κ and σ are measured in the units of $[1/T]$ and $\left[1/\sqrt{T}\right]$, respectively. It is helpful to switch to non-dimensional units. To this end, we introduce $\bar{t} = \kappa t$, $\bar{\sigma} = \sigma/\sqrt{\kappa}$, and rewrite Eq. (36) as follows:

$$
\begin{aligned}
P\left(\bar{t}\right) &= \exp\left(p\left(\bar{t}\right)\right), \\
dp\left(\bar{t}\right) &= -p\left(\bar{t}\right) d\bar{t} + \bar{\sigma} dW\left(\bar{t}\right), \quad p\left(0\right) = 0.
\end{aligned}
\tag{37}
$$

Below we omit overbars for brevity. When σ is small, the log-price is almost deterministic and mean-reverting, when σ is large, it is strongly stochastic. Since small price changes do not result in adjustments of portfolio composition, see Eq. (35), we discretize Eq. (37) with a time step Δt, for instance, 1 day, and rewrite it as follows:

$$
\begin{aligned}
P_{l+1} &= \exp\left(p_{l+1}\right), \\
p_{l+1} &= \left(1 - \Delta t\right) p_l + \sigma \sqrt{\Delta t} \eta_l, \quad p_0 = 0.
\end{aligned}
\tag{38}
$$

Here η_l is the standard normal random variable.

We assume that pool's liquidity provider has a time horizon $T = L \Delta t$ and model the evolution of the system as a whole for L steps. The corresponding dynamics can be described as follows:

$$
\begin{aligned}
p_{l+1} &= \left(1 - \Delta t\right) p_l + \sigma \sqrt{\Delta t} \eta_l, & p_0 &= 0, \\
x_{l+1} &= f\left(\pi_l, p_{l+1}, p_l, x_l, y_l\right), & x_0 &= 1, \\
y_{l+1} &= g\left(\pi_l, p_{l+1}, p_l, x_l, y_l\right), & y_0 &= 1, \\
\pi_{l+1} &= h\left(\pi_l, p_{l+1}, p_l, x_l, y_l\right), & \pi_0 &= 1, \\
P_{l+1} &= \exp\left(p_{l+1}\right), & P_0 &= 1
\end{aligned}
\tag{39}
$$

Here

$$
\begin{aligned}
x_{l+1} &= \begin{cases} \sqrt{\frac{P_{l+1}\pi_l}{(1+\varepsilon)}}, & \frac{(1+\varepsilon)x_l}{y_l} < P_{l+1}, \\ x_l, & \frac{x_l}{(1+\varepsilon)y_l} \le P_{l+1} \le \frac{(1+\varepsilon)x_l}{y_l}, \\ \frac{\pi_l}{y_{l+1}}, & P_{l+1} < \frac{x_l}{(1+\varepsilon)y_l}. \end{cases} \\
y_{l+1} &= \begin{cases} \frac{\pi_l}{x_{l+1}}, & \frac{(1+\varepsilon)x_l}{y_l} < P_{l+1}, \\ y_l, & \frac{x_l}{(1+\varepsilon)y_l} \le P_{l+1} \le \frac{(1+\varepsilon)x_l}{y_l}, \\ \sqrt{\frac{\pi_l}{(1+\varepsilon)P_{l+1}}}, & P_{l+1} < \frac{x_l}{(1+\varepsilon)y_l}. \end{cases} \\
\pi_{l+1} &= \begin{cases} \frac{\left((1+\varepsilon)\left(x_{l+1}-x_l\right)+x_l\right)\pi_l}{x_{l+1}}, & \frac{(1+\varepsilon)x_l}{y_l} < P_{l+1}, \\ \pi_l, & \frac{x_l}{(1+\varepsilon)y_l} \le P_{l+1} \le \frac{(1+\varepsilon)x_l}{y_l}, \\ \frac{\left((1+\varepsilon)\left(y_{l+1}-y_l\right)+y_l\right)\pi_l}{y_{l+1}}, & P_{l+1} < \frac{x_l}{(1+\varepsilon)y_l}. \end{cases}
\end{aligned}
\tag{40}
$$

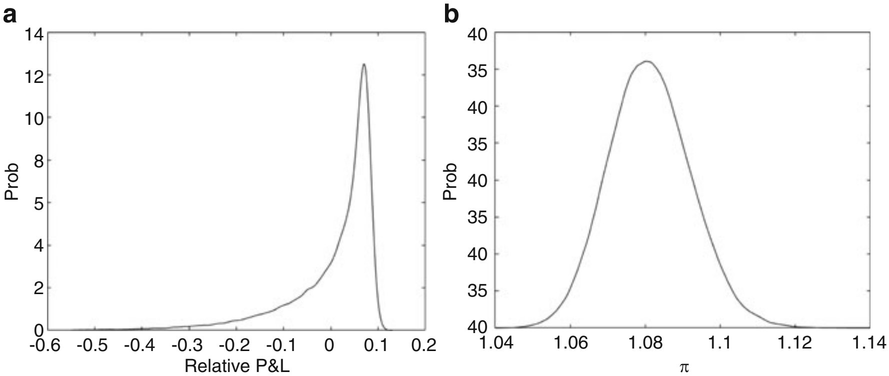

Fig. 7 (**a**) The pdf for P&Loss; (**b**) the pdf for $\pi_L = x_L y_L$. The corresponding parameters are as follows: $L = 200$, $\Delta t = 0.0025$, $\sigma = 1$, $\epsilon = 0.05$

The corresponding $P\&L$ versus the buy-and-hold strategy is given by the following formula:

$$P\&L = (x_L + P_L y_L) - (1 + P_L) . \tag{41}$$

The market making activity makes sense only when $P\&L > 0$. Perhaps, a more informative quantity is the relative $\overline{P\&L}$, which shows the percentage return on market making, compared with the buy-and-hold strategy:

$$\overline{P\&L} = \frac{P\&L}{(1 + P_L)} = \frac{(x_L + P_L y_L)}{(1 + P_L)} - 1. \tag{42}$$

Of course, since the log-price is stochastic, we can only analyze $\overline{P\&L}$ in the probabilistic sense by running Monte Carlo simulations. To this end, we consider M MC paths, which are characterized by a random matrix (η_{ml}), calculate a set of M P&L values, $\left\{\overline{P\&L_m}\right\}$, and study its statistical properties. Figure 7 presents the corresponding results.

It is clear that for the choice of parameters used in Fig. 7 automated liquidity provision is profitable on average. This profitability comes from the fact that the AMM accumulates more tokens at the end of the process than she had in the beginning. This figure illustrates the complicated dependence of $\overline{P\&L}$ on L, σ, ε. The reader, wishing to become an AMM, needs to explore this dependence in detail. One can analyze the P&L generated by the mixed rule AMM by using a similar technique. We leave the corresponding analysis to the reader.

We emphasize that it is exceedingly hard to protect AMMs against meta-threats, such as the underlying SC not being robust enough or the underlying token losing value permanently because of regulatory pressures, poor design, or outright theft of the collateral. Even if none of the above happens, the portfolio consisting of tokens

TN_1, TN_2 can lose value vis-a-vis fiat because both tokens depreciate against fiat currency simultaneously. Engaging in liquidity provision is advantageous to someone who is already invested in tokens and wants to stay invested in them in the long run. Another attractive opportunity is to provide liquidity to the stablecoin universe and view AMM gains (if any) as a way to earn interest on one's investments.

4 Challenges in Blockchain Interoperability

The Internet architecture lends itself to scalability because it (1) permits each AS to employ its own interior routing protocols, while standardizing the interfaces between autonomous systems to permit data-packet flow across these systems; and (2) places the higher layer semantics (content or meaning) of a connection to the edges of the network, where the sender and receiver are located. The intervening autonomous system between the sender and receiver are oblivious to the content of the message being sent. Messages are in fact broken down in to IP packets (datagrams Cerf & Khan, 1974), and each IP packet may traverse differing routes across different sets of autonomous systems end-to-end.

In order for blockchain systems to have a high degree of interoperability, each blockchain system must be free to implement its interior ledger maintenance protocols (e.g. consensus protocol, ledger data structures, etc.), with standardized APIs defined for cross-blockchain transfers of the digital value-representation tokens (assets). Secondly, for digital assets that are blockchain-based the notion of "value" (economic value) must be separated from the blockchain infrastructures that manage the value-representation tokens. That is, if we view on-chain tokens as a counterpart of the IP datagrams on the Internet—where the economic value is discernible only to the sender and receiver at the edges—then a proper interoperable blockchain architecture must permit the tokens to traverse (hop across) multiple blockchain networks in an agnostic manner without loss to the economic value represented by the tokens Hardjono et al. (2019).

A corollary of the last point above is the need to separate the economic value represented by the tokens from the operating costs of the blockchain infrastructure supporting the tokens. More specifically, the degree of user-demand for infrastructure services to transact with a token (i.e. move token/asset ownership from one public-key to another) should not be tied to the economic value of the token, as this renders the infrastructure to becoming too costly to employ. A case in point is Bitcoin (Gogo, 2020; Godbole, 2020) and Ethereum (Shevchenko, 2020), both of which have become too costly to be used for small-amount transactions. Alternative incentives mechanisms for infrastructure (node) operators need to be developed based on this principle of the separation of token economic value from infrastructure operating costs. For example, a traditional subscriber-fee model akin to that used by Internet ISPs has been proposed in Hardjono et al. (2020).

It is worth reflecting that if starting in the 1990s the ISPs charged consumers for Internet access based on the sheer number of IP datagrams routed through their subnets, the Internet would not be the success that it is today.

4.1 Interoperability Across Blockchains and DLT Systems

In considering the use of blockchain technology to represent economic value – through the on-chain value-representation tokens or ledger entries—it is useful to address the problem of interoperability at two levels (Hardjono et al., 2019):

- *Interoperability at the value level*: In order to perform transactions using virtual assets there must be agreement on the notion of value and the standardization of mechanism to represent economic value as tokens within blockchain systems.

  A core part of interoperability at the value level is the standardization of *asset profiles* (asset prospectus) definition. This permits transacting parties to refer to the same definition of the virtual asset to be exchanged. We discuss the notion of asset profiles in Sect.4.3.
- *Interoperability at the mechanical protocol level*: Standard protocols must be developed to perform value conversions (ingress and egress points in Fig. 8) and to perform token-transfers across blockchain networks (transfer points in Fig. 8).

  The standard protocols used to carry out the token-transfers across blockchain networks must be agnostic to the economic value represented by the token. We discuss the technical protocol and its desirable properties in Sect. 4.4.

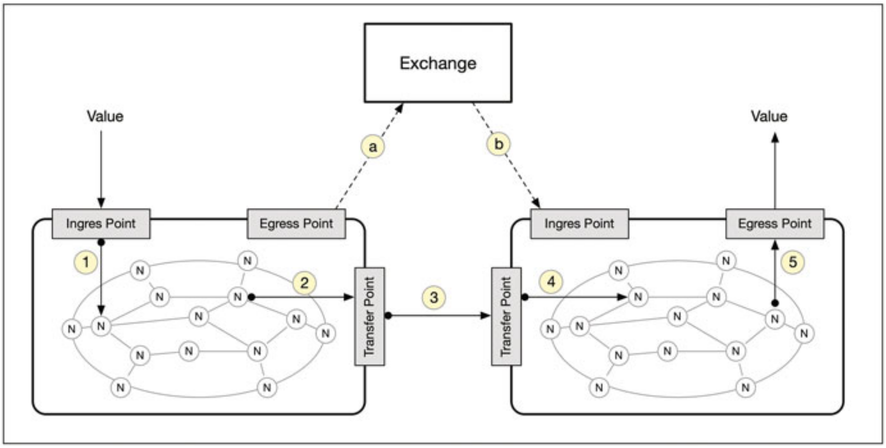

Fig. 8 Overview of interoperability architecture for blockchain networks

4.2 Interfaces for Interoperable Functions

An interoperable blockchain architecture must distinguish between *value-representation conversion points* in the blockchain network from *representation translation points* across blockchain networks (see Fig. 8):

- *Value-representation conversion points*: These are the *ingress points* for value to enter the blockchain system, and the *egress points* where value departs from the blockchain system. When a blockchain system receives economic value from an external source (e.g. a non-DLT representation) via an ingress point, it creates a representation of that value in the form of the token data-structure defined in that blockchain system. This is shown as Step 1 in Fig. 8. Similarly, when a DLT value-representation is to be removed from a blockchain, it departs the blockchain via an egress point. The ledger of the blockchain is marked to indicate that the token no longer exists (i.e. it has been invalidated).

  It is important to note that in some blockchain systems the tokens are an inherent part of the system and that "assets" (their token representation) never enters or leaves the blockchain (e.g. Bitcoin). This is because the value of the virtual asset (i.e. BTCs) not a derivative of an underlying asset (Ankenbrand et al., 2020).
- *Token transfer points*: These are the points in a blockchain network where tokens can be transferred directly from one blockchain to another without any change to the economic value represented by the token. This is shown as Step 3 in Fig. 8. A data structure (format) translation of the token may occur if the two blockchain infrastructures employ differing interior ledger data-structures.

  A successful unidirectional token transfer means that the token data-structure in the origin blockchain (Step 2) is destroyed (or marked as being no longer valid), while a new token data-structure is created (added) in the destination blockchain (Step 4). No change to the economic value presented by the tokens must occur during such transfers.

In cases where circumstances prevent the traversal by a token across blockchain networks (e.g. incompatible ledger data structures, jurisdiction constraints, etc.) then both value-representation conversion and token-format translation must occur with the help of a mediating third party (e.g. crypto-exchanges or similar VASPs). This is illustrated in Step (a) and Step (b) of Fig. 8. This two-stage process consist of the following. The value departs the origin blockchain in Step (a) through an *egress point* to the exchange entity. Here, the value-representation token in the origin blockchain system is extinguished or destroyed. Secondly, the exchange entity must inject that value at the ingress point (Step (b)) at the destination blockchain system—resulting in the creation of a new value-representation token according to the data-structure employed by the destination blockchain. In this case, the third-party exchange entity must be a participant in both blockchain systems, and it must have the means to perform the process (e.g. it holds sufficient fiat currencies).

4.3 Asset Profiles: Standardizing Asset Prospectus Documents

We define the *asset profile* as prospectus of a regulated asset that includes information and resources describing the virtual asset. This includes the asset name/code, issuing authority, denomination, jurisdiction, and the URLs and mechanisms to validate the information. An asset profile document or claim must be digitally signed by the issuer of the virtual asset. It is an assertion or statement regarding the true existence of the virtual asset within a given system (blockchain-based or otherwise), within a legal jurisdiction. The asset profile is independent from the specific instantiation of the asset (on a blockchain or otherwise) and independent from its instance-ownership information.

There are a number of information fields that could be expressed within an asset profile statement or document for a given virtual asset. Some examples include, but not limited to:

- *Issuer*: This information field pertains to the legal entity that issues or creates the virtual asset. The Issuer could be a single corporation, a community of entities, a government, etc.
- *Jurisdiction*: This information field refers to the legal jurisdiction where the virtual asset is defined and recognized.
- *Virtual asset code*: This information field contains globally unique alphanumeric value assigned to the asset. This ensures that users and systems can acurately refer to a virtual asset, without any errors or ambiguities.
- *Virtual asset type*: This refers to the underlying asset or collateral upon which the virtual asset is based. Examples include bankable assets (i.e. fiat currency), company shares (e.g. equity), tangible assets (e.g. real estate), etc.

  Note that this information field in the asset profile could be "none", meaning that the value of the virtual asset is not derived from any underlying asset (as is the case with Bitcoin) (Ankenbrand et al., 2020).
- *Total supply*: This information field refers to total supply of the asset, which could be fixed, conditional (e.g. upon some external factors), or flexible (where the supply of the asset is managed flexibly by authorized parties).
- *Issuance date*: This information field refers to the date when the virtual asset become available. This date should be identical to the date of issuance and signing of the asset profile document (e.g. JSON file).
- *Validation endpoint*: This is the URI/URL where any entity can verify the validity status of the asset profile document (e.g. JSON file).
- *Digital signature of Issuer*: This is the digital signature of the Issuer over the entire asset profile document. Typically the digital signature field includes a copy of the public-key of the Issuer, and the signature is achieved using standard algorithms (US Congress, 2000; Bartel et al., 2015; Jones et al., 2017).

Figure 9 provides an overview of the two levels of interoperability and the need for standardization to occur at each level. Alice seeks to transfer ownership of a virtual asset to Bob who is located in a different blockchain. Both sides have agreed

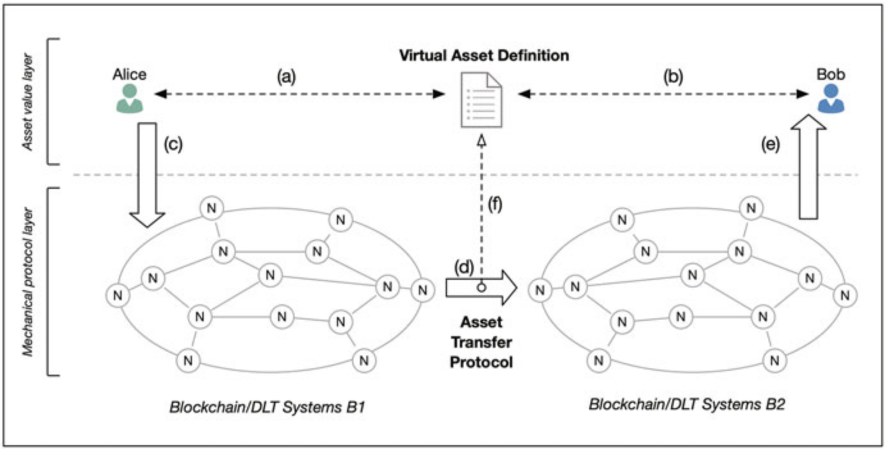

Fig. 9 Asset transfers based on a standard asset definition profile

upon the definition of the asset to be transferred (Steps (a) and (b)). Alice then invokes the transfer protocol (e.g. smart contract, application, etc.) in Step (c), which results in the asset transfer protocol executing between the two blockchain systems B1 and B1 (Step (d)). The protocol itself must have the means to refer to the definition of the asset being transferred (e.g. reference (f) in Fig. 9).

4.4 The Gateway Model: A Framework for Blockchain Interoperability

In this section, we discuss blockchain interoperability following the notion of the *token transfer points* mentioned above, manifesting these transfer points in the form of *blockchain gateways* (Hardjono et al., 2019, 2020). A gateway stands in“ front” of its blockchain systems, and has read/write access to the ledger and other interior resources within its blockchain domain. A gateway in one blockchain interacts with another gateway in a different blockchain in the task of transferring virtual assets between them unidirectional.

Similar to a routing autonomous system being composed of one or more (possibly nested) routing domains, the framework of Hardjono et al. (2019) views a blockchain domain as consisting of *interior nodes* and *gateway nodes*:

- *Interior nodes*: These are nodes and other entities whose main task is maintaining ledger information and conducting transactions within one blockchain domain.

  For certain blockchain configurations (e.g. private or permissioned) the interior nodes are prohibited from engaging external entities without authorization.

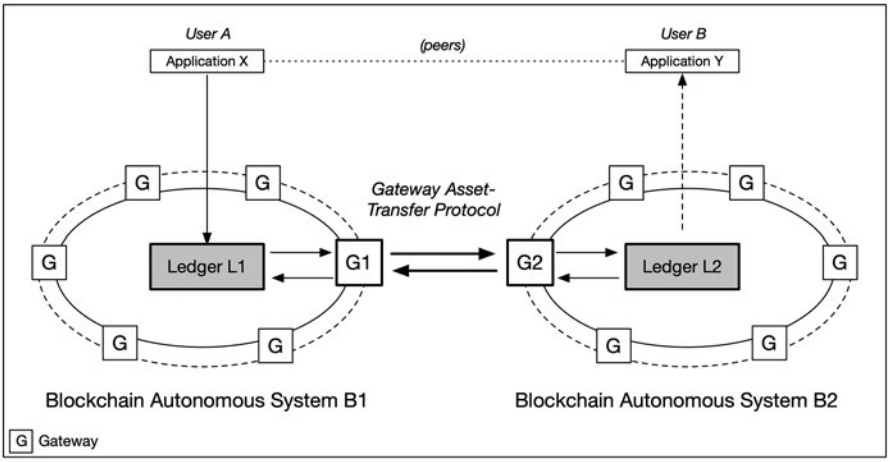

Fig. 10 Overview of the gateway-to-gateway asset transfer protocol

- *Gateway nodes*: These are nodes and other entities whose main task is dealing with cross-blockchain asset transfers involving different blockchain systems.

Figure 10 provides a high level illustration of gateway nodes G within two blockchain domains (interior nodes are not shown). Although Fig. 10 shows a small number of nodes G to be designated as inter-domain gateway nodes, ideally all nodes in a given blockchain system should have the capability (i.e. correct software, hardware, trusted computing base) to become gateway nodes. This allows dynamic groups (subsets) of the population of nodes to become *gateway groups* that act collectively on behalf of the blockchain system as a whole (Hardjono & Smith, 2019).

The following assumptions and principles underlie the interoperability framework of Hardjono et al. (2019), and they correspond to the design principles of the Internet architecture:

- *Opaque blockchain-resources principle*: The interior resources of each blockchain system is assumed to be opaque to (hidden from) external entities. Any resources to be made accessible to an external entity must be made explicitly accessible by a gateway node with proper authorization.

  The opaque resources principle permits the interoperability architecture to be applied in cases where one (or both) blockchain systems are permissioned (private). It is the analog of the autonomous systems principle in IP networking (Clark, 1988), where interior routes in local subnets are not visible to other external autonomous systems.
- *Externalization of value principle*: The gateway-to-gateway protocol must be agnostic (oblivious) to the economic or monetary value of the virtual asset being transferred.

> The value-externalization principle permits asset transfer protocols to be designed for efficiency, speed and reliability—independent of the changes in the perceived economic value of the virtual asset. It is the analog of the end-to-end principle in the Internet architecture (Saltzer et al., 1984), where contextual information (economic value) is placed at the endpoints of the transaction. In the case of virtual asset transfers, the originator and beneficiary at the respective blockchain systems are assumed to have a common agreement regarding the economic value of the asset.

The goal of a blockchain interoperability framework based on interoperable gateways is to permit two (2) gateway nodes belonging to distinct blockchain systems to conduct a virtual asset transfer between them, in a secure and non-repudiable manner while ensuring the asset does not exist simultaneously on both blockchains (double-spend problem). The notion of a gateway is used because we recognize that blockchain technology is evolving and that in many cases the interior technical constructs in these blockchains maybe incompatible with one another. The architecture therefore assumes that certain types of nodes (gateway nodes) will be equipped with an asset transfer protocol and other relevant resources that permits greater interoperability across these incompatible blockchain systems. In a sense, a gateway "hides" the complexity of its blockchain, and in turn exposes standard APIs to other gateways in order to interoperate.

The resources within a blockchain system (e.g. ledgers, public-keys, consensus protocols, etc.) are assumed to be opaque to external entities in order to permit a resilient and scalable protocol design that is not dependent on the interior constructs of particular blockchain systems. This ensures that the virtual asset transfer protocol between gateways is not conditioned or dependent on these local technical constructs. The role of a gateway therefore is also to mask (hide) the complexity of the interior constructs of the blockchain system that it represents. Overall this approach ensures that a given blockchain system operates as a true autonomous system.

It is important to note that the opaque resources (ledgers) principle has implications on smart contract cross-chain conditionals, such as cross-chain hash-locks (Nolan, 2013) and time-locks. Many proposals for cross-chain "atomic swaps" are designed with the underlying assumption that the ledgers on both sides are public or permissionless (e.g. see Ezhilchelvan et al., 2018; Zakhary et al., 2019; Herlihy, 2018). This means that both Alice and Bob are able to read (and invoke) each other's hash-lock smart contract at their respective blockchains. However, we believe that this underlying assumption is unrealistic given the fact that many blockchain systems today are private (permissioned).

The point of the opaque resources principle is to enable the design of cross-blockchain asset transfer protocols under the strictest condition—namely that both blockchains are private and their ledgers and smart-contracts inaccessible to each other. Interaction between them are possible only through their respective gateways. If an asset transfer protocol works for two private blockchains via gateways, then it should also work for cases where one or both of the blockchains are public or permissionless.

4.5 Protocols for Asset Transfers: Desirable Properties

At the mechanical protocol level, there are a number of desirable properties of an asset transfer protocol across blockchain/DLT systems (Hardjono et al., 2020):

- *Atomicity*: An asset transfer must either commit or entirely fail (where failure means there are no changes to asset ownership in the origin blockchain).
- *Consistency*: An asset transfer (commit or fail) must always leave both blockchain systems in a consistent state, where the asset in question is located in one blockchain only. A protocol failure must not result in a "double-existence" of the asset (leading to double-spend in the two blockchain systems respectively).
- *Isolation*: While the asset transfer is occurring, the asset ownership cannot be modified. That is, some kind of temporary disablement of the asset at the origin blockchain must be used (e.g. locking on the ledger, escrow to a gateway, etc.). This is to prevent the owner (who requested the cross-chain transfer) from double-spending the asset locally while the transfer protocol is running.
- *Durability*: Once an asset transfer has been committed on both the origin blockchain and destination blockchain, this commitment must remain true regardless of gateway crashes or blockchain unavailability (e.g. blockchain slowdown as in the case of CryptoKitties in Ethereum BBC News, 2017).

It is crucial to note that these properties must hold true regardless of whether one or both of the blockchain systems are private (permissiond) or public (permissionless).

4.6 Phases of Gateway-to-Gateway Asset Transfers

An asset transfer protocol between two blockchain systems is carried out by two (2) gateway nodes that represent the two respective blockchain systems. A successful transfer results in the asset being extinguished or marked on the origin ledger by the origin-gateway, and for the asset to be introduced by the destination-gateway into the destination ledger. The mechanism to extinguish or introduce an asset from/into a ledger is dependent on the specific blockchain system.

The interaction between the two gateways is summarized in Fig. 11, where the origin blockchain is B1 and the destination blockchain is B2. The gateways are denoted as G1 and G2 respectively.

The gateway nodes must implement one (or more) transactional commitment protocols (in Phase 2) that permit the coordination between two gateways, and the final commitment of the asset transfer. The choice of the commitment protocol

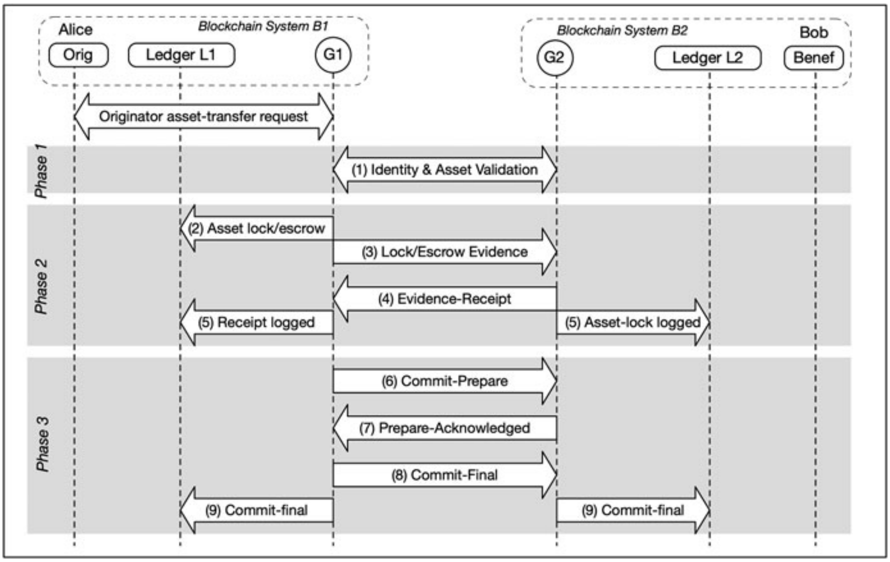

Fig. 11 Overview of the phases of the gateway asset-transfer protocol (after Hardjono et al., 2020)

(type/version) and the corresponding commitment evidence must be negotiated between the gateways during Phase 1. For example, the gateways G1 and G2 may implement the classic 2-Phase Commit (2PC) protocol (Gray, 1981; Traiger et al., 1982) or other variants (e.g. 3PC) as a means to ensure efficient and non-disputable commitments to the asset transfer.

4.6.1 Phase 1: Pre-transfer Asset and Identities Verification

In this phase the gateways G1 and G2 initiate a connection to each other in order to perform a number of validation functions. Some of these are as follows:

- Exchange of parameters for secure channel establishment between G1 and G2.
- Delivery of asset-profile information and asset-holder information, including originator and beneficiary identities and public keys (per the Travel Rule FATF, 2018), and the gateway owner (VASP) identities and public keys
- Exchange of parameters related to the blockchain systems B1 and B2, the commitment mechanism to be employed between G1 and G2, and the form of the asset-lock evidence to be delivered by G1.

4.6.2 Phase 2: Evidence of Asset Locking or Escrow

In this phase, gateway G1 must provide gateway G2 with sufficient evidence that the asset on blockchain B1 is in a locked state (or escrowed) under the control of G1 on ledger L1, and safe from double-spend on the part of its current owner (the originator).

The precise form of the evidence is dependent on the blockchain system in B1, and must be previously agreed upon in Phase 1. The purpose of this evidence is for dispute resolution between G1 and G2 (i.e. entities who own and operate G1 and G2 respectively) in the case that double-spend is later detected.

The gateway G2 must return a signed receipt to G1 of this evidence in order to cover G1 in the case of later denial by G2.

4.6.3 Phase 3: Final Commitment of Transfer

In this phase gateway G1 indicates to G2 its readiness to finally commit to the transfer, and vice versa. Both messages must be signed by G1 and G2 respectively in case of later (post-transfer) disputes.

Gateway G1 marks the ledger L1 that the virtual asset is no longer associated with the public-key of previous owner (originator) and that the asset instance no longer exists on the blockchain system B1. Similarly, gateway G1 marks the ledger L2 in blockchain system B2 to indicate that henceforth the asset is associated with the public-key of the new owner (beneficiary).

Optionally, both G1 and G2 may exchange the local ledger marking information (e.g. block number and transaction number) with each other. This information may aid in future search, audit and accountability purposes from a legal perspective.

4.7 Open Challenges in Interoperability

There are a number of open issues that are related to the asset transfer protocol between gateway nodes. Some of the issues are due to the fact that blockchain technology is relatively new, and that technical constructs designed for interoperability have yet to be addressed. Some of the issues are due to the nascency of the virtual asset industry and lack of conventions, and therefore require industry collaboration to determine these.

- *Global identification of blockchain systems and public-keys*: There is currently no standard nomenclature to identify blockchain systems in a globally unique manner. The analog to this is the AS-numbers associated with IP routing autonomous systems. Furthermore, an address (public-key) may not be unique to one blockchain system. An entity (e.g. user) may in fact employ the same public-key at multiple distinct blockchain systems simultaneously.

- *Standard APIs for Cross Blockchain Transfers* As mentioned previously, standard protocols are needed to perform token-transfers across blockchain networks in a manner that is agnostic to the economic value represented by the token.

  Efforts are underway to begin defining the APIs, messages and payloads for asset transfers across blockchain systems (Hardjono et al., 2020; Hargreaves & Hardjono, 2020).
- *Commitment protocols and forms of commitment evidence*: Commitment protocols for asset transfers across blockchain systems should be standardized based on existing well-deployed transaction systems (e.g. based on 2-Phase Commit Gray, 1981; Traiger et al., 1982).

  The forms of commitment evidence also need to be standardized for families of blockchain systems that employ similar or compatible ledger data structures (e.g. Ethereum and Quorum).

5 Conclusions

This chapter discussed the emerging field of blockchains and distributed ledgers and introduced two important concepts—blockchain intraoperability and interoperability. We showed that intraoperability and interoperability are achievable by using AMMs and gateway asset transfer protocols, respectively. While there are many specific issues left for future research, the general directions are clear. The successful design of intra- and interoperability mechanisms is a must for making blockchain technology a viable tool for solving some of our times' most challenging technological problems.

We are grateful to Prof. Sandy Pentland and Dr. Marsha Lipton for their invaluable help.

References

Angeris, G., Kao, H. T., Chiang, R., Noyes, C., Chitra, & T. (2019). *An analysis of Uniswap markets*. https://arxiv.org/abs/1911.03380

Ankenbrand, T., Bieri, D., Cortivo, R., Hoehener, J., & Hardjono, T. (2020). *Proposal for a Comprehensive (Crypto) Asset Taxonomy* . https://arxiv.org/abs/2007.11877

Bartel, M., Boyer, J., Fox, B., LaMacchia, B., & Simon, E. (2015). *XML Signature Syntax and Processing Version 2.0. W3C Candidate Recommendation, W3C* (2015). Available at http://www.w3.org/TR/2015/NOTE-xmldsig-core2-20150723/

BBC News (2017). *CryptoKitties craze slows down transactions on Ethereum*. BBC News . https://www.bbc.com/news/technology-42237162

Cerf, V. G., & Khan, R. E. (1974). A protocol for packet network intercommunication. *IEEE Transactions on Communications, 22*, 637–648.

Clark, D. (1988). The design philosophy of the DARPA internet protocols. *ACM Computer Communication Review – Proc SIGCOMM 88, 18*(4), 106–114.

Egorov, M. (2019). StableSwap - efficient mechanism for Stablecoin liquidity. *White paper*.

Ezhilchelvan, P., Aldweesh, A., van Moorsel, A. (2018). Non-blocking two phase commit using blockchain. In *Proceedings of the 1st Workshop on Cryptocurrencies and Blockchains for Distributed Systems (CryBlock'18)* (pp. 36–41). New York: Association for Computing Machinery. https://doi.org/10.1145/3211933.3211940

FATF (2018). International standards on combating money laundering and the financing of terrorism and proliferation. *FATF Revision of Recommendation 15, Financial Action Task Force (FATF)* . Available at: http://www.fatf-gafi.org/publications/fatfrecommendations/documents/fatf-recommendations.html

Godbole, O. (2020). Bitcoin Transaction Fees Rise to 28-Month High as Hashrate Drops Amid Price Rally. *Coindesk* https://www.coindesk.com/bitcoin-transaction-fees-hashrate

Gogo, J. (2020). *Bitcoin Transaction Fees Soar 550% in a Month, BCH, Dash Transactions Much Cheaper*. Bitcoin.com. https://news.bitcoin.com/bitcoin-transaction-fees-soar-550-in-a-month-as-price-surges-bch-dash-cheapest-networks

Gray, J. (1981). The transaction concept: Virtues and limitations. In *Very Large Data Bases – Proceedings of the 7th International Conference* (pp. 144–154). Cannes, France

Gupta, M. (2017). *Blockchain for DUMMIES*. Wiley.

Hardjono, T., Hargreaves, M., & Smith, N. (2020). *An Interoperability Architecture for Blockchain Gateways*. Internet-draft draft-hardjono-blockchain-interop-arch-01, IETF . https://www.ietf.org/archive/id/draft-hardjono-blockchain-interop-arch-01.txt

Hardjono, T., Lipton, A., & Pentland, A. (2019). Towards an interoperability architecture blockchain autonomous systems. *IEEE Transactions on Engineering Management, 67*(4), 1298–1309. https://doi.org/10.1109/TEM.2019.2920154

Hardjono, T., Lipton, A., & Pentland, A. (2020). Towards a contract service provider model for virtual assets and VASPs. In *P2P Financial Systems International Workshop 2020* . https://arxiv.org/abs/2009.07413

Hardjono, T., & Smith, N. (2019). Decentralized trusted computing base for blockchain infrastructure security. *Frontiers Journal - Special Issue on Finance, Money & Blockchains, 2*. https://doi.org/10.3389/fbloc.2019.00024

Hargreaves, M., Hardjono, T. (2020). *Open Digital Asset Protocol*. Internet-draft draft-hargreaves-odap-01, IETF. https://www.ietf.org/archive/id/draft-hargreaves-odap-01.txt

Heaven, W. D. (2020). A plan to redesign the internet could make apps that no one controls. *MIT Technology Review*. https://www.technologyreview.com/2020/07/01/1004725/redesign-internet-apps-no-one-controls-data-privacy-innovation-cloud/

Herlihy, M. (2018). Atomic cross-chain swaps. In *Proceedings of the ACM Symposium on Principles of Distributed Computing PODC'18* (pp. 245–254). New York, NY: Association for Computing Machinery. https://dl.acm.org/doi/pdf/10.1145/3212734.3212736

Huang, Z. (2020). China-backed crypto guru wants to unify world's blockchain. *Bloomberg*. https://www.bloomberg.com/news/articles/2020-07-26/china-backed-crypto-guru-wants-to-unify-the-world-s-blockchains

Jones, M., Bradley, J., & Sakimura, N. (2017). *JSON Web Signature (JWS)* . IETF Standard RFC7515 https://tools.ietf.org/html/rfc7515

Lipton, A., & Treccani, A. (2021). *Blockchain and distributed ledgers: Mathematics, technology, and economics. WSPC*, Singapore.

Martin, J. (2020). Dfinity's Internet Computer Could Be a Truly Decentralized Alternative. *Cointelegraph*. https://cointelegraph.com/news/dfinitys-internet-computer-could-be-a-truly-decentralized-alternative

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*.

Nolan, T. (2013). *Alt chains and atomic transfers*. https://bitcointalk.org/index.php?topic=193281.msg2224949-msg2224949

Saltzer, J., Reed, D., & Clark, D. (1984). End-to-end arguments in system design. *ACM Transactions on Computer Systems, 2*(4), 277–288.

Schär, F. (2020). Decentralized finance: On blockchain-and smart contract-based financial markets. *Available at SSRN 3571335*.

Shevchenko, A. (2020). Using a DeFi protocol now costs more than $50 as Ethereum fees skyrocket. *Cointelegraph* (2020). https://cointelegraph.com/news/using-a-defi-protocol-now-costs-more-than-50-as-ethereum-fees-skyrocket

Traiger, I. L., Gray, J., Galtieri, C. A., & Lindsay, B. G. (1982). Transactions and consistency in distributed database systems. *ACM Transactions on Database Systems (TODS), 7*(3), 323–342.

US Congress (2000). *Electronic signatures in global and national commerce act*.

Whittemore, N. (2020). JPMorgan launches JPM coin: Welcome to the private currency Era. *Coindesk*. https://www.coindesk.com/jpmorgan-launches-jpm-coin-private-currency

Zakhary, V., Agrawal, D., & Abbadi, A. E. (2019). *Atomic Commitment Across Blockchains*. https://arxiv.org/pdf/1905.02847.pdf

Zhang, Y., Chen, X., & Park, D. (2018). Formal specification of constant product (xy= k) market maker model and implementation. *White paper*.

Distributed Ledger Technology and Fully Homomorphic Encryption: Next-Generation Information-Sharing for Supply Chain Efficiency

Daniel P. Hellwig and Arnd Huchzermeier

1 Introduction

Information-sharing facilitates the integration of supply chains and is essential for enterprises' survival. While advancements in information and communication technology brought secure information-sharing in reach and have lowered the technological barriers to it, a chronic lack of information-sharing within and across companies remains. Companies also have a direct incentive not to share certain operational attributes (e.g., inventory levels, order sizes, wholesale prices), but the entire ecosystem would benefit should aggregated information (e.g., average inventory levels, average order sizes) become available, as that would facilitate planning, inform decisions, and limit the emergence of operational distortions that result from the bullwhip effect. This chapter outlines promising approaches and methodologies for resolving these chronic pain points by leveraging both distributed ledger ("Blockchain") technology (DLT) and the newly emerging capabilities of fully homomorphic encryption (FHE).

DLT is an internet-based technology that is valued for its ability to publicly validate, record, and distribute transactions in immutable, encrypted distributed ledgers. DLT is arguably uniquely positioned to counteract information distortion and increase the transparency of data flows across supply chains. By enabling selective, secure, anonymous, and automatic information-sharing, DLT can address key concerns among supply chain participants that have been unwilling to engage in selective information-sharing.

FHE allows mathematical operations to be performed on encrypted ciphertexts such that encrypted results on the original input are obtained without knowledge of

D. P. Hellwig (✉) · A. Huchzermeier
WHU – Otto Beisheim School of Management, Vallendar, Germany
e-mail: daniel.hellwig@whu.edu; arnd.huchzermeier@whu.edu

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_2

the encryption key. Thus, FHE lends itself well to preserving privacy in outsourced storage and computation and to encrypting and outsourcing data to third-party environments for processing. The concept of FHE has existed since the late 1970s (Rivest et al., 1978a), but a scheme for its implementation was proposed only in the early 2000s, and viable implementations in real-world applications have been demonstrated only recently (Gentry et al., 2013).

This chapter covers the emergence of both DLT and FHE over the past decade and, without delving into mathematical details, presents an approach that combines DLT with FHE to bridge information-sharing limitations in hyperconnected supply chains without compromising data control. Specifically, the chapter addresses how DLT can facilitate validation of trustless data ownership and how FHE can be leveraged to ensure zero-trust information-sharing. The chapter also introduces real-world applications of both DLT and FHE to illustrate their transformative potential. We then outline the proposed design for an end-to-end approach to supply chain management (SCM) and information-sharing that leverages DLT and FHE to create a process that enables zero-trust and anonymous sharing of summary statistics in multi-echelon supply chains without relying on a third party that can access any of the individual participants' data. We conclude by highlighting the practical of implementing both technologies, as well as their implications for broader adoption.

2 Data-Sharing Limitations in Hyperconnected Supply Chains

The 2020 COVID-19 (C-19) pandemic has caused significant disruptions in global supply operations, the most visible signs of which have been shortages in supplies from face masks to baker's yeast but also resourcing shortages, production interruptions, and delivery delays on a global scale (Kluger, 2020). Disruptions in supply chain operations are not uncommon, as natural disasters like storms and earthquakes pose significant challenges to the local and global flows of materials and interrupt supply chains from both a production and a logistics perspective. However, the C-19 pandemic is unusual in the range, severity, and duration of its impact, and given the hyperconnected nature of our supply chain networks, it is having a significant global cascading effect that has prompted governments to make erratic policy and economic decisions under high levels of uncertainty and led to panic and urgency among the medical community and shoppers in general (Haren & Simchi-Levi, 2020). The ensuing perceived and actual shortages of products have led to irrational consumer behaviours, such as over-buying and hoarding, resulting in cascading volatility throughout global supply chains.

While C-19 has exerted unprecedented pressure on supply chains, the inefficiencies in their operations are a "pre-existing condition" that the pandemic has only amplified. The main culprit is incomplete or distorted information-sharing among supply chain participants, leading to limited knowledge about demand and localized control of the supply chain among customers, retailers, suppliers, and manufacturers, each of which then influences the entire chain with inaccuracies in their forecasted orders. Such lack or distortion of information as it travels from one end of a supply chain to the other can result in significant inefficiencies that manifest as excessive inventory investments, lacklustre revenues, backlogs, mediocre customer service, missed production and delivery schedules, and waste (Lee et al., 1997). Consider the case of the egg supply in Singapore in March and April 2020, when eggs were frequently missing from grocery shelves, and distributors in the import-dependent city state responded by increasing their stocks. Fast forward to June, and the distributors had to throw away more than 250,000 eggs because of oversupply (https://www.us.jll.com/en/trends-and-insights/investor/too-many-eggs-covid-19-turns-focus-on-the-bullwhip-effect). Another prominent example was the widespread demand for toilet paper in early 2020, when the bullwhip effect swung from under-stocking to over-stocking: Because of concerns about supply during the C-19 pandemic, customers hoarded toilet paper, causing many retailers to run out of stock. However, since there was no actual lack of toilet paper, demand for the product dropped when it was back in stock as consumers used their existing stocks. Therefore, given the dynamic nature of supply chains and the often-conflicting interests of its participants, transparent, rapid, and secure information transmission has become more critical than ever for the accurate supply, production, and distribution of inventory (Lee et al., 1997).

2.1 The Bullwhip Effect

A particularly well-documented case of systematic information distortion is the bullwhip or whiplash effect, in which small variations in demand at the retail level are amplified along the supply chain, increasing fluctuations at the manufacturer level (Bray & Mendelson, 2012a). The bullwhip effect occurs when retailers are particularly reactive to positive variations in consumer demand, causing them to change their expectations about future demand, an effect that propagates along the supply chain (Isaksson & Seifert, 2016) until variance in upstream orders exceeds that of downstream sales. For example, if a store perceives a coming shortage of a certain good, and the retail demand is 10 units, the store may choose to order 20 units to be prepared, which can prompt the distributor to order 30 units from the manufacturer, who may then decide to produce 50 units just to be safe, thus amplifying the increase in demand from 20 to 50 units. Therefore, the upstream manufacturer who considers only its immediate order data will be misled as a lack of information-sharing regarding the inventory levels of adjacent supply chain members and customers' point-of-sale data forces each entity to operate independently using incomplete information, which results in optimization of local, rather than global, operational efficiencies (Chaharsooghi et al., 2008). The expenses incurred by a myopic view of supply chain dynamics include the cost of excess raw materials and supplies, unplanned manufacturing and warehousing expenses, and additional transportation costs.

The bullwhip phenomenon plagues many markets. At Hewlett-Packard, orders that resellers place with the printer division have much larger swings and variations than customer demand does, and the swings in the orders to the company's integrated circuit division have been even worse (Lee et al., 1997). Similarly, Procter & Gamble found that variability in distributors' diaper orders cannot be accounted for by fluctuations in consumers' demand (Taylor & Brunt, 2010). The dynamic random-access memory (DRAM) market has also seen high volatility relative to the computer market (Lee et al., 1997). Case studies that analyse hospital capacity and laboratory scheduling have similarly demonstrated the bullwhip effect in the healthcare sector (Sethuraman & Tirupati, 2004). More recently, the bullwhip effect has been observed in the operations of the financial system (e.g., following the 2008 global financial crisis), which caused disruptions in the global equity markets and supply chain financing. As a result, companies were forced to reduce their working-capital targets and to find new means of financing their business plans, initiatives that caused a significant shock in the supply chains across the world, leading to an inventory-driven and financial bullwhip effect (Tate & Ellram, 2019).

2.2 Causes of the Bullwhip Effect

The dynamics of supply chain behaviour were first formulated by the Massachusetts Institute of Technology's Jay Wright Forrester (1958), who offered one of the most compelling illustrations of the bullwhip effect, the "beer distribution game". In its original form, the game simulates a traditional four-stage supply chain in which players take the roles of retailer, wholesaler, distributor, and factory. To play, each sub-group must fulfil the incoming orders of beer (i.e., the demand). To win, players must keep their costs along the supply chain as low as possible while maximizing their sales revenue—that is, they must minimize fluctuations along the supply chain by balancing supply and demand. The rules of the game are simple: information exchanges in terms of order quantities can occur only between successive layers in the supply chain (e.g., between a retailer and wholesaler). Therefore, each layer is cognizant only of its own local state. As the game unfolds, variabilities at upstream sites are inevitably amplified compared to those that are downstream, thus capturing the bullwhip effect in action and illustrating that it is conditions within, rather than outside, the system that reduce supply chain efficiencies.

Experimental and theoretical studies have proposed four factors as major contributors to the bullwhip effect (Lee et al., 1997): (1) order-batching, where the company accumulates or batches demand prior to placing an order instead of ordering frequently, leading to spikes and lulls in demand; (2) demand forecast updates, where each company in the supply chain bases its future demand pattern, material requirements, production scheduling, and inventory on its own customer order history, so each upstream company adjusts its own demand forecasts; (3) price variations (or promotions) that result in the placement of large orders that do not reflect true customer demand; and (4) product rationing, which causes customers to exaggerate their demand for rationed products in times of short supply, resulting in a flurry of cancellations and lacklustre demand when products are back in ample supply. While improvements in operational inefficiencies, poor chain design, and channel misalignment can mitigate several of these factors, our focus is on how enhanced information-sharing can improve supply chain performance.

2.3 Current Approaches for Information-Sharing

Because of increasing competition in today's global markets and the bullwhip effect's outsized effects on supply chains, the bullwhip effect has garnered considerable attention from academics, economists, and policymakers who seek to diminish its prevalence and reduce costs for supply chains across industries (Cachon & Fisher, 2000).

Traditional means of dampening fluctuations along supply chains have had varying degrees of success. In production-smoothing, one such approach, the production rate is maintained at a relatively constant level throughout the year, leading to inventory accumulation during times of low demand in anticipation of periods of high demand. However, empirical studies have found that production at firms that have excess inventory was more variable than sales were (Blanchard, 1983). The vendor-managed inventory (VMI) approach, which is based on an agreement between a vendor and a buyer for the vendor to manage the inventory and initiate orders for the buyer, has also been attempted. In such a setup, the vendor has more visibility of the product's demand than the buyer does. Made popular in the 1980s and now widely employed, VMI has been the topic of empirical studies based on item-level data that have shown that VMI benefits downstream firms by reducing inventory and stock-outs, while upstream firms benefit from reduced bullwhip effects. Still, their principal limitation is loss of control for retailers, so it is suitable only for certain kinds of supply chains. Individualized approaches have also been proposed and tested. For example, the heavy-equipment manufacturer Caterpillar Inc. designed customized solutions to counteract inventory volatility (Bray & Mendelson, 2012b) and mitigate the bullwhip effect by waging a multi-pronged attack on its components. In the earlier days of the internet, Caterpillar was among the first companies to undertake high-speed sharing of sales data between its product-design department and its suppliers (Songini, 2000). The company also mitigated bullwhip effects that resulted from short lead time by fixing orders such that an order remained unchanged for 3 months after being placed (Aeppel, 2010). While these efforts proved effective (Katz, 2011), Caterpillar's approaches are not readily generalizable to all other industries, especially not those that manage non-physical flows of materials. Given the heterogeneity of the various markets that are plagued by the bullwhip effect, a standardized and effective approach is much needed but has yet to be fully envisioned, much less implemented.

Theoretical and experimental studies have shown that timely and undistorted information-sharing can enhance integrated supply chains' performance (Yu et al., 2002). One means of counteracting the bullwhip effect that results from product forecasting is to make downstream data available to upstream sites, thus enabling both to update their forecasts using the same data (Steckel et al., 2004). Forecasting is critical for mitigation of risk and uncertainty, yet management decisions are often made in a climate that is characterized by insufficient or undependable data. Artificial neural networks (ANN) are currently employed to solve logistic problems like optimizing transport routes but can also be leveraged to generate accurate demand forecasts. ANN's ability to learn from abundant data enables supply chain networks' behaviour to be modelled, making forecasting network behaviours and testing the effects of network manipulation and modifications possible. One limitation to this approach is that it requires finding and extensively filtering data, as the dataset has a significant effect on the solution. Behaviours that underlie volatility across supply chains, such as decision bias and over-reactions that contribute to operational complications in supply chains, are also dampened through information-sharing initiatives that enable upstream chain participants to anticipate demand fluctuations (Croson & Donohue, 2006).

With the advent of secure and robust technologies and platforms for information-sharing, rapid and secure information-sharing has become increasingly viable, driving firms to explore new avenues for cooperation across industries and geographically separated operations (Funda & Robinson, 2002). Electronic data interchange (EDI) systems and more recent SaaS-based systems are two examples that have been shown to facilitate and promote greater transparency while standardizing the attributes of supply-and-demand data between manufacturers and customers (Machuca & Barajas, 2004). Similarly, enterprise logistics software (e.g., SAP) has enabled companies to maintain and disseminate information to supply chain participants on a common database, a feature that has been thought to facilitate significant cost savings.

The use of EDI and SaaS-based approaches has undoubtedly facilitated information-sharing; however, insufficient data security, cumbersome data mobility, issues with software integration and cross compatibility, and high costs have prevented these approaches' wide-scale adoption. Furthermore, despite the benefits of information-sharing, supply chain participants remain reluctant to cooperate and lack incentive to share their data fully in an increasingly competitive global market (Baihaqi & Sohal, 2012). As a result, the need for technologies that enable secure, reliable, anonymous information-sharing and coordination and that align the incentives of supply chain participants remains unfilled.

3 Technology Innovations

Two key limitations to data-sharing among supply chain participants pertain to the requirement for centralization of data-interaction processes and the reluctance

Technology	Distributed Ledger Technology (DLT)	Fully Homomorphic Encryption (FHE)
Summary	*DLT represents a varied, adaptive group of consensus-based technologies with unique features and benefits. It is ideally suited to performing a range of capital markets functions.*	*A new form of 'secret' computing will allow parties to analyze and monetize data while still encrypted, protecting its value and remaining compliant with data protection rules.*
Enablers	• Tokenization (i.e., digital representation) • Identify Verification (e.g., history)	• Data Privacy (e.g., for analytics) • Data Sharing (e.g., for API services)
Awareness	*Very high*	*Low*
Maturity	*Medium*	*Low*
Potential	*High*	*High*
Applications	• Cross-border Payments • Trade / Supply Chain Finance • Stable Coins	• Data Exchanges (*e.g., among network nodes*) • Client Analytics (*i.e., as a product*)

Fig. 1 Overview of DLT and FHE

of participants to share their data. We introduce two key technology innovations that can address these challenges: DLT and FHE. The key tenets of these two technologies are summarized in Fig. 1.

We propose that DLT can bridge the tear in the data-integration fabric by providing a way to facilitate information-sharing without a centralized party, making DLT-based solutions prime candidates for enhanced sharing of select information in the supply chain network, and guaranteeing its anonymization. Furthermore, newly emerging FHE schemes have the potential to be transformational in providing the confidence that supply chain participants need to engage in processes that require select information-sharing across the board (e.g., summary statistics at various granularities) by providing verification that their identifiable raw inputs are not passed on to any other participant or actor.

3.1 Decentralized Information-Sharing Using DLT

Blockchain is an internet-based technology that is prized for its ability to validate, record, and distribute transactions publicly in immutable, encrypted ledgers. To illustrate the workings of this technology, consider its role in supporting transactions in Bitcoin, a digital cryptocurrency that operates independent of a central bank (Hellwig et al., 2020). Blockchain technology provides the platform for creating and distributing the ledger, or record, of every Bitcoin transaction to thousands, if not millions, of computers that are linked to networks in all parts of the world.

Because the transactions and ledgers are encrypted, blockchain technology offers more security than the banking model, and its instantaneous transmission via the internet eliminates banks' two- to three-day clearing process and accompanying costs for transferring money from one account to another. The term "blockchain" is derived from the "blocks" of validated and immutable transactions and how they link in chronological order to form a chain. Although the technology was first invented to support transactions in Bitcoin, the many other applications include cross-border transfers, tokenization, asset tracking, and commodity trading. While most of today's supply chains operate at scale, without blockchain technology, the technology has inspired diverse research projects and prompted established IT players and start-ups to initiate promising pilot projects. These projects have included large US-based retailers like Walmart's proof-of-concept (POC) test runs to trace pork in China and produce in the US and to authenticate commercial transactions and monitor the accuracy and efficiency of record-keeping (Kamath, 2018), Maersk and IBM's projects to facilitate cross-border transactions (Miller, 2019), and BHP's project to replace internal and external tracking of goods (Hamilton, 2016).

DLT is uniquely positioned to counteract distortion and increase the transparency of information flows across supply chains, as it enables selective information-sharing, zero-knowledge information verification, and autonomous rule-driven interactions. Examples of selective information-sharing include directional averaged indicators across participating parties and temporally aggregated inventory statistics (e.g., data on average demand at the retail level during the current period compared to the previous one that mask participants' unique identities). Zero-knowledge proofs enable all stakeholders to verify information without requiring that the actual information be shared. For example, consider the birthday paradox problem, which requires determining whether any two individuals in a group share the same birthday (Chazelle, 2007). This condition can be verified in a "zero-knowledge" fashion by hashing the birth date of each individual and comparing the hashes of the birth dates, rather than the actual dates in plaintext. Finally, executions of smart contract-enabled autonomous orders can establish a marketplace-like platform while enabling new forms of participant interactions. A smart contract is a computer protocol that can digitally facilitate, verify, or enforce the negotiation of a contract without third-party involvement. DLT-based solutions are generalizable and allow for selective, secure, anonymous, and automatic information-sharing, thereby addressing the key concerns of supply chain participants that are unwilling to share information with other members.

While the initial use case for DLT was the distributed, irreversible recordings of transactions that are tamper-proof and permit ownership to be tracked, the technology has since become the fabric for trust-supporting data-sharing platforms. Users of a blockchain-based ledger can easily reconstruct when a change to the ledger occurred, what information was modified, and where in the network the change originated. Users can also eliminate paper documents and set up an efficient digital infrastructure to record asset ownership and transfer information with verifiable origins. However, a key problem that remains is data privacy; as blockchains are transparent platforms by design, how to protect data is a critical consideration in the blockchain industry, with FHE emerging as the most promising solution.

As is often the case with potentially disruptive innovations, adoption proceeds slowly as actors become familiar with the technology, determine how to harness its potential, and integrate it into existing infrastructure ecosystems. Blockchain applications beyond cryptocurrency have yet to reach the mass market, but there are signs that a new phase in the adoption cycle has begun, not least by technology industry leaders. Tokenization and asset-tracking are two examples of DLT applications that illustrate the mechanisms that are central to the proposed information-sharing approach and framework (section "Looking Ahead: DLT Meets FHE"). These mechanisms enable two key features of DLT, control and transparency, that make DLT attractive to those who seek to encourage and facilitate information-sharing, especially in structured environments like supply chains.

3.1.1 Tokenization

Tokenization describes the process of converting rights to an asset into a digital representation. Consider an asset, such as real estate, that is worth \$1000,000. Tokenization can digitally represent and transform it into 1000 tokens (an arbitrary number), such that each token represents a 0.1% share of the asset. The tokens are subsequently issued on a platform like Ethereum that supports smart contracts, thus providing a mechanism for the tokens to be traded on various exchanges. One area that already leverages blockchain-enabled tokenization is the art industry. Art transactions today are complex and often delicate, as both buyers and sellers may want to remain anonymous (e.g., because some art may have questionable origins, and/or owners may not want their ownership of a valuable asset—or their ability to buy it—to be public), and the availability of a buyer's funds is difficult to validate *a priori*.

Established auction houses' ventures into the digital world have been largely unsuccessful, but the art market is now gradually leveraging blockchain. During Christies' of New York's sale of the Barney A. Ebsworth Collection in November 2018, potential buyers were able to access the blockchain-based system prior to attending the auction event to view the entire provenance and transaction history of the works on offer. However, especially relevant to the art market is that blockchain makes transactions more transparent. For example, using blockchain, the auction house can validate whether the buyer has the necessary funds. What's more, the entire transaction sequence, including the ownership transfer, can, in theory, be executed via a smart contract, which can be used to determine whether the seller really owns the art, as the digital contract cannot be altered, so the chain of sales transactions that have taken place can be verified. The contracts are also accessible to all, thus facilitating the traceability of provenances. Finally, blockchain also gives artists a way to prove that they are the true creator of their work, thereby curtailing incidences of counterfeit. However, an increase in transactional transparency also entails a loss of anonymity, so the extent to which this technology will be incorporated into the art market will depend on whether it can comply with ever-stricter money-laundering regulations while granting the degree of anonymity buyers and sellers demand.

3.1.2 Asset-Tracking

Traceability of provenances can be applied to other goods, such as commodities. Consider the Smartcane Best Management Practice (SBMP) project, an initiative run by the Queensland Cane Growers Association that seeks to improve the traceability of sugar products from farms to factories to suppliers to retailers. Using blockchain, the agency's goal is to ensure that Australia's sugar supply is traceable and understandable so consumers know where their sugar comes from and Australia's sugar products can be distinguished from imports. Here, a blockchain-enabled system ensures a database is protected from tampering by unscrupulous parties who are likely to benefit from a lack of consumer-driven inventions or transparency. Blockchain also allows Australia's sugar industry to track the global market's reaction to Australian imported sugar with an eye to beating competitors. Finally, retail consumers can use the blockchain to track products' (e.g., sugar packets) individual codes and to uncover data like when the sugar was made and where it came from (e.g., fair-trade certificates). While experts expect blockchain to increase the cost of sugar production, the Queensland government and farmers predict that the difference will be offset by the expected increase in demand when the blockchain becomes operational.

3.2 Oblivious Information-Sharing Using FHE

Encryption preserves the privacy of information. With the advent of cheap cloud computing and storage, traditional encryption methods have become exceptionally fast, allowing data to be stored conveniently in encrypted form. However, an inherent limitation of most encryption techniques is the inability of information systems that work with encrypted data to perform operations prior to decrypting the data, delaying operations while data is decrypted before being used and raising security concerns; at most, encrypted data can be stored or retrieved. FHE addresses these challenges by allowing data to be processed and used without prior decryption. As such, FHE holds significant promise for industries like health care and financial services that are characterized by heavy regulation and data security.

Rivest, Adleman, and Dertouzos proposed the concept of an FHE scheme, a privacy homomorphism, in 1978, a conceptual feat that came only 2 years after the Diffie-Hellman public-key-exchange algorithm was proposed for secure exchange of cryptographic keys through a public communications channel, and only shortly following Rivest, Adelman, and Shamir's multiplicative FHE scheme, RSA (Rivest

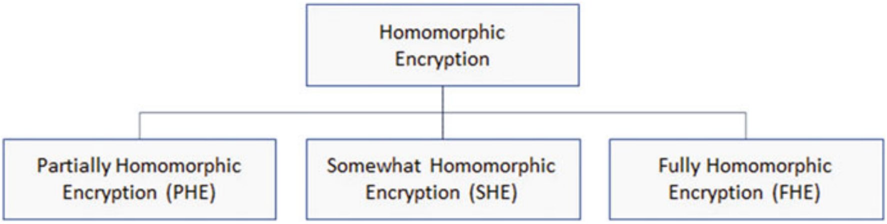

Fig. 2 Types of homomorphic encryption schemes

et al., 1978b). Like other forms of encryption, FHE uses a public key to encrypt the data, but unlike other forms of encryption, it allows functions to be performed on still-encrypted data, so it produces results that are encrypted as well.

When the problem of constructing an FHE scheme was proposed in 1978, academia and industry embarked on a search for a solution, but it was not until 2009 that Craig Gentry, a PhD candidate at Stanford, presented the first viable construction of an FHE scheme in the ground-breaking paper, "Fully Homomorphic Encryption Using Ideal Lattices". Gentry's construction was based on an encryption scheme that could evaluate low-degree polynomials homomorphically—that is, a somewhat homomorphic encryption (SWHE)—by leveraging lattice-based cryptography (Gentry, 2009). A limitation of Gentry's original construction is that the scheme is restricted to low-degree polynomials because noise grows with polynomial degrees, which compromises both security and efficiency. In 2011, Brakerski et al. (2011) (BGV) proposed an improved FHE-levelled scheme in which noise growth is limited, growing logarithmically, rather than linearly in the degree of the evaluated function and resulting in bootstrappable schemes that can be made fully homomorphic. In 2013, Brakerski et al. (Gentry et al., 2013) proposed a new, levelled FHE scheme that is based on approximate eigenvectors of matrices. These techniques improve the efficiency of the new schemes (commonly referred to as second-generation FHE) while basing security on more standard hardness assumptions like learning with errors (LWE).

The three types of homomorphic encryption—partial, somewhat, and fully—are differentiated as shown in Fig. 2. The main limitation of partial homomorphic encryption is that addition and multiplication are the only mathematical operations that can be executed on the data. With SWHE, the limitations pertain to the number of times that operations can be repeated, while FHE has no limitations in regard to either mathematical operations or subsequent re-evaluations of homomorphically obtained results.

The complexity of the computations and the significant computing power required have been stumbling blocks in FHE's evolution from theory to application. Industry has made numerous algorithmic and hardware advancements in attempts to improve the efficiency of and make more practical the BGV and the GSW schemes. The differences in process speed depend heavily on the type of data queried and processed and on the amount of data in a given process; however, current target-state estimates are approximately 50:1 computing penalty and approximately 20:1 memory penalty compared to traditional computation methodologies.

3.2.1 FHE Applications

FHE is widely expected to advance to where it is fast enough to be useful; when that happens, applications like oblivious processing, secure delegated computing, and oblivious set intersections may benefit most. For the purposes of this chapter, and with a focus on applications of FHE in the realm of supply chain information-sharing, we consider two applications of the FHE-enabled approach to secure delegated computing: oblivious processing of medical records and set intersections for attribution of retail advertisements.

3.2.2 Oblivious Processing

In a private database query system, an oblivious query allows a client to search the database and obtain results without learning anything else about the database and without the server's learning about the query (Boneh et al., 2013). Applications in which organizations cannot share plaintext data with each other include business competition, privacy regulations, and liability concerns, yet collaboration among competing organizations is necessary and could lead to new applications (Dave et al., 2020). Secure delegated computing, perhaps the most general application of FHE, pertains to third parties' processing of private data—that is, the ability for a third party to compute secret inputs using FHE schemes (Fig. 3). The motivation for this process is not new: When a single computer's computation power does not suffice, resources must be shared to manipulate and manage data through clouds. However, delegating computations or storing data with a third party (i.e., the cloud provider) risks revealing the data during computations, a risk that can be addressed by carrying out the computations without first decrypting the data (Gupta & Sharma, 2013).

Under the cloud-computing architecture, information is permanently stored in servers on the internet and cached temporarily on clients' computers (Hewitt, 2008); both users and service providers as well as data owners and custodians are separated, so the data owners do not have full control over their own data. The cloud service provider can access the data that is in the cloud at any time and may also share private information with third parties without the data owner's permission, which inevitably causes some new privacy issues (Wang et al., 2018). For example, depending on a given jurisdiction, governments may require cloud service providers to share data as part of criminal investigations.

Another process that is enabled through FHE provides a solution to the "sysadmin problem": If computations are performed on a system that is managed by a third party, the root-privileged operators at the third party have access to the data. Encryption of data that is stored in the database while on one of the servers' hard

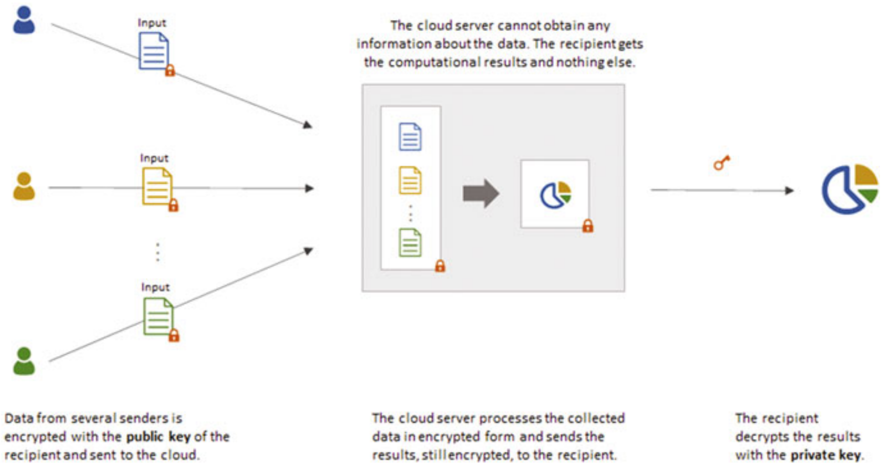

Fig. 3 Secure delegated computing (illustration)

drives prevents access to the data outside the scope of whatever computation is being done at that moment. However, with root privileges, a system operator can scan or alter the contents of RAM to gain access to whatever data is being used. With FHE, those calculations can be performed without the actual (decrypted) data's being revealed to the remote system or to its operators, which would solve the sysadmin problem.

One obvious application of secure delegated computing is in the processing of medical records. Conventional encryption methods provide restricted or even no access to encrypted data without decrypting it first. Precision medicine, a medical model that proposes the customization of healthcare, stands to benefit significantly from applications of FHE technologies. For example, sharing genomics data can facilitate precision medicine by uncovering the nature and significance of genetic variants that may underlie disorders or disabilities, leading to improved treatments and quality of life; however, genotypic and phenotypic data must be shared and analysed for variant frequency across thousands of institutes and clinics worldwide, and sharing data from advances in RNA and DNA sequencing while preserving patients' privacy remains a limitation in the field of genomics. FHE provides a tool for handling such computations without decrypting the data or having the decryption key (Bos et al., 2014); in addition, genomic data lends itself well to FHE, as it relies on relatively simple operations on the data (Hewitt, 2008).

3.2.3 Oblivious Set Intersections

Oblivious (or abstracted) set intersections allow two or more parties to discover the intersection set of their separately held datasets without either party's revealing

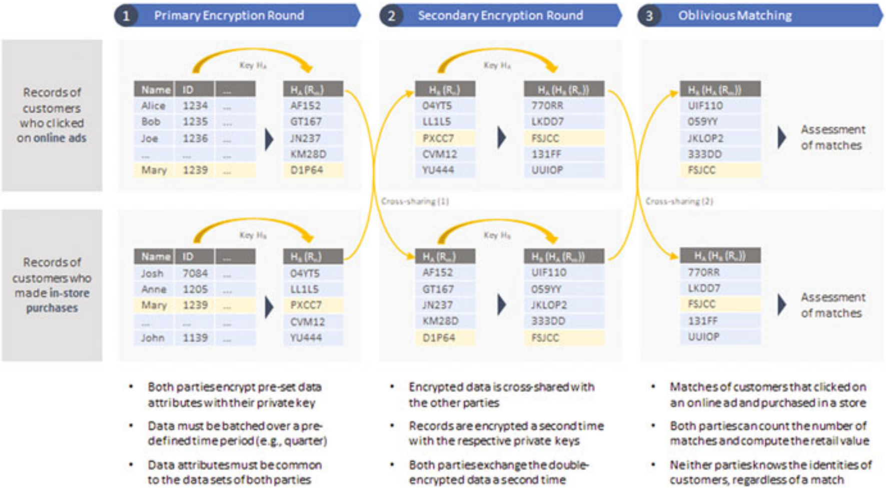

Fig. 4 FHE-enabled oblivious set intersection (illustration)

the contents of their data. One of the most transformative uses of this process was provided by Google and Mastercard in 2018 (Fig. 4). Using oblivious set intersection enabled by an FHE scheme, certain Google advertisers could track whether the ads they ran online affected retail sales at physical stores in the United States (Bergen & Surane, 2018). This method is not without faults, as customers often make purchases for reasons other than the advertisements that bombard them. However, given the magnitude of available data, statistically significant samples are obtainable, and FHE may make some degree of valid attribution between online ad exposure and retail purchase behaviour possible for the first time without the need for browser cookies.

4 Looking Ahead: DLT Meets FHE

When combined, DLT and FHE have the potential to address comprehensively some of the data-sharing limitations in the supply chain ecosystem. DLT can unambiguously validate data's ownership and origin (e.g., where inventory-level data come from), while FHE can be leveraged to ensure decentralized zero-knowledge-based computation of attributes (e.g., averages and means).

As an illustration, consider a group of three people who want to determine their average salary without sharing their individual salaries with each other or involving a trusted third party that would act as an intermediary (Fig. 5). In this two-round, multi-party circular calculation round, each of the participants (Alice, Bob, and Chris) splits his or her salary into three additive pieces and shares two of the pieces, one with each of the others. Then each party calculates an average value based on the

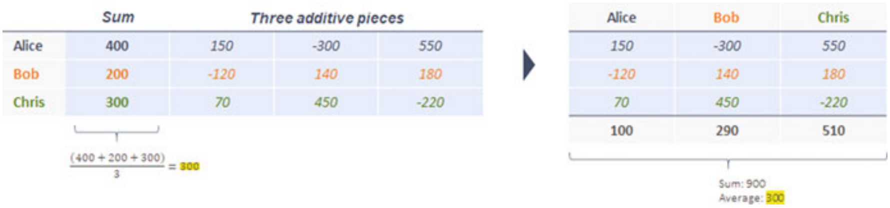

Fig. 5 Example of a zero-knowledge calculation

three inputs they hold, including their own. After the new intermediate values are calculated, the participants exchange them freely, and each party can individually derive the overall average, without having learned any information regarding the salaries of the other parties, other than whether their average salary is higher or lower than their own.

This section presents a conceptual end-to-end model for applications in supply chains. It provides a framework for zero-trust, anonymous sharing of summary statistics that may also be used for multi-echelon supply chains without their relying on a third party or a trusted intermediary that has access to the individual participants' data. The proposed approach allows data-sharing and facilitates smart decision steps in supply chains. As in the salary-averaging example, applications include calculations of summary statistics, determinations of the most competitive price points, and processing of geolocations (e.g., for route optimization), all without revealing any data to other supply chain participants and without the involvement of a third party.

4.1 An Integrated Framework

We propose a framework for the supply chain realm that leverages DLT for the execution of smart contracts, combined with a multi-party FHE scheme for data encryption and oblivious exchange.

Each entity in the supply chain is represented on the blockchain by means of an identity key. In the first step, the multi-party evaluation process is captured in a smart contract that also issues a public key that allows the parties to encrypt their data inputs. Then the participants share their encrypted data with the smart contract, which contains the evaluation key for executing FHE-based evaluations like average calculations for summary statistics and 'max' and 'min' functions for other attributes, such as geolocation and pricing information. The protocol for smart supply chain data exchanges consists of seven steps:

1. One of the supply chain parties originates a proposal for a multi-party smart contract.

2. The smart contract is launched and subsequently approved by all participating parties.
3. Once active, the smart contract shares one public key for data encryption with all parties.
4. The parties encrypt their data and send it to the smart contract for processing.
5. The smart contract calculates average levels within the trusted computing environment.
6. Encrypted results are returned to all participants once all parties have provided input.
7. Parties use the evaluation key to decrypt the average (or other) summary statistics.

The proposed protocol is expected to enable decentralized, oblivious, multi-party data exchanges for the purpose of improving supply chain transparency. However, the concrete implementation for a real-world POC, including the specification of the most suitable DLT and encryption environments and the proposed key strength, are beyond the scope of this introductory chapter. Nevertheless, capabilities that are available today are sufficient to demonstrate a POC using real-world supply chain data. As a next step, the original beer game using both DLT and FHE capabilities can be employed to quantify the efficiency gains and behavioural implications of a decentralized and trustless environment for the purpose of information-sharing within multi-echelon supply chain networks.

4.2 Limitations

Here we consider the challenges associated with driving the real-world implementations of DLT and FHE, and what is involved in moving these technologies from the theoretical to the applied space—that is, determining whether their application can be sufficiently fast and sufficiently amenable to integration with existing frameworks. While the technology is ready for POC implementations, certain technological limitations with regard to the capacity and scalability of DLT and FHE pose practical limitations for large-scale implementations in the realm of supply chain operations.

With the emergence and growth of permissionless, blockchain-based computer environments like Ethereum and the widespread availability of FHE software development kits (SDKs), the basic pre-requisites for real-world adoption are in place. Scalability and performance are a key consideration for DLT networks when it comes to applications like supply chains, as the ability to handle high transaction throughput and transaction peaks is a decisive criterion. The scalability of DLT solutions depends on the consensus mechanism chosen, which may result in DLT solutions that require far more data storage, data instructions, and time to complete a single transaction than traditional information systems do. The Bitcoin network processes around 350,000 transactions at peak times worldwide each day and is largely used against the background of current specifications. However, for widespread adoption of the current Bitcoin network infrastructure by supply chain networks, this volume is too low.

Similarly, FHE strategies are inefficient in the sense that calculations are significantly slower than calculations without encryption. Using FHE always comes at a cost, as all operations follow their own paradigms. Consider oblivious database queries: When a regular database receives a query, the underlying engine does not perform a full-text search on all contents, as tables are indexed to accelerate most operations; however, when a search is run using an FHE-encrypted value, the full text of the encrypted query is compared to every row in the relevant tables so a result can be retrieved without sharing any context with the database operator, which is a much more cumbersome process. While algorithmic development remains the critical bottleneck, advances in computing capabilities will make these computation processes faster.

For the proposed framework, while a smart contract can generate a public key, questions about the origination of this smart contract remain. Multi-sig confirmation of a smart contract origination transaction—that is, a digital signature scheme that allows participating supply chain entities to review and confirm the smart contract code before its launch—is possible, but its practicality must be evaluated in a real-world setting to determine participants' comfort with a digital-signature-based solution.

The capacity constraints for public permissionless DLT networks and the performance penalties for FHE are significant, but with recent advances, they are now well below the required threshold for usefulness. IBM has estimated that the minimum efficiency for FHE to be useful in the real world would have to be on the order of 1000:1 compared to traditional encryption schemes. With penalties now well under 100:1, real-world applications are feasible, and POCs are already underway, primarily for analytical operations in the realm of financial services.

References

Aeppel, T. (2010). 'Bullwhip' hits firms as growth snaps back. *The Wall Street Journal*. https://www.wsj.com/articles/SB10001424052748704509704575019392199662672

Baihaqi, I., & Sohal, A. S. (2012). The impact of information sharing in supply chains on organisational performance: An empirical study. *Production Planning & Control, 24*(8–9), 743–758. https://doi.org/10.1080/09537287.2012.666865

Bergen, M., & Surane, J. (2018). *Google and Mastercard cut a secret ad deal to track retail sales*. https://www.bloomberg.com/news/articles/2018-08-30/google-and-mastercard-cut-a-secret-ad-deal-to-track-retail-sales

Blanchard, O. (1983). The production and inventory behavior of the American automobile industry. *The Journal of Political Economy, 91*(3), 365–400.

Boneh, D., Gentry, C., Halevi, S., Wang, F., & Wu, D. J. (2013). Private database queries using somewhat homomorphic encryption. In M. Jacobson, M. Locasto, P. Mohassel, & R. Safavi-Naini (Eds.), *Applied Cryptography and Network Security. ACNS 2013* (Lecture notes in computer science, Vol. 7954). Springer. https://doi.org/10.1007/978-3-642-38980-1_7

Bos, J. W., Lauter, K., & Naehrig, M. (2014). Private predictive analysis on encrypted medical data. *Journal of Biomedical Informatics, 50*, 234–243. https://doi.org/10.1016/j.jbi.2014.04.003

Brakerski, Z., Gentry, C., & Vaikuntanathan, V. (2011). Fully Homomorphic encryption without bootstrapping. *ACM Transactions on Computation Theory (TOCT), 6*(3), 1–36.

Bray, R., & Mendelson, H. (2012a). Information transmission and the bullwhip effect: An empirical investigation. *Management Science, 58*(5), 860–875.

Bray, L., & Mendelson, H. (2012b). Information transmission and the bullwhip effect: An empirical investigation. *Management Science, 58*(5), 860–875. https://doi.org/10.1287/mnsc.1110.1467

Cachon, G. P., & Fisher, M. (2000). Supply chain inventory management and the value of shared information. *Management Science, 46*(8), 1032–1048.

Chaharsooghi, S. K., Heydari, J., & Zegordi, S. H. (2008). A reinforcement learning model for supply chain ordering management: An application to the beer game. *Decision Support Systems, 45*(4), 949–959. https://doi.org/10.1016/j.dss.2008.03.007

Chazelle, B. (2007). The security of knowing nothing. *Nature, 446*, 992–993. https://doi.org/10.1038/446992a

Croson, R., & Donohue, K. (2006). Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science, 52*(3), 323–336. https://doi.org/10.1287/mnsc.1050.0436

Dave, A., Leung, C., Popa, R. A., Gonzalez, J. E., & Stoica, I. (2020). Oblivious coopetitive analytics using hardware enclaves. *Proceedings of the Fifteenth European Conference on Computer Systems*. https://doi.org/10.1145/3342195.3387552

Forrester, W. (1958). Industrial dynamics: A major breakthrough for decision makers. *Harvard Business Review, 36*, 37–66.

Funda, S., & Robinson, E. P. (2002). Flow coordination and information sharing in supply chains: Review, implications, and directions for future research. *Decision Sciences, 33*, 505–536.

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. *Symposium on the Theory of Computing (STOC)*, 169–178.

Gentry, C., Sahai, A., & Waters, B. (2013). Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. In *Advances in cryptology – CRYPTO 2013* (Lecture notes in computer science, pp. 75–92). Springer. https://doi.org/10.1007/978-3-642-40041-4_5

Gupta, C. P., & Sharma, I. (2013). A fully homomorphic encryption scheme with symmetric keys with application to private data processing in clouds. In *2013 fourth international conference on the network of the future (NoF), Pohang* (pp. 1–4). https://doi.org/10.1109/NOF.2013.6724526

Hamilton, A. (2016). Bhp to use blockchain for a different kind of mining. *IBS Intelligence*. https://ibsintelligence.com/ibs-journal/ibs-news/bhp-to-use-blockchain-for-a-different-kind-of-mining

Haren, P., & Simchi-Levi, D. (2020). *How coronavirus could impact the global supply chain by mid-March*. https://hbr.org/2020/02/how-coronavirus-could-impact-the-global-supply-chain-by-mid-march

Hellwig, D. P., Karlic, G., & Huchzermeier, A. (2020). *Build your own blockchain: A practical guide to distributed ledger technology*. Springer Nature.

Hewitt, C. (2008). *ORGs for scalable, robust, privacy-friendly client cloud computing*. IEEE Educational Activities Department.

https://www.us.jll.com/en/trends-and-insights/investor/too-many-eggs-covid-19-turns-focus-on-the-bullwhip-effect

Isaksson, O., & Seifert, R. (2016). Quantifying the bullwhip effect using two-echelon data: A cross-industry empirical investigation. *International Journal of Production Economics, 171*, 311–320.

Kamath, R. (2018). Food traceability on blockchain: Walmart's pork and Mango pilots with IBM. *The Journal of the British Blockchain Association, 1*(1), 1–12. https://doi.org/10.31585/jbba-1-1-(10)

Katz, J. (2011). Creating high-value supply chains. *Industry Week*. https://www.industryweek.com/leadership/companies-executives/article/21944979/creating-highvalue-supply-chains

Kluger, J. (2020). *In the wake of the coronavirus, here's why Americans are hoarding toilet paper*. September 22, 2020, from https://time.com/5803273/hoarding-toilet-paper

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*(4), 546–558. https://doi.org/10.1287/mnsc.43.4.546

Machuca, J. A., & Barajas, R. P. (2004). The impact of electronic data interchange on reducing bullwhip effect and supply chain inventory costs. *Transportation Research Part E: Logistics and Transportation Review, 40*, 209–228. https://doi.org/10.1016/j.tre.2003.08.001

Miller, R. (2019). IBM-Maersk blockchain shipping consortium expands to include other major shipping companies. *TechCrunch*. https://techcrunch.com/2019/05/28/ibm-maersk-blockchain-shipping-consortium-expands-to-include-other-major-shipping-companies

Rivest, R. L., Adleman, L., & Dertouzos, M. L. (1978a). On data banks and privacy homomorphisms. In *Foundations of secure computation* (pp. 169–180). Academia Press.

Rivest, R. L., Shamir, A., & Adleman, L. (1978b). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM, 21*(2), 120–126. https://doi.org/10.21236/ada606588

Sethuraman, K., & Tirupati, D. (2004). Evidence of bullwhip effect in healthcare sector: Causes, consequences, and cures. *International Journal of Services and Operations Management, 1*(4), 372. https://doi.org/10.1504/ijsom.2005.007499

Songini, M. L. (2000). Operations via the Web. *Computer World*. https://www.computerworld.com/article/2588777/caterpillar-moves-to-revamp-supply-chain-operations-via-the-web.html

Steckel, J. H., Gupta, S., & Banerji, A. (2004). Supply chain decision making: Will shorter cycle times and shared point-of-sale information necessarily help? *Management Science, 50*(4), 458–464. https://doi.org/10.1287/mnsc.1030.0169

Tate, L., & Ellram, L. (2019). *Supply chain finance: Risk management, resilience and supplier management*. Kogan Page.

Taylor, D. A., & Brunt, D. (2010). *Manufacturing operations and supply chain management: The lean approach*. Cengage Learning.

Wang, Q., Zhou, D., & Li, Y. (2018). Secure outsourced calculations with homomorphic encryption. *Advanced Computing: An International Journal, 9*, 1–14. https://doi.org/10.5121/acij.2018.9601

Yu, Z., Yan, H., & Cheng, T. C. (2002). Modelling the benefits of information sharing-based partnerships in a two-level supply chain. *Journal of the Operational Research Society, 53*(4), 436–446. https://doi.org/10.1057/palgrave.jors.2601255

Tutorial on Blockchain Applications in Supply Chains

Volodymyr Babich and Gilles Hilary

1 Introduction: Supply Chain Management Challenges

Supply chains comprise firms, organizations, and individuals who are independent and self-interested but are linked through physical, informational, and financial flows. Their activities enable products and services to be created and consumed. Consider, for instance, supply chains for personal protective equipment (PPE), as illustrated in Fig. 1.

The final PPE consumers are doctors and nurses, lab technicians and dentists, as well as individual consumers. Doctors and nurses and other medical or research professionals obtain PPE through their organizations, such as hospitals and medical offices. Individual consumers order PPE through retailers, such as Amazon, Walmart, and CVS. These retailers and hospitals receive PPE from distributors, e.g., a company such as McKesson. Because the packaging and volume of PPE that goes through regular retailers and to commercial uses differ, distributors may specialize in either consumer or commercial PPE or they may supply both. Distributors receive PPE from manufacturers (e.g., a company such as 3M). Even though the final PPE consumers may be located in one country (e.g., the USA), manufacturers have production plants around the world. Manufacturers procure various components from different contractors. A large percentage of PPE production capacity is located in China. The physical flow of PPE is enabled by logistics providers, such as ocean shippers, trucking companies, airlines, and rail companies. Warehouse operators and port and customs authorities also contribute to the physical flow of PPE. Governments not only set the rules that govern trade and protect industry and consumer interests, but they also participate in PPE supply chains directly by

V. Babich (✉) · G. Hilary
McDonough School of Business, Georgetown University, Washington, DC, USA
e-mail: vob2@georgetown.edu; gilles.hilary@georgetown.edu

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_3

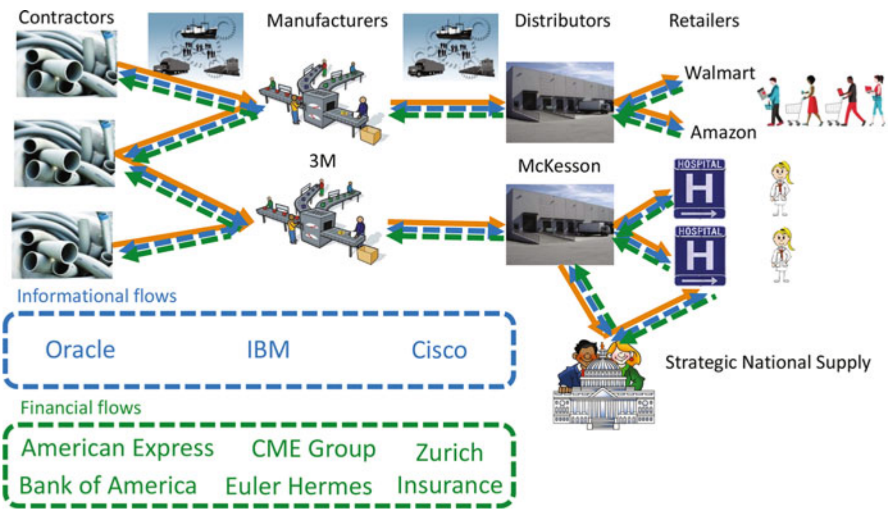

Fig. 1 Personal protective equipment (PPE) supply chains

purchasing PPE and managing strategic national reserves. In a complex procurement system, in the USA, PPE is purchased not only by the federal government but also by states, counties, and municipalities. Information technology (IT) vendors and telecom companies (e.g., Oracle, IBM, and Cisco) facilitate informational flows in PPE supply chains. Banks, transaction processors, and commodity exchanges (e.g., American Express, Bank of America, CME Group) facilitate financial flows. Insurance companies (e.g., Euler Hermes and Zurich Insurance Group) insure against disruptions to both the physical and financial flows.

An important observation from this example is that even for simple products, such as face masks, medical gowns, and surgical gloves, supply chains are global and complex. *Complexity* is the first challenge of supply chain management. Despite this complexity, PPE supply chains operate seamlessly most of the time. The majority of the final consumers rarely experience shortages and do not think about global supply chains.

Unfortunately, the second challenge of supply chain management is *uncertainty*. In 2020, the COVID-SARS-2 pandemic shocked PPE supply chains. Simultaneously, with factories in China shuttering due to lockdowns, the demand for PPE around the world spiked. Border closings, reduced flight frequencies, and labor shortages impeded the physical flow of goods. Declining economies and the increasing threat of payment defaults threatened financial flows. Misinformation, panic, the lack of response plans, and poor communication among the participants in supply chains interfered with informational flows. Many countries, including the US, experienced persistent shortages of PPE.[1]

Responding to the growing demand, declining supply, and persistent shortages, procurement teams desperately tried to secure PPE for their organizations by either inflating orders or over-ordering from multiple distributors. This made it difficult for manufacturers to plan their capacity use and make investment decisions. There was fierce competition between procurement teams at the hospital, state, and federal levels, seizures of supplies intended for other customers, and hoarding. In part, this destructive behavior was due to unreliable and missing information about production capacity and inventory. When they were unable to procure PPE from their usual suppliers, buyers turned to untested vendors. While some of the new vendors were able to deliver,[2] many failed to do so, and worse, many delivered defective and counterfeit products.[3] The selfish behavior of buyers and sellers was entirely predictable and natural, but its presence highlighted the third supply chain management challenge, i.e., the *misalignment of incentives*. The problems were exacerbated by the fourth challenge: *lack of visibility and transparency*. In fact, in most supply chains, there is little visibility of their structure; firms do not have the data to evaluate the capabilities of their supply chain partners, and there is no real-time information about the state of production processes, the location of goods, or the conditions in which these goods are in. In most supply chains, there is a lack of transparency about the chain of custody of goods, the processes used to manufacture and transport goods, and the incentives of supply chain participants.

This lack of transparency extends to financial flows in supply chains. For example, S&P Global Ratings called supply chain financing (in particular, "reverse factoring") a "sleeping risk" and argued that poor disclosures could hide the deterioration of a firm's financial situation and result in mispricing of risk or the misallocation of capital.[4] A case in point is that of Carillion. The company used to be a large British facilities management and construction services firm before its liquidation in 2018. It made extensive use of reverse factoring as a source of financing but did not disclose the exact amount. When the company's policy of pursuing low-margin government contracts to maintain revenue growth finally caught up with it, the firm was destabilized by its lopsided capital structure. Naturally, when one important player in a supply chain is adversely affected, the entire chain suffers (see Cohen and Frazzini, 2008; Hertzel et al., 2008; Kolay et al., 2016; Agca et al., 2021), and the issue goes beyond the industrial partners.

---

[1] An article in the WSJ on November 4, 2020 had the title "Face Masks Are Again in Short Supply as Covid-19 Cases Surge," https://www.wsj.com/articles/face-masks-are-again-in-short-supply-as-covid-19-cases-surge-11604499588. The first case of COVID 19 in the US was recorded on January 20, 2020, approximately 9 months earlier.

[2] For example, Ford produced N95 masks with clear panels to help the deaf population and ease communications. https://www.cbsnews.com/news/ford-is-making-n95-masks-with-clear-panels/.

[3] The US DHS seized 11 million counterfeit masks in just one sting operation, https://www.ice.gov/news/releases/dhs-prevents-millions-counterfeit-n95-masks-reaching-hospital-workers-first.

[4] Reverse Factoring: Why It Matters, S&P Global Ratings, March 10, 2020.

For example, when the large supply chain financing company Greensill filed for bankruptcy in 2021, 50,000 jobs were said to be at risk, and some of the 40 clients covered by the firm were faced with becoming insolvent.[5]

An illustration of the importance of the integration between the physical and financial supply chain flows is the example of the state of Illinois purchasing face masks.[6] Like other states and local governments, Illinois was eager to acquire PPE but was facing global scarcity and competition from other buyers. The state made a public call for help. One entrepreneur was able to create an ad hoc supply chain that included an intermediary in the US, a shipper based in Germany, and spare production capacity at various facilities in China. The make-shift supply chain could deliver the badly needed PPE but only if a payment could be made promptly. However, the state of Illinois is prohibited from making wire transfers to a foreign bank and has a long lead time on "cutting checks" because of the industrial-grade process of printing checks. Facing extremely tight time constraints and in the desperate need for PPE, the state abandoned the usual payment process (in fact, all check printing was stopped). A check for PPE was cut, physically driven to a midway rendezvous point in a parking lot between the offices of the State of Illinois and the American intermediary, and then deposited to the bank account of this intermediary, who took a picture of the deposit slip and sent the image to the factory owner in China. The rest of the supply chain worked on trust and the limited assurance provided by this picture. This was a highly unorthodox way of conducting business for a state government and a deviation from the usually much more regimented process of international trade.

The multitude of supply chain failures observed during the COVID-19 crisis is not unique to PPE supply chains or this particular crisis. Mehrotra et al. (2020) listed these problems for the COVID-19 pandemic, while Patel et al. (2017) reported the very same supply chain failures as lessons learned from the 2009 H1N1 pandemic and the 2014 Ebola epidemic. Generally, OM researchers will immediately map many of the failures in PPE supply chains to the discussion in the seminal paper by Lee et al. (1997b) on the bullwhip effect.

The champions of blockchain technology argue that blockchain can help to overcome at least two of the causes of supply chain challenges, namely, the *misalignment of incentives* and the *lack of visibility*, even if this technology is not a panacea for every ill afflicting supply chains.[7] Sceptics of blockchain technology counter that the promises related to blockchain in supply chain management are just hype. In the next section, we review recent applications of blockchain in supply chains to establish the provenance and the chain of custody, track assets, streamline

[5]https://wolfstreet.com/2021/03/04/softbank-fintech-unicorn-greensill-on-verge-of-collapse/.

[6]The following paragraph is largely based on an interview of several protagonists broadcasted on National Public Radio on April 17, 2020. https://www.npr.org/transcripts/837216447.

[7]Blockchain is an example of distributed ledger technology (DLT). In this article, we focus on blockchain as it is currently the most common form of DLT, but nearly all of our comments apply to all currently available DLT.

supply chain transactions, and facilitate supply chain finance. These examples will motivate our reflection on the following two important questions: (1) Is blockchain technology superior to other solutions, such as databases or other technologies for digitizing supply chains? (2) What are the open and interesting research questions in this domain?

In this tutorial, we build on Babich and Hilary (2020, 2019), who reviewed the very early development of blockchain and other distributed ledger technologies to operations, to address these questions. We do not describe the technology itself and rely instead on the discussion in Babich and Hilary (2020) and Hilary (2022). We revisit their early findings; more specifically, we review recent industry developments (especially in light of the early examples they provided) in Sect. 2, update the conceptual framework in Sect. 3 and discuss possible academic developments in Sect. 4. We conclude in Sect. 5.

2 Blockchain Technology and Supply Chains in 2021

Blockchain started as a technology to support cryptocurrencies but has gone beyond this use and now extends into healthcare, energy, sustainability, supply chains, identity management, finance, and other applications (e.g., see Hilary and Liu, 2021 for applications in finance). Like most new technologies, blockchain has generated initial enthusiasm verging on hype. In particular, claims have been made that supply chain management is on the verge of the “blockchain revolution.”[8] Such claims have been met with skepticism because similar claims have been made about every new technology related to supply chains (e.g., ERP, RFID, 3D printing, AI, and robotics).

Numerous pilot studies of blockchain applications in supply chains started in the mid-2010s, but mixed results have been achieved. The end of 2017 appears to have been the peak of the blockchain hype cycle. For example, according to Google Trends, the number of worldwide searches for the word “blockchain” peaked in December 2017, which coincides with a peak in Bitcoin price. From 2010 to mid-2019, the graph of the number of Google searches for “blockchain” mimicked the price of Bitcoin almost perfectly, which suggests that interest in blockchain technology was driven by Bitcoin and perhaps by Bitcoin speculation. In 2018, the research and advisor company Gartner placed blockchain on their hype cycle graph between the “peak of inflated expectations” and the “trough of disillusionment”; in 2019 and 2020, blockchain technology slid into the “trough of disillusionment.” If Gartner’s hype cycle prediction works out, the technology will mature and go through the “slope of enlightenment” and settle on the “plateau of productivity.” We believe that when that plateau is reached, the value of blockchain technology will be established along three dimensions: (1) information, (2) automation, and (3) tokenization. In the next two sections, we review recent developments in technology

[8] See, for example, Tapscott (2020) and Fenech (2018).

as they relate to supply chain management, and supply chain finance to identify areas where the technology has created value.

2.1 Applications of Blockchain Technology to Supply Chain Management (SCM)

Babich and Hilary (2020) described initial pilot applications of blockchain to SCM. Thus, it is instructive to revisit their work to assess where this practice stands two years later (since the initial article was written) and see if these initial projects were a byproduct of hype or if they created value.

We can divide the current applications of blockchain in SCM into three categories. The first one is **establishing the provenance of the goods and the chain of custody**. This application is instrumental in addressing the problem of food and medicine counterfeiting and adulteration. The WHO has reported that one in ten medical products are either substandard or counterfeit,[9] while the "2019 Food Safety Insights Survey" stressed the importance of economically motivated food adulteration.[10] Recent examples of application in this category include establishing the chain of custody for the following:

1. Fish, "ocean-to-table." Bumble Bee uses SAP blockchain to trace fish.[11]
2. Berries, "vine-to-table." SAP blockchain tracks blueberries for Naturipe.[12]
3. Coffee, "bush-to-brew." Smucker's Folgers brand started using IBM's Farmer-Connect system for coffee traceability in 2020.[13] This is a more developed application compared with the original example of Denver's Coda Coffee Co.[14]
4. Pork, chicken, and vegetables, "farm-to-fork." In this area, IBM has formed a consortium of large companies, which is dubbed the Food Trust Initiative, to develop a platform supported by Hyperledger Fabric that tracks different products. Since its inception in August 2017, the private network has expanded to more than 80 members and tracked over 1300 products in 2020.[15] Carrefour has implemented blockchain tracking for 20 products and observed increased sales. The retailer has also noticed a halo effect for other products (i.e., if consumers

[9] https://www.who.int/news-room/detail/28-11-2017-1-in-10-medical-products-in-developing-countries-is-substandard-or-falsified.

[10] http://fsns.com/news/an-update-on-food-fraud.

[11] https://youtu.be/n_XnEKJVKXo.

[12] https://www.youtube.com/watch?v=1FhbNI_9a2Y.

[13] https://www.techrepublic.com/article/ibm-blockchain-connects-folgers-drinkers-to-the-origin-of-their-brew/.

[14] https://www.bext360.com/.

[15] https://www.forbes.com/sites/biserdimitrov/2019/12/05/how-walmart-and-others-are-riding-a-Blockchain-wave-to-supply-chain-paradise/#308bbf797791.

can trust Carrefour with chicken, they feel they can trust Carrefour with cheese and milk).[16]

5. Diamonds and gemstones, "mine-to-shine." Everledger has recorded the provenance and the chain of custody of millions of diamonds on a blockchain since 2015. As of 2020, the company is advertising on their website the ability to track other gemstones, wine, art, luxury goods, e-recycling, and insured assets, although it stopped publicly sharing examples of the information contained in their blockchains.[17]
6. Airplane parts, "market-to-maintenance." Honeywell operates the GoDirect marketplace for airplane parts, where parts' authenticity and ownership are recorded on a blockchain.

The second category of blockchain applications **validates ethical and sustainable sourcing and fair-trade practices**. For this category, not just the provenance but also compliance with economic and social norms and laws must be documented. Examples in this category include Denver's Coda Coffee, which instantly pays farmers for coffee beans and records the prices paid on a blockchain.[18] Everledger stores images of Kimberley process certificates for diamonds. These certificates are intended to reduce sales of "conflict diamonds." OpenSC uses blockchain to certify that toothfish were caught in legal fishing zones. Records also contain information about the type of fuel fishing vessels and transport vessels are using, as well as the amount of carbon emitted to maintain cold storage conditions.[19] This amount is counted against the carbon-offset investments made by the company that owns the fleet. Together, this information is used to claim that the process is carbon neutral. Information on the blockchain is shared with consumers and anyone who needs access to it. Solara.io plans to record generation data from individual solar panels on the EnergyWeb blockchain and provide those records as proof of sustainable generation. In an earlier article, Babich and Hilary (2020) discussed the example of Veridium Labs and IBM who announced in 2018 that they were working on a blockchain solution to improve the operations of carbon credit markets.[20] Alas, there have been limited public updates on this project since then.[21]

The third category of applications is **improving processes, establishing asset ownership and tracking assets, and generating data for sale**. An example of this category is the TradeLens blockchain platform. It is a joint venture between Maersk and IBM. Over 150 supply chain operators (carriers, ports, terminal operators, 3PLs, freight forwarders, and shippers), accounting for almost half of the world's

---

[16] https://www.reuters.com/article/us-carrefour-Blockchain-idUSKCN1T42A5.

[17] https://www.reuters.com/article/us-carrefour-Blockchain-idUSKCN1T42A5.

[18] See https://www.bext360.com/ and https://youtu.be/Mn9dM_roD1A.

[19] https://opensc.org/product-example.

[20] https://techcrunch.com/2018/05/15/veridium-labs-teams-with-ibm-and-stellar-on-carbon-credit-Blockchain/.

[21] See https://medium.com/@robertgreenfieldiv/Blockchain-enabled-carbon-credit-markets-1a195520f0e1 for a discussion of the issues associated with blockchain-enabled carbon credit markets.

oceans freight data and at least one major bank (Standard Chartered), have joined this platform, whose goal is to streamline the international shipping processes by reducing the amount of physical paperwork that is required to move goods. The TradeLens system uses sensor data for monitoring many variables, from temperature control to container weight, and records information on a blockchain. Another example in this category is the Tianjin Port blockchain pilot. In 2015, a Tianjin container storage facility exploded, killing 173 people (the explosion registered as a 2.9 magnitude earthquake). To date, the exact composition of the chemicals that exploded is unknown. To mitigate the risk of the reoccurrence of such an event, the port authority developed a blockchain platform that will be used for the confirmation of rights, certificates of bills, trading, finance, logistics, and overall supervision. The third example in this category is Solara.io, which plans to sell data generated on solar panels, batteries, IoT devices, and local electricity usage to system operations, energy users, and traders.

2.2 Applications of Blockchain Technology to Supply Chain Finance (SCF)

Working capital is tied up in supply chains. SCF allows organizations to use their working capital more effectively and reduces financing costs by streamlining transactions among trading partners and financial institutions by, among other things, facilitating the transfer of information and verified documentation about procurement and logistics events, such as invoices, bills of lading, proof of delivery and payments. SCF allows suppliers to access financing using more favorable terms that correspond to the credit risk of the buyers. This optimizes the use of credit capacity in supply chains and lowers capital costs. Traditional modes of international trade financing include letters of credit and escrow services[22] (their relations to blockchain are discussed in Babich and Hilary, 2020), but there are many other types of trade financing (Hofmann et al., 2017). For example, if a supplier would like to receive an advance in the future payment from a buyer, a supplier can use factoring, which is a sale of an account receivable. Factoring transactions are initiated by the supplier, and the buyer may or may not be aware of them. In contrast, reverse factoring is initiated by the buyer (typically a large enterprise). Once the buyer guarantees payments to suppliers, intermediaries (e.g., financial institutions) offer financing for a fee. A supplier can elect to receive an early payment, and when the invoice becomes due, the buyer pays the bank directly. An advantage of this approach is that it centralizes some of the steps of financial transactions with one financial entity and digitizes the information transferred. Two-

[22]Letters of credit guarantee that the buyer of goods or services has the financial resources to pay. With letters of credit and escrow accounts, a third party (e.g., banks) holds the money until sale conditions have been met.

thirds of European companies with revenues larger than $750 million run a supply chain finance program.[23] Revenues derived from trade finance reached between $50 billion and $75 billion in 2019, according to the estimates of the International Chamber of Commerce.[24]

One of the challenges of these financing agreements is that they are typically negotiated bilaterally between a buyer and a seller (plus a financial institution), whereas supply chain resources are distributed across multiple firms and jurisdictions. Thus, it is difficult to access capital that is distributed across multiple supply chain tiers. Obstacles to creating supply chain-wide financing systems include the lack of standardization, the lack of visibility into pairwise transactions, and the lack of documentation sharing across transactions in multiple tiers. The seemingly simple steps of verifying documents and transactions and onboarding in a computerized system require significant time and attention.

Blockchain can potentially help in the three dimensions of process efficiency, visibility, and tokenization. All too often, trade financing processes remain at least partially manual, and integration with enterprise resource planning (ERP) systems may take months. Third-party integration is a notoriously buggy process, as data sharing across organizations relies on incompatible systems and involves extensive manual data wrangling. Although blockchain does not solve every incompatibility problem, it can improve process efficiency by facilitating the exchange of documentation in a fast and secure way and providing proof of its veracity. For example, Komgo, which is a consortium of financial, trading companies, and oil companies, claims a 99.58% reduction in time to issue a digital letter of credit (from 10 days to 1 h).[25] It is difficult to share supply chain resources if the firms in supply chains do not have visibility beyond their immediate supply chain partners. Blockchain provides visibility to extended supply chains. Finally, blockchain can create verifiable and transferable digital claims ("tokens") corresponding to both real and financial assets that can be shared across multiple supply chain tiers and even with entities outside of supply chains. Such tokens could be used as collateral for loans and be traded as bundles in the same way that credit card debts are traded.

Several organizations have started to take advantage of these possibilities. In early 2019, the Chinese company Ant Financial announced a new blockchain supply chain finance subsidiary (called Ant Duo-Chain) based on a pilot conducted in 2018. Similarly, Ping An's OneConnect launched its supply chain finance platform (called "One Enterprise Chain").[26] China Everbright Bank (CEB), a major Chinese commercial bank, joined the OneConnect platform in 2020.[27] This partnership has

[23] https://www.pwc.com.au/publications/pdf/supply-chain-finance-jul17.pdf.

[24] https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/supply-chain-finance-grows-amid-pandemic-but-faces-stark-risk-warnings-58841608.

[25] https://consensys.net/Blockchain-use-cases/finance/komgo.

[26] https://www.ledgerinsights.com/alipay-ant-double-chain-Blockchain-supply-chain/.

[27] https://cointelegraph.com/news/china-everbright-bank-uses-ant-financials-dlt-for-supply-chain-finance.

been praised by the United Nations as a way to increase fund accessibility for SMEs.[28] Marco Polo is a technology project that launched in September 2017 and is led by TradeIX and a consortium of banks to build an open account trade finance platform.[29] SME Early Pay has a novel onboarding approach that uses blockchain for identity management to address know-your-customer (KYC) frictions.[30] LinkLogis, which is backed by Tencent and Standard Chartered, offers multi-tier supply chain finance (this financing is also offered to the supplier's supplier).[31]

Babich and Hilary (2020) noted several issues that may slow down the development of blockchain solutions for SCF. For example, weak suppliers may feel that greater integration with a large buyer may allow the buyer to take advantage of them (now or in the future). Firms or customers may prefer that supply chains remain opaque, for example, to protect proprietary information or maintain gray markets. The tokenization of supply chain assets across multiple tiers of suppliers and the transfer of tokens to entities outside of the supply chain is difficult. For example, external verification of the value of corporate assets is needed before they can be used as collateral, but as the financial crisis of 2008 showed, even for standardized securities such as mortgages, collateralization makes it more difficult for market participants and regulators to understand risk exposure. The problem is more acute when assets are not standardized, as is often true for supply chain assets.

Shibuya and Babich (2021) compared a blockchain-based SCF system with traditional bank-based SCF solutions using a three-tier supply chain model. They pointed out that the key advantage of a blockchain-based system is its ability to create claims to collateral assets at higher supply chain tiers. A surprising finding in their paper is that even if firms that can provide financing to a supplier are identical in their financial health, supply chains prefer financing from the immediate customer of the supplier.

[28] https://cryptonews.com/news/un-praises-ant-financial-s-Blockchain-support-during-covid-1-6111.htm.

[29] https://www.gtreview.com/news/fintech/marco-polo-completes-its-biggest-Blockchain-trade-finance-trial/.

[30] https://www.ledgerinsights.com/mastercard-marco-polo-to-launch-Blockchain-supply-chain-finance-for-smes/.

[31] https://www.ledgerinsights.com/mastercard-marco-polo-to-launch-Blockchain-supply-chain-finance-for-smes/.

3 A 5-and-5 Framework of Strengths and Weaknesses of Blockchain Technology in Supply Chain Management Applications

In this section, we discuss the strengths and weaknesses of blockchain technology concerning supply chain management.

Table 1 summarizes these strengths and the corresponding weaknesses and is similar to the framework reported in Babich and Hilary (2020); however, the accompanying discussion has been expanded to reflect discussions with practitioners that have been undertaken since the completion of the earlier article. For strengths, blockchain technology is not revolutionary in the sense that other technologies can achieve the same results. However, blockchain technology is cheaper and more effective in some applications.[32] Conversely, other technologies may suffer from the same weaknesses as blockchain, but the problems we highlight are particularly acute for blockchain because of its design features.

3.1 Trust in Data vs "Garbage In Garbage Out" (GIGO)

The key blockchain technology promise is that once information is written in a distributed ledger, it is resistant to tempering.[33] This is a key difference between blockchain and traditional data storage technologies. Although there are ways of making traditional databases secure, blockchain improves on them through consensus and conflict resolution protocols. The ability to trust data on a blockchain underlies all blockchain technology applications in SCM and SCF. This is true whether they are in establishing the provenance and the chain of custody for fish, coffee, tea, diamonds, pork, airplane parts, and energy, or in managing resource allocation by verifying identities and keeping records of processes, in creating

Table 1 Blockchain technology strengths and weaknesses

	Blockchain strengths	Blockchain weaknesses
(1)	Trust in data	Garbage in, garbage out (GIGO)
(2)	SC visibility	Lack of privacy
(3)	Data aggregation	Lack of standardization
(4)	Process automation	Black box effect
(5)	System resiliency	Inefficiency

[32] Babich and Hilary (2020) discuss the cost structure of different technologies (See Table 3, in particular).

[33] This is known as the "Byzantine tolerance" feature. See Hilary (2022) in this book for a discussion of this point.

digital claims assets in supply chains, such as inventory or carbon offsets, or in verifying that trade transaction terms have been met.

This strength of blockchain technology is also its greatest weakness because of the "garbage in, garbage out" (GIGO) problem. Applications of blockchain technology to SCM require a reliable link between the physical state and the information recorded in the distributed ledger, and this link can be broken in many ways. For example, incorrect information about the physical state can be introduced into the distributed ledger at the point of information inception. This can be done by mistake, even when the stakes of making mistakes are high.[34] For example, maintaining land registries on a blockchain may help to improve data integrity, but the benefit is questionable if the initial land survey is inaccurate, as is the case in sub-Saharan Africa where only 1% of the land is under formal government registration (Adiaba et al., 2011).[35] Incorrect information can also be introduced intentionally by a rogue agent or by an organization with incentives to misreport information either directly or by bribing a party responsible for entering such information in a ledger. Recall that the misalignment of incentives is one of the challenges of SCM. This "state-zero" corruption of records may be less of an issue when the asset is natively digital (e.g., pollution rights, intellectual property, digital record of ownership), but it is a significant problem for physical assets. A perhaps more insidious manifestation of the GIGO problem is the fact that the state of physical assets constantly changes, while the information in the distributed ledger is not updated. Again, recall that uncertainty is one of the SCM challenges. Constantly monitoring and updating digital records in response to events in the physical world is a costly task that requires investments in monitoring technology and managerial effort, which can erode the cost advantages of using blockchain technology. To be fair, other data storage technologies are subject to the same challenges.

There are several ways of mitigating the GIGO problem. If a digital record captures a transaction among multiple participants and if all parties to this transaction certify the veracity of records, then this reduces the likelihood of both mistakes and intentional record manipulation. Visibility regarding the identities of organizations or persons who submit records to a distributed ledger may serve as a deterrent to fraud.

Using multiple sources of data and relying on automated processes and sensors also reduce opportunities for mistakes and fraud. As we discussed earlier, this is the approach taken by OpenSC,[36] which uses blockchain to certify that toothfish were caught in legal fishing zones. OpenSC analyzes data on a ship's location, movements, and speed to determine whether the ship was fishing in a protected area.

[34] To use an example not involving a blockchain, the Chilean Mint misspelled their country's name as "Chiie" on a batch of 50-peso coins in 2008 (https://www.telegraph.co.uk/news/newstopics/howaboutthat/7219088/Chilean-mint-spells-countrys-name-wrong-on-coins.html).

[35] https://medium.com/coreledger/land-registry-on-blockchain-a0da4dd25ea6.

[36] https://opensc.org/product-example.

All fish caught are tagged with RFID tags and barcode tags, whose information is recorded on a blockchain.

A potential weak point in the OpenSC process is the tags placed on the fish. How can one ensure that the asset that has been recorded on a blockchain has not been swapped for another asset (e.g., a toothfish is not substituted for a less-valuable fish or a toothfish caught elsewhere)? Companies are working on developing tags that are resistant to tempering and that use special inks. For example, Sicpa, whose main line of business is the security ink used in banknotes, is now offering secure markers embedded at a molecular level to tag oil and gas in pipelines.[37]

3.2 *Supply Chain Visibility vs Lack of Privacy*

Supply chain visibility allows supply chain participants to follow the flow of products, information, and money through the entire supply chain. Currently, for most supply chains, the ordering party may be able to monitor some aspects of operations at its tier-1 suppliers but rarely at tier-2 or beyond. Even the identities of higher-tier suppliers are mysteries. By overlaying a blockchain network on the supply chain network, there is an opportunity to learn the identities of the firms and individuals involved. Furthermore, depending on what information is collected, it is possible to learn what processes firms in higher tiers follow, verify how much they pay for inputs, monitor what information they exchange, observe operations of supply chains, and identify unnecessary delays and inefficiencies.

Supply chain visibility comes with costs. There are consumer privacy law compliance costs, and one must overcome the incentives of the firms in a supply chain to share information. For example, the European General Data Protection Regulation (GDPR) enshrines as one of the individual rights *the right to have one's data erased*. The core design principle of blockchain technology is that new information is combined with old information when blocks are added to a blockchain. This makes it costly to honor requests for records to be erased or even modified. The entire blockchain from the moment the record in question was added would need to be revalidated, which can be time, energy, and capital-consuming.

Similar to individual consumers, firms have secrets that they wish to protect. For example, despite its relative success, few shippers signed up on the TradeLens platform, which was created for their competitor, Maersk. The shippers' hesitation is understandable. What would Maersk do with information about their customers, shipment details, and pricing? It is conceivable that this information could be used by Maersk to compete against other shippers. Convincing suppliers to share their supply chains and bills of materials is often an impossible task. For example, after massive disruptions to supply chains from the 2011 Tohoku earthquake and tsunami, Toyota undertook a major campaign to map out its supply chains and learn the

[37] https://www.sicpa.com/solutions/oil-and-gas-integrity-management.

identities of higher-tier suppliers. However, the first-tier suppliers of Toyota delayed sharing the identities of their suppliers. Six months after the earthquake, Toyota reported that they still did not know the identities of more than half of their second-tier suppliers. Why was there such reluctance from the first-tier suppliers? They were concerned that Toyota could bypass them and contract directly with second-tier suppliers or that by knowing who the second-tier suppliers were, Toyota could better assess the cost structure of the first-tier suppliers and gain edges in future contract negotiations.

There are technology solutions to guarding the privacy of consumers and the private information of firms. Data can be encrypted, and only a carefully curated subset of data can be shared on a blockchain. Rules can be placed on who may access which data and how much control users have over their data being shared. However, these solutions increase the computational burden and system complexity. For instance, zero-knowledge proof verification and fully homomorphic encryption (see Hellwig and Huchzermeier, 2021 in this book) require deliberate design choices and are not easy to implement. They also add to the "black box" weakness of blockchain, as they make the records more difficult to audit. When records are encrypted, it is possible to verify that certain records are on a ledger, but it is more difficult to disprove that alternative records have not been recorded as well.[38]

3.3 Data Aggregation vs Lack of Standardization

Information on a blockchain can come from a variety of sources, e.g., firms, customers, logistics providers, financial institutions, regulators, and from a variety of media, e.g., databases, the Internet, and sensors. Information comes in multiple formats. For example, Everledger's diamond blockchain contains videos, photos, IDs, certificates, and geolocations. Information has a temporal dimension. This promotes a "Big Data" environment. Artificial intelligence (AI) tools for video and natural language processing and other analytics tools, in general, can convert these aggregate data into valuable information.

[38]For example, a ledger may contain two SHA256 hashes: d3a0fba49f3044ad1a178d89fced943f66e644fd2422ebc526fbc349156edb9cAnd ad754cb9164953095622022dfccc33e8a7affa0e39892eadc136edbef15666e6.The first corresponds to the statement "Alice predicted on 10/29/2019 that the Nationals would win the World Series" and the second corresponds to the statement "Alice predicted on 10/29/2019 that the Astros would win the World Series". Both hashes can be recorded prior to the outcome of the 2019 World Series final game. After the outcome is known, Alice can choose which of the statement she allows to be verified by Bob to prove her prophetic skills. It would be difficult for Bob to verify that another statement, with a different message, does not exist on the ledger (Alice and Bob are fictional characters often used in examples describing how cryptography works).

The current obstacle to achieving the full potential of data aggregation is the lack of standards. Blockchain is not a unique technology but rather an umbrella that encompasses a portfolio of protocols. These protocols are not yet stable, and as they age, they will become obsolete, creating legacy issues. This is a particularly acute problem for a technology developed to keep permanent records. The lack of standardization fosters technological uncertainty. If every sensor connected to a blockchain network operates on a different blockchain protocol, it is difficult to combine their readings on one platform. If Walmart, CVS, Target, Walgreens, and Costco all have proprietary blockchains that use different protocols, a supplier that sells to all five companies may have to train workers to use all five systems.

To remedy the lack of standards, industry alliances have emerged that set one of their goals to establish common protocols. For example, the Blockchain in Transport Alliance (BiTA), which was formed in 2017, is working on common protocols for the transportation industry.[39] EnergyWeb, which was also formed in 2017, is developing a common protocol for energy applications of blockchain technology.[40]

3.4 Process Automation vs the “Black Box” Effect

Some of the blockchain implementations can execute transactions automatically in response to prespecified conditions, using what is known as “smart contracts.”[41] For example, some blockchain protocols include pieces of code that can make payments at the component level when subcomponents have been fully integrated and delivered to the final customer. In another example, orders for replacement parts can be automatically placed throughout the entire supply chain when a machine is brought in for repairs. Skuchain offers automation for performing trade transactions.[42]

The downside of automating transactions is the “black box” effect. Blockchain can remove the need to trust a counterparty in some circumstances, but it requires “meta-trust” in the blockchain concept (i.e., trust in a protocol or a distributed system but not in a specific company, individual, or government entity). Users of the system need to trust the integrity of the process without understanding the technical underpinnings. Even if one trusts the concept of distributed ledgers, one needs to trust the specific implementation as well. Returning to the example of TradeLens, the participants might question the incentives of the owners of that blockchain.

The use of smart contracts relies on the feasibility of anticipating future contingencies in supply chains. The greater the task that a smart contract is supposed to automate, the higher the probability that non-anticipated events will occur (recall

[39] http://www.bita.studio/.

[40] http://www.energyweb.org/.

[41] See Hilary (2022) for a discussion of smart contracts.

[42] https://www.skuchain.com/.

the complexity and uncertainty challenges of supply chain management). Real-life supply chain contracts are necessarily incomplete in that contracts cannot specify what is to be done in every possible contingency. Except for highly regimented and narrow applications, it is unlikely that smart contracts on a blockchain can govern general supply chain interactions without occasional human interventions.

A possible remedy to the "black box" effect and the incompleteness of supply chain interactions is to rely on partial automation with humans both checking the decisions of the system and handling exceptions. For example, in Skuchain's "bracket" system, human managers need to approve transactions. It is difficult to find a balance between automation and manual work that does not increase the cognitive load of managers. The development of self-driving cars highlighted the problem that humans struggle to turn on their attention instantly and step in to handle a complex situation.

3.5 System Resiliency vs Inefficiency

Although this is not the main point of the technology, a blockchain database is resilient to system disruptions by design because it is replicated on multiple nodes. This allows for a better recovery after a natural disaster or a cyber-incident, such as a distributed denial of service (DDOS) attack. As Babich and Hilary (2020) discussed, resiliency to computer malware is poorly understood at this point.

The cost of resiliency is inefficiency. There are at least two reasons for inefficiency. First, some protocols, such as proof of work, are inefficient by design to deter tempering. Second, the entire blockchain network needs to be updated once information is validated at one of the nodes. Such inefficiency is not a bug; it is a feature. There are some ways of addressing it, however. For instance, one can restrict what is recorded on the blockchain. Instead of saving the actual video, which consumes significant space, a hash of the video can be recorded on the blockchain instead. The system can be designed so that only a part of the network is updated when blocks are validated. Faster and more efficient consensus mechanisms, such as the proof of stake and proof of authority, can be used instead of proof of work. However, this increases the complexity of design and does not work for all applications.

4 Examples of Research Problems in SCM Involving Blockchain Technology

Babich and Hilary (2020) discussed a range of research questions for which blockchain technology provides a new perspective. Some are classical OM problems, while others suggest new business models. In this section, we expand on the discussion in Babich and Hilary (2020) of the bullwhip effect and use the example of PPE supply chains from the Introduction section to illustrate both the application of our 5-and-5 framework of strengths and weaknesses and research opportunities. Many, but not all, problems in PPE supply chains can be attributed to bullwhip effect causes. We point out those additional problems at the end of this section.

We begin with a few preliminaries. First, Babich and Hilary (2020) identified (1) information, (2) tokenization, and (3) automation as research themes of blockchain in operations management. The *information* research theme explores the strengths and weaknesses of blockchain technology concerning trust in data, supply chain visibility, and data aggregation. The *tokenization* research theme focuses on the benefits and costs of creating digital claims (tokens) to corporate and supply chain assets. The *automation* research theme explores the consequences of automating decisions, orders, and payments. We shall use these tags in the following discussion.

Second, OM researchers are familiar with the causes and consequences of the bullwhip effect (BWE), especially after the publication of the seminal paper by Lee et al. (1997b). Other readers may be less familiar with these issues. Therefore, for the benefit of these readers, we review the key ideas from Lee et al. (1997b) in some detail. Furthermore, a managerial introduction is given in Lee et al. (1997a), while a review of Bullwhip effect research can be found in Wang and Disney (2016).

The bullwhip effect (BWE) refers to the phenomenon that the variance of orders that a company places, is greater than the variance of demand this company observes, e.g., the variance of order to the suppliers is greater than the variance of sales of the buyer and that this distortion propagates upstream. An increasing variance increases operational and financing costs (see Babich and Birge, 2021 and Babich and Birge, 2022) and reduces firm value because higher variance makes decision making (e.g., how much capacity to build, how much to produce, how much to procure, where and how much to stock inventory) more difficult and increases the likelihood of a mismatch between supply and demand. In the context of PPE supply chains, the mismatch between supply and demand translates into PPE shortages and PPE overstocking and waste of potentially life-saving resources. Lee et al. (1997b) pointed out four possible causes of the bullwhip effect: (1) demand signal processing, (2) rationing games, (3) order batching, and (4) price variations.

Demand signal processing (cause (1) of BWE) refers to the phenomenon in which the retailer who uses the optimal inventory-order policy, while facing serially correlated demand and applying appropriate demand forecasting techniques, will place orders whose variance is greater than the variance of the actual demand process. The mathematical analysis shows that such inflation of the variance increases the order replenishment lead times and the uncertainty about the replenishment lead times.

Lee et al. (1997b) proposed several countermeasures to demand signal processing effects. One is to provide upstream firms with point-of-sales data. Another countermeasure is to delegate control over inventory to upstream firms or delegate control over production to downstream firms, for example, using arrangements such as a vendor-managed inventory. The third countermeasure is to reduce the replenishment lead time. These three countermeasures easily map into information, tokenization, and automation themes of blockchain research.

The first countermeasure–sharing point-of-sales data–naturally falls under the *information theme* of blockchain research. It might be less obvious that the third countermeasure–control over replenishment lead times–can also be accomplished with information. One reason for lead time uncertainty is that buyers do not know how much capacity is available upstream, how much inventory is en route, and when the orders will arrive. By sharing supply information, this uncertainty is reduced. However, why use blockchain? Many technological solutions enable data sharing; for example, hospitals can be linked to the ERP systems of distributors or manufacturers. As we discussed in our 5-and-5 framework, the key strength of blockchain technology is promoting *trust in data*. High-trust requirements might not be necessary during normal times when incentives to manipulate records are low. However, when supply shortages are extreme, procurement managers have strong incentives to manipulate data to increase the share allocated to their hospital or state. Conversely, suppliers of PPE may prefer to sustain buyers' beliefs that supply is limited to maintain high prices. Alternatively, suppliers may want to quell panic buying and lie that the supply is getting ready to increase. Either way, suppliers have incentives to manipulate their data. This is an opportunity for technology that promotes trust in data. However, as we discussed, the flip side of immutable blockchain records is the *GIGO* weakness and the disconnect between the physical world and digital records. Recall that a way to address the *GIGO* problem is by combining blockchain with other technologies, such as the Internet of Things (IoT) (for automating data collection) and artificial intelligence (AI, for processing text and images and fraud detection), in ways that create more reliable records.

Blockchain technology affords other solutions, which are omitted from Lee et al. (1997b) to the bullwhip effect caused by demand signal processing (discussed above). As explained in our 5-and-5 framework, blockchain facilitates data aggregation among multiple sources, e.g., across hospitals, retailers, and doctors' offices. For manufacturers of PPE, access to such industry-level data (and not only from their customers) can help to better forecast future demand and make capacity investment decisions. Because of the dynamics of the pandemic, some areas of the country were affected before others. Having data from regions that have already been affected helps hospitals in unaffected regions prepare for future developments. Aggregated and trustworthy data can also help state and federal agencies assess the extent of the crisis and take mitigation steps.

The second countermeasure—unifying control over the procurement process—may be accomplished through the *automation theme* of blockchain research. Instead of giving control to a supplier or buyers, procurement decisions can be delegated to smart contracts, which would place an order when certain conditions are met. Unfortunately, as we discussed in our 5-and-5 framework, this gives rise to the "black box" effect and requires all the participants in a supply chain to trust not only that the data are trustworthy but also that the automatic system is fair. In a pandemic, automation faces additional challenges because of the unpredictable business environment the pandemic creates. New sources of supply are added to supply chains, and greater flexibility in procurement and payment protocols (as examples in the Introduction illustrate) is needed.

Rationing games (cause (2) of BWE) occur when suppliers allocate scarce products in a manner that is proportional to the orders placed by the buyers. This creates an incentive for the buyer to inflate their orders relative to their true demand (hence causing the bullwhip effect) by anticipating a shortage and hoping to secure some of the supply. The problem is exacerbated if buyers are unaware of the extent of the shortage and if there are no restrictions on how much they can order, cancel, and return. The countermeasures proposed by Lee et al. (1997b) are to allocate products based on past sales (if there is a shortage), to share capacity and supply information, and to reduce the flexibility of the buyers to order and return (e.g., require nonrefundable capacity reservations). In PPE supply chains, there are numerous examples of order inflation and duplication among multiple suppliers, as well as cancellations.

Having trustworthy information about the production capacity and inventory across a supply chain can help to alleviate panic buying, over-ordering, and control rationing games. This solution fits with the *information theme* or blockchain research. However, similar to demand information, suppliers may have concerns about sharing their private information and losing their strategic advantage relative to buyers.

Information about the procurement process of retailers can also be shared on a blockchain. Then, suppliers can observe not only the actual demand but also whether a retailer has already placed numerous orders with other suppliers. While useful for keeping order inflation in check, this raises competitive concerns (other retailers may use this information to their advantage).

Switching to the *tokenization theme* of blockchain research, suppliers may require buyers to place cryptocurrency or digital claims on assets in digital escrow accounts and program smart contracts to transfer deposits to the supplier or the buyer when certain conditions are met. This approach can reduce the flexibility of the buyers, as Lee et al. (1997b) suggested. What is not considered in Lee et al. (1997b), however, is that blockchain tokens can be issued as claims on scarce production resources and inventory. These tokens can then be traded with customers, thereby creating markets and leading to a more efficient allocation of resources. One challenge in designing such markets is that the buyers with the deepest pockets, rather than those with the greatest medical needs, can end up owning most of the tokens. Furthermore, the existence of contracts, whether physical or digital, does not guarantee the delivery of PPE. As was seen during 2020, vendors can fail to deliver, or some buyers can "hijack" scarce resources, thus ignoring outstanding claims. Another challenge is that the products are not necessarily fungible or interoperable (as the experience with the Ebola outbreak demonstrated). PPE equipment may have unique features, and the gowns from one vendor may not work with the pants or gloves from another. The market design must account for this feature.

Order batching (cause (3) of BWE) occurs when procurement managers wait until a sufficient volume of orders has been accumulated (e.g., enough for a full truck) or when certain calendar dates occur (e.g., the beginning of the month) before placing orders and other procurement managers follow the same algorithm. Lee et al. (1997b) proposed the following countermeasures: use automated systems to reduce order processing costs, consolidate orders with other customers by using third-party logistics providers, and have regular delivery appointments.

These measures fit the *automation theme* of blockchain research, and smart contracts can be used to implement them. Furthermore, by relying on the visibility strength of blockchain technology (see the 5-and-5 framework), the consumption of PPE can trigger orders not just from PPE distributors but all the way upstream of the supply chain to manufacturer and raw material providers. The data aggregation strength of blockchain technology can help manufacturers (e.g., 3M) aggregate consumption data even if PPE is delivered through different retail channels.

Fluctuating prices (cause (4) of BWE) create the bullwhip effect because retailers strategically decrease their purchases during times of high wholesale prices and increase their purchases during times of low wholesale prices. Fluctuating prices do not appear to be a major factor in PPE supply chains, probably because shortages at the time were so severe that hospitals did not have the luxury of waiting until prices would fall. However, in general, blockchain technology can be used to remove the strategic waiting behavior of buyers. For example, suppliers can implement "price match" guarantees using blockchain so that they automatically refund buyers if wholesale prices drop in the future.

Last, switching the focus of our discussion from the bullwhip effect to counterfeiting and adulteration risk, we note that Babich and Hilary (2020) discussed the role of blockchain technology at every step of the supply chain risk management process. Intuitively, knowing the provenance and the chain of custody of PPE can go a long way toward preventing counterfeit and defective products from making their way to consumers. However, just knowing the identities of the firms in supply chains might not be enough if those companies have very short-term objectives and are willing to risk their reputation. A scandal at some of Kobe's and Mitsubishi's steel and aluminum plants illustrates this point (McLain, 2017). Managers at these plants intentionally modified the formulas for alloys to reduce production costs. Although customers (which included most major automakers) knew the supplier's identity, this did not stop their adulteration. Therefore, visibility into the process and inputs being enabled by blockchain technology may be required. Again, as explained in the 5-and-5 framework, the GIGO problem is the main weakness. While it appears relatively easy to record identities on a blockchain, establishing a data collection process to track production inputs and production steps and making this process resistant to tempering is a nontrivial task.

5 Conclusions

Blockchain is expected to improve supply chain operations by providing visibility, information aggregation, information validation, contract automation, and system resiliency. Babich and Hilary (2020) reviewed some preliminary applications of these features. In this chapter, we take stock of the situation two years after that article was written. By and large, the process has been evolutionary rather than revolutionary. For example, blockchain has facilitated the integration of supply chains and finance. As expected with new technology, many initially promising projects have failed, but new applications have been initiated in their place. Some organizations have gone beyond pilots, but few have completely revamped their strategies and operations around the new technology. Certain natural applications of technology have not yet been developed but remain viable options for the future. For example, applications relying on tokenization in supply chains remain limited as of the time this chapter was written (although financial speculation in nonfungible tokens (NFT) may attract interest to more mundane supply chain and supply chain finance applications). Applications exploiting the ability of blockchain technology to integrate heterogeneous systems are lagging as well. Some of the Industry 4.0 elements, such as cyber-physical systems or the Internet of Things (IoT), stand to benefit from better integration. However, these applications, particularly those that connect loosely related actors, are not yet mature. They may become more common with the development of 5G networks. If this evolution appears, it will most likely require significant investment, and it will be interesting to see if this leads to the emergence of new actors or if the existing players will be able to capture these new markets. However, one thing is clear: more research and analysis are required.

References

Adiaba, S., Hammond, F., Proverbs, D., Lamond, J., Booth, C, & Bloch, R. (2011). Land registration and information asymmetry problem in sub-saharan Africa urban real estate markets. School of Engineering and the Built Environment, pp. 877–893.

Agca. S., Babich V., Birge J. R., Wu J. (2021). Credit shock propagation along supply chains: Evidence from the CDS market. Management Science, forthcoming, https://doi.org/10.1287/mnsc.2021.4174

Babich, V., & Birge, J. R. (2021). The interface of finance, operations, and risk management. *Foundations and Trends® in Technology, Information and Operations Management, 15*(1–2), 1–203.

Babich, V., & Birge, J. R. (2022). Interface of operations and finance: A tutorial. In: V. Babich, J. R. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations*. Berlin: Springer.

Babich, V., & Hilary, G. (2019). Blockchain and other distributed ledger technologies in operations. *Foundations and Trends® in Technology, Information and Operations Management 12*(2–3), 152–172.

Babich, V., & Hilary, G. (2020). Distributed ledgers and operations: What operations management researchers should know about blockchain technology. *Manufacturing & Service Operations Management, 22*, 223–240.

Cohen, L., & Frazzini, A. (2008). Economic links and predictable returns. *The Journal of Finance, 63*(4), 1977–2011.

Fenech, G. (2018). The supply chain revolution on the blockchain. Section: Crypto & Blockchain. https://www.forbes.com/sites/geraldfenech/2018/12/18/the-supply-chain-revolution-on-the-blockchain/

Hellwig, D., & Huchzermeier, A. (2021). Distributed ledger technology and fully homomorphic encryption: Next-generation information-sharing for supply chain efficiency. In V. Babich, J. R. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations*. Berlin: Springer.

Hertzel, M. G., Li, Z., Officer, M. S., & Rodgers, K. J. (2008). Inter-firm linkages and the wealth effects of financial distress along the supply chain. *The Journal of Finance, 87*(2), 374–387.

Hilary, G. (2022). Blockchain and other and other distributed ledger technologies, an advanced primer. In: V. Babich, J. R. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations*. Berlin: Springer.

Hilary, G., & Liu, L. X. (2021). Blockchain and other distributed ledger technologies in finance, springer books. In R. Rau, R. Wardrop, & L. Zingales (Eds.), *The Palgrave handbook of technological finance* (pp. 243–268). Springer.

Hofmann, E., Strewe, U. M., & Bosia, N. (2017). *Supply chain finance and blockchain technology: The case of reverse securitisation*. Berlin: Springer.

Kolay, M., Lemmon, M., & Tashjian, E. (2016). Spreading the misery? Sources of bankruptcy spillover in the supply chain. *Journal of Financial and Quantitative Analysis, 51*(6), 1955–1990.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997a). The bullwhip effect in supply chains. *Sloan Management Review, 38*, 93–102.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997b). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*(4), 546–558.

McLain, S. (2017). Mitsubishi materials' quality scandal kept quiet for months. WSJ, Nov 24, 2017. Retrieved February 26, 2018 from https://goo.gl/C3Hdsu

Mehrotra, P., Malani, P., & Yadav, P. (2020). Personal protective equipment shortages during covid-19–supply chain–related causes and mitigation strategies. *JAMA Health Forum 1*(5), e200,553–e200,553.

Patel, A., D'Alessandro, M. M., Ireland, K. J., Burel, W. G., Wencil, E. B., & Rasmussen, S. A. (2017). Personal protective equipment supply chain: Lessons learned from recent public health emergency responses. *Health Security, 15*(3), 244–252.

Shibuya, Y., & Babich, V. (2021). Multi-tier supply chain financing with blockchain. Available at SSRN 3787044.

Tapscott, D. (2020). *Supply chain revolution: How blockchain technology is transforming the global flow of assets*. Toronto: Barlow Publishing.

Wang, X., & Disney, S. M. (2016). The bullwhip effect: Progress, trends and directions. *European Journal of Operations Research, 250*(3), 691–701.

Impact of Blockchain-Driven Accountability in Multi-Sourcing Supply Chains

Yao Cui, Ming Hu, and Jingchen Liu

1 Introduction

With the rapid development of globalization, supply chains become more complex than ever. Nowadays, it is prevailing for companies to source the same material or product from multiple suppliers, who can even locate in different countries. Since suppliers' efforts to improve product quality are usually unobservable and non-contractible, multi-sourcing makes managing supply chain quality more difficult and may result in more product failures. Product defects and recalls are widely observed in various industries, such as agri-food, pharmaceutical, automobile, and smartphones.[1]For example, food recalls due to salmonella contamination have frequently occurred in peanut, milk, beef, salad, turkey, melon, and cereal (Basu, 2015; Goldschmidt, 2018; Karimi & Goldschmidt, 2018). Product defects not only cause consequential financial and reputational damages to the companies concerned but could also pose a safety risk to the public. When facing the risk of product

[1] See https://www.kiplinger.com/slideshow/investing/T052-S000-10-biggest-product-recalls-of-all-time/index.html.

Y. Cui
Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, NY, USA
e-mail: yao.cui@cornell.edu

M. Hu
Rotman School of Management, University of Toronto, Toronto, ON, Canada
e-mail: ming.hu@rotman.utoronto.ca

J. Liu (✉)
School of Business, Nanjing University, Nanjing, China
e-mail: jingchenliu@nju.edu.cn

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_4

defects, a firm needs to design a mechanism to transfer the resulting market loss to the suppliers at fault. However, without the proper technology, it may be difficult for a firm to hold its suppliers accountable for their own faults. For instance, there may not exist a third party who can guarantee the identification to be credible, and this difficulty can be further magnified if the suppliers are small companies or reside in a foreign country (Babich & Tang, 2012). Moreover, the suppliers may not agree to a payment scheme that is contingent on the actual product defects upfront, because they may fear that the buying firm would misreport the defects and overcharge them for compensations. Because of all these challenges, a buying firm and its suppliers may not be able to reach an agreement on a contract that penalizes particular suppliers who the buying firm identifies to be defective. In such cases, if the buying firm wants to implement a penalty term, it may instead opt to charge penalties on all suppliers when a defect occurs in the end product, so that all suppliers will be paid or penalized at the same time; such a mechanism has been studied by the previous literature as the "group-warranty contract" (see, e.g., Baiman et al., 2004; Li, 2012).

The recent development in the blockchain technology can help firms overcome the challenges that arise from a lack of accountability in supply chains. Blockchain is a decentralized digital ledger technology that can record transactions efficiently, verifiably, and permanently. When applied to supply chains, blockchain can be used to record quality-related information along the entire production and logistics process (Nash, 2016; Bajpai, 2019; Metcalfe, 2019; Hellwig & Huchzermeier, 2021). When an end-product failure occurs, such credible quality-related information can be used to verify the quality outcomes of the suppliers. More importantly, a smart contract, implemented along with blockchain, is a self-enforcing transactional protocol relying on tamper-proof consensus on contingent outcomes (Hilary, 2021). Once agreed upon and stored in a blockchain, smart contracts are irreversible, and can automatically execute contract terms without the need for a third-party intermediary (Chu, 2016, 2017). It can thus help facilitate automation of payment and penalization in a supply chain based on the quality outcomes of the suppliers that are identified from the information recorded in the blockchain. Therefore, along with smart contracts, blockchain can significantly improve the accountability of supply chains and may enforce new payment structures.

The accountability enabled by blockchain can impact supply chain quality contracting. As is common in agri-food and pharmaceutical industries, a firm may source the same material or product from multiple suppliers and convert it into the final product to sell to the consumers. Traditionally, because the quality outcomes are not verifiable, the firm may not be able to hold the suppliers accountable for their own defects. In this case, the firm may have to penalize all suppliers at the same time by collecting some pre-determined amount of penalties from them when a defect occurs. In practice, as required by U.S. Food and Drug Administration, manufacturers and suppliers are responsible for any costs associated with all product recalls (Nath, 2020). Some firms (e.g., Wal-Mart, Kroger, and Albertsons) even charge their suppliers additional product removal processing fee for product recalls of any reason (Prevor, 2008). Nevertheless, blockchain can enable the firm to hold the suppliers accountable for their own faults and execute supplier-specific penalties accordingly. Specifically, the information recorded in the blockchain can make the quality outcomes verifiable, and the smart contracts can enable the firm to automatically execute the payment and penalization based on the pre-defined protocol and the information recorded in the blockchain. In this case, the firm can penalize only the failure-causing suppliers when a defect occurs.

Given the rapid development of blockchain and its promise in improving supply chain accountability, this research aims to understand its impact in supply chain quality contracting. In particular, we are going to investigate the following research questions. First, how does the accountability enabled by blockchain impact the suppliers' quality decisions and the supply chain contract? Second, how would the impact of accountability be affected by the complexity of the supply chain and the buyer's ability to penalize the suppliers?

To answer the above research questions, we consider a *multi-sourcing* supply chain that consists of a buyer and an arbitrary number of suppliers. All suppliers jointly determine the quality of the end product. Although the actual quality outcome is subject to uncertainty, suppliers can exert an effort to improve the probability that their output is non-defective. The buyer uses contracts, where payments are contingent on the realization of quality, to induce the suppliers to choose the desired quality levels. The buyer offers each supplier a contract, and the payment scheme depends on whether the supply chain is accountable or not. If the supply chain is not accountable, the payments to all suppliers are contingent on the quality outcome of the end product (i.e., all suppliers receive a payment if the end product is non-defective and a penalty if the end product is defective). If the supply chain is accountable, the payment to each supplier is contingent on his own quality outcome (i.e., each supplier receives a payment if he is non-defective and a penalty if he is defective). By comparing the equilibrium contracts, we develop insights into the value of accountability in the multi-sourcing supply chain.

We highlight our main findings and contributions as follows. In the multi-sourcing supply chain without accountability, the equilibrium contract will result in over-penalization for the suppliers, because the buyer will penalize all suppliers when the end product is defective even if one supplier is non-defective himself. This, in turn, requires the buyer to offer higher wholesale prices to the suppliers when the end product is non-defective so that they are willing to participate in the first place. As a result, if the retail price is not sufficiently high, the buyer's total wholesale price payment to the suppliers can exceed the retail price he earns, implying that the buyer's cash flow may be negative, referred to as being infeasible, when the end product is non-defective. Thus, without accountability, the equilibrium contract will likely result in a payment scheme that is difficult to implement in practice. However, if the supply chain is accountable, the buyer can penalize a supplier only if he is defective himself. Thus, the suppliers are no longer over-penalized, and the buyer can induce all suppliers to participate with lower wholesale prices. As a result, the buyer's total wholesale price payment to the suppliers is always below the retail price, so he no longer faces a cash flow infeasibility issue. Therefore, in the multi-sourcing supply chain, accountability creates value by *ensuring cash flow feasibility* for the buyer, so that the equilibrium contract becomes more implementable in practice and first-best becomes more readily achievable.

Furthermore, we find that as the number of suppliers increases (so that the supply chain becomes more complicated), the value of accountability is *strengthened* in the multi-sourcing supply chain. In particular, as the number of suppliers increases, the buyer's cash flow is more likely to be infeasible without accountability, whereas it is always feasible with accountability. In contrast, we find that when suppliers face limited liability constraints, the value of accountability is *weakened* in the multi-sourcing supply chain. Specifically, due to suppliers' limited liability, the buyer cannot charge high penalties, and thus he has no incentive to offer exorbitantly high wholesale prices, leading to less likely cash flow infeasibility in the case without accountability. Since the issue without accountability is mitigated by limited liability constraints, the value of accountability is weakened accordingly.

This work is related to the literature that studies quality contracting in multi-supplier supply chain settings. For example, Baiman et al. (2004) and Li (2012) consider an assembly supply chain in which the buyer assembles an end product using outsourced parts from multiple suppliers. Baiman et al. (2004) compare the two contracts that require individual testing with the group-warranty contract, and Li (2012) further studies the optimal group-warranty contract in different scenarios. While both papers show that the group-warranty contract works well in an assembly supply chain, we find that this kind of contract will likely result in a payment scheme that is difficult to implement in practice. Mu et al. (2016) examine two quality-testing strategies, individual testing and mixed testing, to curb deliberate adulteration by milk farmers in a multi-sourcing supply chain. Besides quality management, the operations management literature has also studied many other issues in multi-sourcing supply chains. Jiang and Wang (2010), Fang et al. (2014), and Hu and Qi (2018) study the optimal procurement contract design of an assembly supply chain. Gümüş et al. (2012) and Ang et al. (2017) explore the optimal sourcing problem with the consideration of disruption risk on the supply side. Chen et al. (2020) investigate how to manage suppliers' social and environmental responsibility when the buying firm sources from multiple suppliers. This work is also related to the growing body of literature that investigates the operational and financial impacts of blockchain (Babich & Hilary, 2020). Some researchers study the operational impacts of the information recorded in the blockchain in combating counterfeits (Pun et al., 2021), signaling firms' quality to lenders (Chod et al., 2020), revealing the network visibility of a supply chain with competing firms (Cui et al., 2020), and enabling end-to-end product traceability in a multi-tier supply chain (Cui et al., 2021). Other researchers focus on the financial impacts of blockchain and study the tokenization of blockchain-based platforms (Chod et al., 2021), the effectiveness of token-weighted voting (Tsoukalas & Falk, 2020), and the design of token floating and pricing for Initial Coin Offerings (ICOs) (Gan et al., 2021).

2 The Model

We consider a multi-sourcing supply chain that consists of a buyer and $n \geqslant 2$ suppliers. Following the quality contracting literature (e.g., Balachandran & Radhakrishnan, 2005; Hwang et al., 2006; Nikoofal & Gümüş, 2018), we normalize the market demand to one. The buyer procures $1/n$ unit of product from each supplier and converts it into one unit of final product to sell to the consumers.

Supplier $i \in \{1, 2, \ldots, n\}$ needs to make his own quality decision $q_i \in [0, 1]$, which indicates the probability that the product made by him is non-defective. Although the actual product outcome is subject to uncertainty, suppliers can incur additional costs to improve the chance of having a non-defective product. In particular, supplier i's quality cost function is $C_i(q_i)$. Following the quality contracting literature (e.g., Baiman et al., 2001; Balachandran & Radhakrishnan, 2005; Chao et al., 2009; Plambeck & Taylor, 2016), we assume that $C_i(q_i)$ is twice continuously differentiable on [0, 1], and convexly increasing in quality, i.e., $C_i'(q_i) > 0$ and $C_i''(q_i) > 0$, for $q_i \in (0, 1]$. Note that our analysis does not require the suppliers to have the same quality cost function. The quality of the end product sold to consumers is determined by the quality of all suppliers involved.[2] Specifically, the end product is non-defective with probability $\prod_{i=1}^{n} q_i$, and defective with probability $1 - \prod_{i=1}^{n} q_i$. Such a feature is referred to as the "weakest link" property in the literature, and has been adopted to study multi-sourcing supply chains (e.g., Baiman et al., 2004; Mu et al., 2016).

The buyer sells the end product to consumers for a price of p per unit without knowing its exact quality. After the sales, if the end product is non-defective, then no further action is needed. Conversely, if the end product is defective, then the buyer has to refund the price p to consumers and will also incur a loss l per unit, where $p, l > 0$. The loss l may be incurred due to product recalls, customer dissatisfaction, reputation damage, and market loss (e.g., Balachandran & Radhakrishnan, 2005; Hwang et al., 2006; Chao et al., 2009). For example, in the Ford-Firestone recall case in 2000, l would include the costs of replacing all tires, advertising to restore lost reputation, and lawsuits (Isidore, 2001).

The quality level q_i chosen by supplier i is neither observable nor verifiable to outside parties. Hence, it is infeasible for the buyer to offer a contract contingent on q_i to supplier i. It is easy to verify that a simple wholesale contract in this case would result in a failure of the supply chain, because in equilibrium suppliers would exert no effort in improving quality and the buyer would offer the lowest wholesale price possible. We thus consider contracts where the buyer's payments to the suppliers are contingent on the quality outcome of the end product. Denote Π_{B} and Π_{S_i} as the expected profit of the buyer and supplier i, respectively. With the outside

[2] We focus on the suppliers' quality decisions and do not consider the buyer's quality decision, which is consistent with the literature (e.g., Baiman et al., 2000; Hwang et al., 2006; Babich & Tang, 2012; Rui & Lai, 2015; Nikoofal & Gümüş, 2018). The main insights of the paper would carry through if the buyer's quality decision is incorporated.

opportunity cost normalized to zero, the buyer has an incentive to offer contracts if and only if $\Pi_{\mathrm{B}} \geqslant 0$, and supplier i has an incentive to accept the contract if and only if $\Pi_{\mathrm{S}_i} \geqslant 0$. All proofs are relegated to the appendix.

2.1 First-Best Scenario

Under this setting, we first characterize the first-best equilibrium when quality is contractible. Note that the first-best scenario corresponds to a centralized supply chain in which the buyer makes all quality decisions. Denote $\mathbf{q} \equiv (q_1, q_2, \ldots, q_n)$. The buyer's first-best problem is formulated as follows:

$$\max_{\mathbf{q}} \ \Pi_{\mathrm{B}}(\mathbf{q}) = p \prod_{i=1}^{n} q_i - l \left(1 - \prod_{i=1}^{n} q_i\right) - \sum_{i=1}^{n} C_i(q_i). \tag{1}$$

The first two terms correspond to the expected revenue from selling the product, and the last term is the total quality cost from all suppliers. The following result characterizes the solution to (1) under some regularity assumptions (see details below), with superscript "$*$" denoting the first-best optimum.

Assumption 1 (Existence of Interior Solution) $C_i(0) = C_i'(0) = 0$*,* $0 \leqslant C_i''(0) < p + l$*, and* $C_i'(1) > p + l$*, for* $i \in \{1, 2, \ldots, n\}$*.*

Assumption 2 (Uniqueness of Interior Solution) $C_i''(q_i^*) > (n-1)(p+l)$*, for* $i \in \{1, 2, \ldots, n\}$*.*

Lemma 1 (First-Best Equilibrium) *Under Assumptions 1 and 2, the first-best quality levels* $\mathbf{q}^* = (q_1^*, q_2^*, \ldots, q_n^*)$ *are given by the unique solution to the following system of equations:*

$$(p+l) \prod_{j=1, j \neq i}^{n} q_j^* = C_i'(q_i^*), \quad i \in \{1, 2, \ldots, n\}, \tag{2}$$

and $q_i^* \in (0, 1)$ *for* $i \in \{1, 2, \ldots, n\}$*. Besides,* q_i^* *decreases in* n *if suppliers are symmetric in the sense of having the same quality cost functions. Furthermore, if* $C_i'(q) \leqslant C_j'(q)$ *for all* $q \in (0, 1)$*, where* $i, j \in \{1, 2, \ldots, n\}$ *and* $i \neq j$*, then* $q_i^* \geqslant q_j^*$*.*

Lemma 1 shows that at the first-best solution, the marginal value of quality improvement is equal to the marginal cost. Moreover, under Assumptions 1 and 2, the existence of a unique interior first-best solution is guaranteed, i.e., $q_i^* \in (0, 1)$ for $i \in \{1, 2, \ldots, n\}$. To be specific, Assumption 1 contains technical conditions for the optimal solution to be achieved at an interior point; whereas it is common in the literature to assume $C_i'(1) = \infty$ (e.g., Baiman et al., 2000, 2001; Balachandran

& Radhakrishnan, 2005; Chao et al., 2009), we do not require $C_i'(q_i)$ to approach infinity as q_i approaches one. Assumption 2 ensures that a stationary point is also a local maximum; note that this assumption only needs to hold at the stationary point. These assumptions can be satisfied by a wide range of cost functions, such as the logarithmic function and power function. We proceed the analysis with these assumptions, unless otherwise stated. Moreover, when all suppliers are symmetric in the sense of having the same quality cost functions, the equilibrium quality level q_i^* will be lower as the supply chain becomes more complex (i.e., as n increases). Lastly, Lemma 1 also indicates that the equilibrium quality level of a more efficient supplier (i.e., one with a lower marginal quality cost) is higher than that of a less efficient supplier.

3 Equilibrium Analysis

In this section, we analyze the equilibria of the main model. We first analyze the contract without accountability, in which case the buyer can only penalize all suppliers at the same time when a product failure occurs. Then, we analyze the contract with accountability, in which case the buyer can hold the suppliers accountable for their own faults and penalize only the supplier(s) who caused the product failure.

3.1 Equilibrium without Accountability

We first consider the case without accountability. In this case, the buyer cannot hold the suppliers accountable for their own faults. Thus, he can only penalize all suppliers at the same time when a defect occurs, i.e., the buyer employs a group-warranty contract. The contract is characterized by contingent payments $(\mathbf{w}, \mathbf{t})$, where $\mathbf{w} \equiv (w_1, w_2, \ldots, w_n)$ and $\mathbf{t} \equiv (t_1, t_2, \ldots, t_n)$. Specifically, the buyer offers contract (w_i, t_i) to supplier $i \in \{1, 2, \ldots, n\}$. If contract (w_i, t_i) is accepted, supplier i produces the product with a chosen quality level q_i. The buyer pays the suppliers based on the realization of the end-product quality. If the end product is non-defective, the buyer receives a revenue p and pays each supplier a wholesale price w_i. If the product is defective, the buyer incurs a loss l and pays each supplier t_i (or equivalently, supplier i "pays" $-t_i$ to the buyer; hence $t_i < 0$ means that the buyer charges supplier i a penalty if the end product is defective). Note that the contract is equivalent to one where the buyer pays w_i upfront and then penalizes $w_i - t_i$ if the end product turns out to be defective.

The buyer's contracting problem in the multi-sourcing supply chain without accountability is formulated as follows:

$$
\max_{\mathbf{w},\mathbf{t}} \quad \Pi_{\mathrm{B}}(\mathbf{w},\mathbf{t}|\mathbf{q}) = p\prod_{i=1}^{n} q_i - l\left(1-\prod_{i=1}^{n} q_i\right) - \left(\sum_{i=1}^{n} w_i\right)\prod_{i=1}^{n} q_i - \left(\sum_{i=1}^{n} t_i\right)\left(1-\prod_{i=1}^{n} q_i\right)
$$
$$
\text{s.t.} \quad \begin{cases} \Pi_{\mathrm{S}_i}(q_i|w_i,t_i,q_{-i}) \geqslant 0, & (\mathrm{IR}_i) \\ q_i = \arg\max\limits_{q_i'} \; \Pi_{\mathrm{S}_i}(q_i'|w_i,t_i,q_{-i}), & (\mathrm{IC}_i) \end{cases} \tag{3}
$$

where $\Pi_{\mathrm{S}_i}(q_i|w_i,t_i,q_{-i}) = w_i q_i \prod_{j=1,j\neq i}^{n} q_j + t_i\left(1 - q_i \prod_{j=1,j\neq i}^{n} q_j\right) - C_i(q_i)$, for $i \in \{1,2,\ldots,n\}$. The buyer maximizes his expected profit subject to supplier i's individual rationality constraint (IR_i) and incentive compatibility constraint (IC_i), for $i \in \{1,2,\ldots,n\}$. Given the contract $(\mathbf{w},\mathbf{t})$ offered by the buyer, the suppliers choose their quality levels simultaneously as self-interested profit maximizers. The first two terms of the buyer's expected profit function correspond to the expected revenue of selling the product, and the last two terms are the expected total payment to suppliers. The first two terms of supplier i's expected profit function are the expected payment from the buyer, and the last term is his own quality cost. The following proposition characterizes the solution to (3), with superscript "N" denoting the case without accountability and "†" denoting the equilibrium.

Proposition 1 (Equilibrium without Accountability) *In the multi-sourcing supply chain without accountability, the unique equilibrium contract is characterized by the following: the buyer offers contingent payments* $w_i^{\mathrm{N}\dagger} = C_i(q_i^{\mathrm{N}\dagger}) + (p + l)\left(1-\prod_{j=1}^{n} q_j^{\mathrm{N}\dagger}\right)$ *and* $t_i^{\mathrm{N}\dagger} = C_i(q_i^{\mathrm{N}\dagger}) - (p+l)\prod_{j=1}^{n} q_j^{\mathrm{N}\dagger}$ *to supplier* $i \in \{1,2,\ldots,n\}$*. Moreover,* $w_i^{\mathrm{N}\dagger} > 0$ *and* $t_i^{\mathrm{N}\dagger} < 0$*. In equilibrium, the supply chain achieves the first-best (i.e.,* $q_i^{\mathrm{N}\dagger} = q_i^*$ *and* $\Pi_{\mathrm{B}}^{\mathrm{N}\dagger} = \Pi_{\mathrm{B}}^*$*).*

Proposition 1 shows that in equilibrium, $w_i^{\mathrm{N}\dagger} > 0$ and $t_i^{\mathrm{N}\dagger} < 0$ always hold, which implies that the buyer charges all suppliers a penalty if the end product is defective. Moreover, Proposition 1 also shows that because the buyer employs both wholesale price payment and penalty as instruments, he is able to incentivize all suppliers to choose the desired quality levels and achieve the first-best. Finally, note that we do not impose limited liability constraints for the suppliers in the main model. We explore the case with suppliers' limited liability as an extension in Sect. 4.3.

3.2 Equilibrium with Accountability

We next consider the case with accountability. In this case, because the buyer can hold the suppliers accountable for their own faults, he only penalizes the defect-causing supplier(s). The contract is characterized by contingent payments $(\mathbf{w},\mathbf{t})$, where $\mathbf{w} \equiv (w_1, w_2, \ldots, w_n)$ and $\mathbf{t} \equiv (t_1, t_2, \ldots, t_n)$. Specifically, the buyer offers contract (w_i, t_i) to supplier $i \in \{1,2,\ldots,n\}$. If contract (w_i, t_i) is accepted, supplier i produces the product with a chosen quality level q_i. The buyer pays each supplier based on the realization of their own quality. If the end product is non-defective, the buyer receives a revenue p and pays each supplier w_i. If the end product is defective, the buyer incurs a loss l, and due to accountability, the buyer pays t_i to the defective supplier(s) (or equivalently, the defective supplier(s) "pay" $-t_i$ to the buyer). If supplier i turns out to be non-defective, he still receives w_i from the buyer even if the end product is defective.

The buyer's contracting problem in the multi-sourcing supply chain with accountability is formulated as follows:

$$\max_{\mathbf{w},\mathbf{t},\mathbf{q}} \quad \Pi_{\mathrm{B}}(\mathbf{w},\mathbf{t}|\mathbf{q}) = p\prod_{i=1}^{n} q_i - l\left(1-\prod_{i=1}^{n} q_i\right) - \sum_{i=1}^{n} w_i q_i - \sum_{i=1}^{n} t_i(1-q_i)$$
$$\text{s.t.} \quad \begin{cases} \Pi_{\mathrm{S}_i}(q_i|w_i,t_i) \geqslant 0, & (\mathrm{IR}_i) \\ q_i = \arg\max\limits_{q_i'} \ \Pi_{\mathrm{S}_i}(q_i'|w_i,t_i), & (\mathrm{IC}_i) \end{cases} \tag{4}$$

where $\Pi_{\mathrm{S}_i}(q_i|w_i,t_i) = w_i q_i + t_i(1-q_i) - C_i(q_i)$, for $i \in \{1,2,\ldots,n\}$. Similar to the previous case without accountability, the buyer maximizes his expected profit subject to the suppliers' individual rationality and incentive compatibility constraints, and given contract $(\mathbf{w},\mathbf{t})$ offered by the buyer, the suppliers choose their quality levels simultaneously. However, it is easy to see the difference between (3) and (4). Note that in (4), one supplier's quality decision does not enter another supplier's profit function, indicating that the suppliers' quality decisions are independent of each other. The following result characterizes the solution to (4), with superscript "A" denoting the case with accountability.

Proposition 2 (Equilibrium with Accountability) *In the multi-sourcing supply chain with accountability, the unique equilibrium contract is characterized by the following: the buyer offers contingent payments* $w_i^{\mathrm{A}\dagger} = C_i(q_i^{\mathrm{A}\dagger}) + (p+l)\prod_{j=1,j\neq i}^{n} q_j^{\mathrm{A}\dagger}\left(1-q_i^{\mathrm{A}\dagger}\right)$ *and* $t_i^{\mathrm{A}\dagger} = C_i(q_i^{\mathrm{A}\dagger}) - (p+l)\prod_{j=1}^{n} q_j^{\mathrm{A}\dagger}$ *to supplier* $i \in \{1,2,\ldots,n\}$. *Moreover,* $w_i^{\mathrm{A}\dagger} > 0$ *and* $t_i^{\mathrm{A}\dagger} < 0$. *In equilibrium, the supply chain achieves the first-best (i.e.,* $q_i^{\mathrm{A}\dagger} = q_i^*$ *and* $\Pi_{\mathrm{B}}^{\mathrm{A}\dagger} = \Pi_{\mathrm{B}}^*$*).*

Proposition 2 shows that in equilibrium, $w_i^{\mathrm{A}\dagger} > 0$ and $t_i^{\mathrm{A}\dagger} < 0$ always hold, implying that the buyer charges a penalty to supplier i if he is defective. Moreover, Proposition 2 shows that the buyer is also able to achieve the first-best in terms of quality under the optimal contract with accountability. We next compare the equilibrium contracts with and without accountability in the multi-sourcing supply chain.

4 Value of Blockchain-Driven Accountability

In this section, we study the impact of blockchain-driven accountability in the multi-sourcing supply chain. First, by comparing the equilibrium contracts, we develop insights into the value of accountability for the main model. Second, we study how the complexity of the multi-sourcing supply chain would impact the value of blockchain-driven accountability. Third, we incorporate suppliers' limited liability constraints and explore how it would impact the value of blockchain-driven accountability.

4.1 Equilibrium Comparison

Given the above characterized equilibrium contracts with and without accountability, we now study how accountability impacts the multi-sourcing supply chain. Although the buyer is able to achieve the first-best in both cases, the equilibrium outcomes are different. The following corollary summarizes the comparison of equilibrium contracts for the cases with and without blockchain, which follows immediately from Propositions 1 and 2.

Corollary 1 (Comparison of Contracts)

(1) $t_i^{\mathrm{A}\dagger} = t_i^{\mathrm{N}\dagger}$ *for* $i \in \{1, 2, \ldots, n\}$.
(2) $w_i^{\mathrm{A}\dagger} < w_i^{\mathrm{N}\dagger}$ *for* $i \in \{1, 2, \ldots, n\}$.
(3) $w_i^{\mathrm{N}\dagger} - w_i^{\mathrm{A}\dagger} \geqslant w_j^{\mathrm{N}\dagger} - w_j^{\mathrm{A}\dagger}$ *if* $C_i'(q) \leqslant C_j'(q)$ *for all* $q \in (0, 1)$, *where* $i, j \in \{1, 2, \ldots, n\}$ *and* $i \neq j$.

Corollary 1 shows that in the multi-sourcing supply chain, accountability does not change the amount of penalty (i.e., $t_i^{\mathrm{A}\dagger} = t_i^{\mathrm{N}\dagger}$ for $i \in \{1, 2, \ldots, n\}$). Although each supplier pays the same amount of penalty when he is penalized by the buyer, each supplier stands a higher chance of being penalized in the case without accountability, because he will also be penalized if any of the other suppliers is defective. By contrast, in the case with accountability, each supplier is penalized if and only if he is defective himself. Thus, the suppliers are charged higher penalties in expectation in the case without accountability. In order to incentivize the suppliers to participate, the buyer needs to offer higher payments to the suppliers when the product is non-defective (i.e., $w_i^{\mathrm{A}\dagger} < w_i^{\mathrm{N}\dagger}$ for $i \in \{1, 2, \ldots, n\}$). Therefore, accountability will decrease the wholesale prices that the buyer offers to the suppliers. Additionally, recall that our model allows the suppliers to be heterogeneous. Corollary 1 further shows that accountability lowers the wholesale price by a greater amount for a more efficient supplier (i.e., one with a lower marginal quality cost).

Proposition 3 (Impact of Blockchain-Driven Accountability) *In the multi-sourcing supply chain, accountability always guarantees the buyer's cash flow* *feasibility. In particular, without accountability, there exists a threshold $\bar{p} > 0$ such that $\sum_{i=1}^{n} w_i^{\mathrm{N}\dagger} > p$ if $p < \bar{p}$, implying that the buyer's cash flow is infeasible when p is not sufficiently large. By contrast, with accountability, $\sum_{i=1}^{n} w_i^{\mathrm{A}\dagger} < p$ always holds, implying that the buyer's cash flow is always feasible.*

We have seen that in the case without accountability, the equilibrium contract will result in over-penalization for the suppliers, which in turn requires the buyer to offer higher payments to the suppliers so that they are willing to participate in the first place. Proposition 3 further reveals that the high wholesale prices can create an issue in the buyer's cash flow in the case without accountability. In particular, if the retail price is not sufficiently high, the buyer's total payment to the suppliers will exceed the retail price (i.e., $\sum_{i=1}^{n} w_i^{\mathrm{N}\dagger} > p$ if $p < \bar{p}$), indicating that the buyer's cash flow will be infeasible when the end product is actually non-defective. This issue will significantly limit the implementability of the equilibrium contract. By contrast, in the case with accountability, the buyer's total payment to the suppliers is always below the retail price (i.e., $\sum_{i=1}^{n} w_i^{\mathrm{A}\dagger} < p$), and hence the buyer does not face the cash flow issue as in the case without accountability.

Therefore, in the multi-sourcing supply chain, the value of accountability lies in the buyer's ability to *guarantee cash flow feasibility*. In the multi-sourcing supply chain that is not accountable, the buyer may face a dilemma. On the one hand, the buyer cannot induce first-best by using a simple wholesale contract with only one lever of payment. On the other hand, adding a second lever of penalty may lead to a contract that is impractical because the buyer may have to pay the suppliers more than the retail price when the end product is non-defective.[3] However, if the supply chain is accountable, the buyer no longer faces such a dilemma. Accountability eliminates over-penalization for the suppliers when the end product is defective and guarantees cash flow feasibility for the buyer when the end product is non-defective. This suggests that because of the enabled accountability, the emerging blockchain technology can help improve the implementability of contracts in multi-sourcing supply chains so that first-best is more readily achievable.

Example 1 We now provide an example with the logarithmic quality cost functions following Corbett and DeCroix (2001) and Chao et al. (2009). Suppose $n = 2$ and $C_i(q_i) = -\alpha_i\big[\ln(1 - q_i) + q_i\big]$, where $\alpha_i > 0$ for $i \in \{1, 2\}$. The expressions of the equilibrium qualities as well as the optimal contracts with and without accountability are as follows:[4] for any $i \in \{1, 2\}$,

[3]It is easy to see that if a cash flow feasibility constraint (i.e., $\sum_{i=1}^{n} w_i \leqslant p$) is explicitly incorporated into the model, the buyer will not be able to induce the first-best if $p < \bar{p}$ in the case without accountability.

[4]In this example, Assumptions 1 and 2 can be satisfied as long as $(p+l)(1-q_i^*)^2 < \alpha_i < p+l$ for $i \in \{1, 2\}$ and $(p+l)(p+l+\alpha_1)^2 > \alpha_1(p+l+\alpha_2)^2$.

$$q_i^{\mathrm{N}\dagger} = q_i^{\mathrm{A}\dagger} = q_i^* = \frac{(p+l)^2 - \alpha_i \alpha_{3-i}}{(p+l)(p+l+\alpha_i)},$$

$$w_i^{\mathrm{N}\dagger} = \alpha_i \ln\left[\frac{(p+l)(p+l+\alpha_i)}{\alpha_i(p+l+\alpha_{3-i})}\right] + \frac{\alpha_{3-i}(p+l+\alpha_i)}{p+l+\alpha_{3-i}},$$

$$t_i^{\mathrm{N}\dagger} = \alpha_i \ln\left[\frac{(p+l)(p+l+\alpha_i)}{\alpha_i(p+l+\alpha_{3-i})}\right] - \frac{(p+l)^2 - \alpha_i \alpha_{3-i}}{p+l+\alpha_{3-i}},$$

$$w_i^{\mathrm{A}\dagger} = \alpha_i \ln\left[\frac{(p+l)(p+l+\alpha_i)}{\alpha_i(p+l+\alpha_{3-i})}\right],$$

$$t_i^{\mathrm{A}\dagger} = \alpha_i \ln\left[\frac{(p+l)(p+l+\alpha_i)}{\alpha_i(p+l+\alpha_{3-i})}\right] - \frac{(p+l)^2 - \alpha_i \alpha_{3-i}}{p+l+\alpha_{3-i}}.$$

Recall that with general quality cost functions, Proposition 3 has identified a threshold market price level, $\bar{p}$, below which the buyer's cash flow is infeasible in the case without accountability. With logarithmic quality cost functions and $n = 2$ in this example, we can further prove that $p < \bar{p}$ is a sufficient and necessary condition for $w_1^{\mathrm{N}\dagger} + w_2^{\mathrm{N}\dagger} > p$.

Figure 1 displays some numerical results with logarithmic quality cost functions. The solid curve corresponds to the threshold $\bar{p}$; we can also see that $\bar{p}$ increases as l increases, indicating that when the buyer faces a greater loss from the market if the end product is defective, the cash flow infeasibility problem will be severer in the case without accountability. Furthermore, Fig. 1 also incorporates the possibility that if the retail price p is too small, the buyer's first-best profit will become negative.

4.2 Impact of Supply Chain Complexity

We now study how the complexity of the multi-sourcing supply chain would impact the value of accountability. Note that the number of suppliers in the supply chain can measure the complexity of the multi-sourcing supply chain. The following proposition summarizes the main results.

Proposition 4 (Impact of Supply Chain Complexity) *The value of accountability increases as the multi-sourcing supply chain becomes more complex. In particular, if suppliers are symmetric (i.e., they have the same quality cost function), as the number of suppliers n increases, the buyer's cash flow is more likely to be infeasible without accountability. By contrast, with accountability, the buyer's cash flow is always feasible.*

Proposition 4 shows that in the multi-sourcing supply chain, the value of accountability is *strengthened* as the supply chain becomes more complicated. As stated by Proposition 4, in the multi-sourcing supply chain without accountability,

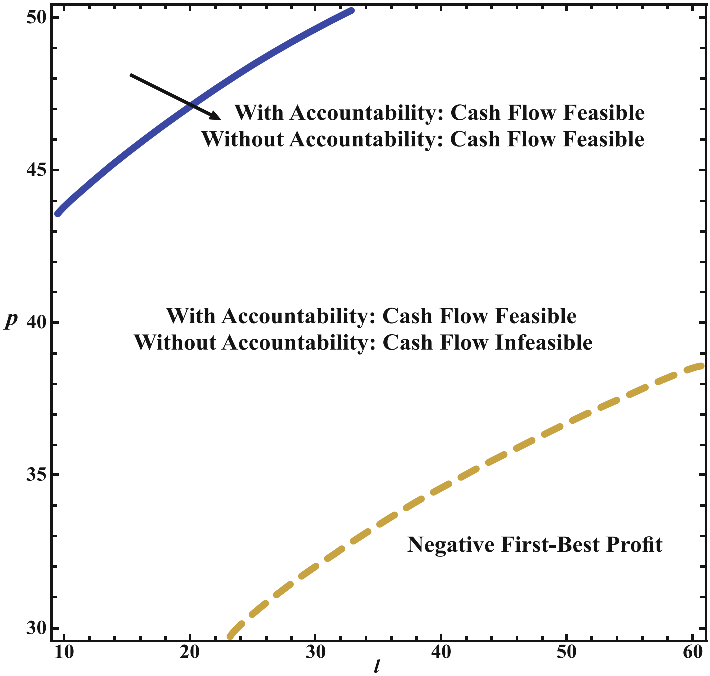

Fig. 1 Impact of blockchain-driven accountability on buyer's cash flow feasibility ($n = 2, \alpha_1 = 5, \alpha_2 = 10$)

as the number of suppliers increases, the buyer's cash flow is more likely to be infeasible. This is because as the number of suppliers increases, the probability that the end product is non-defective decreases. This implies that the suppliers will be over-penalized to an even greater extent. Thus, in order to induce the suppliers to participate, the buyer would need to offer even higher wholesale prices, and hence the cash flow is more likely to be infeasible. However, with accountability, as Proposition 4 verifies, the buyer's cash flow is always feasible. Therefore, in a more complex multi-sourcing supply chain, accountability is more likely to create value by improving the cash flow feasibility for the buyer.

4.3 Impact of Suppliers' Limited Liability

In our main model, we have not put any restriction on the amount that the suppliers can be penalized. In this subsection, following the literature (e.g., Chu & Lai, 2013; Dong et al., 2016), we consider a model extension with limited liability constraints for the suppliers. The constraint for supplier i, $t_i \geqslant -b$, specifies that supplier i can only be penalized up to a certain amount $b \geqslant 0$. With suppliers' limited liability constraints, we find that the first-best is not always achievable. In particular, when b is small enough, the suppliers' limited liability constraints would be binding in equilibrium, in which case the first-best is not achieved. Moreover, we also find that the suppliers' limited liability constraints *weaken* the value of accountability in the multi-sourcing supply chain, which we explain below in detail.

In the multi-sourcing supply chain without accountability, the suppliers' limited liability constraints would mitigate the over-penalization by the buyer. Since the buyer now charges smaller penalties to the suppliers, he can induce the suppliers to participate using lower wholesale prices. Thus, when suppliers face limited liability constraints, the buyer is less likely to face the issue of cash flow infeasibility. Since the buyer's cash flow issue is mitigated without accountability, the value of accountability in ensuring cash flow feasibility for the buyer would be weakened in the multi-sourcing supply chain. Figure 2 displays some numerical results using the same example as Fig. 1. In Fig. 2, the region where the buyer's cash flow is infeasible without accountability shrinks because of the suppliers' limited liability constraints. Also as Fig. 2 shows, incorporating the suppliers' limited liability constraint may also result in a negative second-best profit for the buyer in some cases.

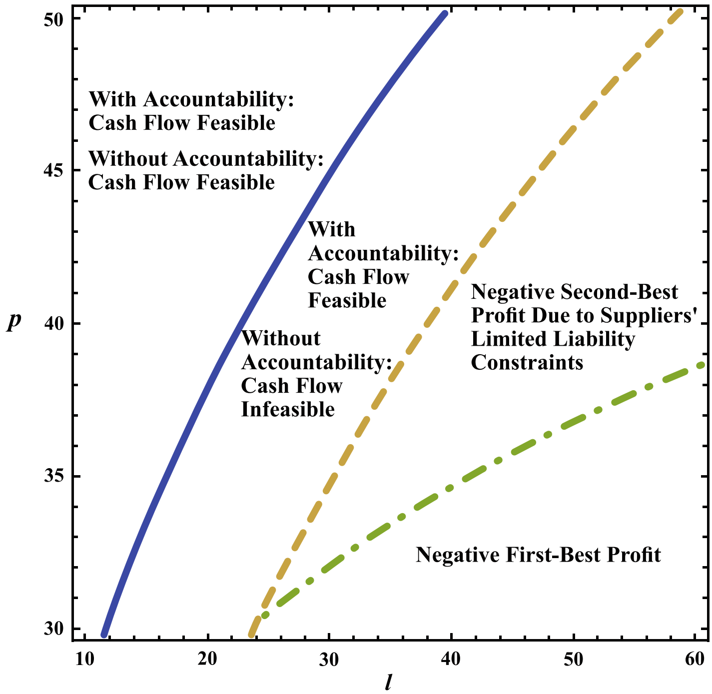

Fig. 2 Impact of supplier's limited liability constraints ($n = 2, \alpha_1 = 5, \alpha_2 = 10, b = 5$)

5 Conclusion

Traditionally, supply chains in certain industries (such as agri-food and pharmaceutical) may face the issue of inability to hold the suppliers accountable for their own faults when defects occur in the end products. Even if the buying firm knows who is the failure-causing supplier(s), he may not be able to penalize only that supplier without the accountability of the entire supply chain. By recording information in a transparent, verifiable, and permanent way, the recently emerged blockchain technology can naturally improve the accountability of such supply chains. As an essential application of blockchain, smart contracts can enable the buying firm to automatically execute the payment and penalization based on the pre-defined protocol and the information recorded in the blockchain.

In this work, we study the value of accountability in multi-sourcing supply chains. We find that a critical value of accountability in the multi-sourcing supply chain is that it guarantees cash flow feasibility for the buyer, so that the supply chain can more readily achieve the first-best through (smart) contracting. In particular, if the multi-sourcing supply chain is not accountable, the inability to make the failure-causing suppliers liable for their own faults results in the buyer over-penalizing all suppliers when the end product is defective and offering excessively high wholesale prices to all suppliers when the end product is non-defective. In fact, the buyer's total wholesale price payment to all suppliers can exceed the retail price he earns when the end product is actually non-defective, leading to an infeasible cash flow for the buyer. However, if the supply chain is accountable, the buyer is able to penalize only the supplier(s) who is (are) identified to be defective. Since the buyer no longer over-penalizes the suppliers, he offers lower wholesale prices to the suppliers, the sum of which never exceeds the retail price. Moreover, we further find that the value of accountability in guaranteeing cash flow feasibility for the buyer is strengthened as the multi-sourcing supply chain becomes more complicated, while weakened when suppliers face limited liability constraints.

Appendix: Proofs

Proof of Lemma 1 The first-order conditions of $\Pi_{\rm B}(q_1, q_2, \dots, q_n)$ are

$$\left.\frac{\partial \Pi_{\rm B}(q_1, q_2, \dots, q_n)}{\partial q_i}\right|_{(q_1=q_1^*, q_2=q_2^*, \dots, q_n=q_n^*)} = (p+l)\prod_{j=1, j\neq i}^{n} q_j^* - C_i'(q_i^*) = 0, \tag{5}$$

where $i \in \{1, 2, \dots, n\}$. Taking the second-order derivatives of $\Pi_{\rm B}(q_1, q_2, \dots, q_n)$ w.r.t. q_i and q_j yields $\frac{\partial^2 \Pi_{\rm B}(q_1, q_2, \dots, q_n)}{\partial q_i^2} = -C_i''(q_i)$ and $\frac{\partial^2 \Pi_{\rm B}(q_1, q_2, \dots, q_n)}{\partial q_i \partial q_j} = (p + l)\prod_{k=1, k\neq i,j}^{n} q_k$, where $i, j \in \{1, 2, \dots, n\}$ and $i \neq j$. The Hessian of $\Pi_{\rm B}(q_1, q_2, \dots, q_n)$ is:

$$H(\mathbf{q}) = (p+l)\begin{bmatrix} -\frac{C_1''(q_1)}{p+l} & \prod_{k=1,k\neq 1,2}^{n} q_k & \cdots & \prod_{k=1,k\neq 1,n}^{n} q_k \\ \prod_{k=1,k\neq 1,2}^{n} q_k & -\frac{C_2''(q_2)}{p+l} & \cdots & \prod_{k=1,k\neq 2,n}^{n} q_k \\ \vdots & \vdots & \ddots & \vdots \\ \prod_{k=1,k\neq 1,n}^{n} q_k & \prod_{k=1,k\neq 2,n}^{n} q_k & \cdots & -\frac{C_n''(q_n)}{p+l} \end{bmatrix}.$$

By Assumption 1, the solution to (5) is either $(0, 0, \ldots, 0)$ or an interior point $(q_1^*, q_2^*, \ldots, q_n^*)$, where $q_i^* \in (0, 1)$ for $i \in \{1, 2, \ldots, n\}$. Specifically, it is impossible for any q_i^* to be 1 since $C_i'(1) > p+l$, and it is also impossible for some but not all q_i^* equals 0 to be the solution because the n equations in (5) cannot be satisfied at the same time. In addition, Assumption 1 also guarantees that $(0, 0, \ldots, 0)$ cannot be a local maximum, because the Hessian of $\Pi_{\mathrm{B}}(q_1, q_2, \ldots, q_n)$ is not negative definite at $(0, 0, \ldots, 0)$ due to $0 \leqslant C_i''(0) < p+l$.

Next, we prove the existence of an interior solution $(q_1^*, q_2^*, \ldots, q_n^*)$ satisfying (2) by induction. First, consider $n = 2$. We can view q_1^* as a function of q_2^* and view q_2^* as a function of q_1^*, i.e., $q_2^* = \frac{C_1'(q_1^*)}{p+l}$ and $q_1^* = \frac{C_2'(q_2^*)}{p+l}$. Thus, to prove the existence of an interior solution (q_1^*, q_2^*) is equivalent to showing the existence of an interior intersection point of the following two lines: $q_2 = F_1(q_1) = \frac{C_1'(q_1)}{p+l}$ and $q_2 = F_2(q_1)$, where $q_2 = F_2(q_1)$ is the inverse function of $q_1 = \frac{C_2'(q_2)}{p+l}$. We have $F_1'(q_1) = \frac{C_1''(q_1)}{p+l} > 0$ and $F_2'(q_1) = \frac{p+l}{C_2''(q_2)} > 0$ for any $q_1 > 0$. By Assumption 1, we have $F_1(0) = F_2(0) = 0$, $F_1'(0) = \frac{C_1''(0)}{p+l} < 1$, and $F_2'(0) = \frac{p+l}{C_2''(0)} > 1$. Thus, there must exist an infinitesimal $\epsilon > 0$ such that $F_1(\epsilon) < F_2(\epsilon)$. On the other hand, since $\frac{C_2'(1)}{p+l} > 1$ by Assumption 1, in order for $F_2(q_1) = 1$, q_1 must be strictly greater than 1. Then, we have $F_2(1) < 1$ because $F_2'(q_1) > 0$ for any $q_1 > 0$. Thus, $F_1(1) = \frac{C_1'(1)}{p+l} > 1 > F_2(1)$. Hence, following the monotonicity of $F_1(q_1)$ and $F_2(q_1)$, they must have an intersection within $(0, 1)$. Therefore, the existence of an interior solution in the case of $n = 2$ is proved. Then, consider $n > 2$. Suppose an interior solution $(q_1^*, q_2^*, \ldots, q_n^*)$ exists in the case of $n > 2$, which satisfies

$$(p+l)\prod_{j=1, j\neq i}^{n} q_j^* = C_i'(q_i^*), \quad i \in \{1, 2, \ldots, n\}. \tag{6}$$

Then, regarding the case of $n+1$, considering $\tilde{q}_i(q_{n+1})$ as functions of q_{n+1}, we have

$$(p+l)\prod_{j=1, j\neq i}^{n} \tilde{q}_j(q_{n+1})q_{n+1} = C_i'(\tilde{q}_i(q_{n+1})), \quad i \in \{1, 2, \ldots, n\}, \tag{7}$$

and

$$(p+l)\prod_{j=1}^{n}\tilde{q}_j(q_{n+1}) = C'_{n+1}(q_{n+1}). \tag{8}$$

Hence, the interior solution characterized in (6) can be viewed as a special case of (7), where q_{n+1} is considered as an exogenous parameter and $q_{n+1} = 1$. Since $(q_1^*, q_2^*, \ldots, q_n^*)$ characterized in (6) is an interior solution in the case of $n > 2$, we have $\tilde{q}_i(1) = q_i^* < 1$ for $i \in \{1, 2, \ldots, n\}$. Then, by (7), we have

$$\begin{aligned}(p+l)\Bigg[\prod_{j=1,j\neq i}^{n}\tilde{q}_j(q_{n+1}) + \sum_{k=1,k\neq i}^{n}\prod_{j=1,j\neq i,k}^{n}\tilde{q}_j(q_{n+1})\frac{d\tilde{q}_k(q_{n+1})}{dq_{n+1}}q_{n+1}\Bigg]\\ = C''_i(\tilde{q}_i(q_{n+1}))\frac{d\tilde{q}_i(q_{n+1})}{dq_{n+1}},\end{aligned} \tag{9}$$

for $i \in \{1, 2, \ldots, n\}$. By (8), we have

$$(p+l)\sum_{k=1}^{n}\left[\prod_{j=1,j\neq k}^{n}\tilde{q}_j(q_{n+1})\frac{d\tilde{q}_k(q_{n+1})}{dq_{n+1}}\right] = C''_{n+1}(q_{n+1}). \tag{10}$$

Solving (9) and (10), we have $\frac{d\tilde{q}_i(q_{n+1})}{dq_{n+1}} = \frac{C''_{n+1}(q_{n+1})q_{n+1}+(p+l)\prod_{j=1}^{n}\tilde{q}_j(q_{n+1})}{C''_i(\tilde{q}_i(q_{n+1}))\tilde{q}_i(q_{n+1})+(p+l)\prod_{j=1,j\neq i}^{n}\tilde{q}_j(q_{n+1})q_{n+1}} > 0$, for any $q_{n+1} > 0$ and for $i \in \{1, 2, \ldots, n\}$. Moreover, from (7), we can obtain $\tilde{q}_i(0) = 0$. Combining with $\frac{d\tilde{q}_i(q_{n+1})}{dq_{n+1}} > 0$, we have $0 = \tilde{q}_i(0) < \tilde{q}_i(q_{n+1}) < \tilde{q}_i(1) < 1$, for any $q_{n+1} \in (0, 1)$ and for $i \in \{1, 2, \ldots, n\}$. Then, combining with Assumption 1, we have $C'_{n+1}(0) = 0 < (p+l)\prod_{i=1}^{n}\tilde{q}_i(q_{n+1}) < p+l < C'_{n+1}(1)$. Hence, there must exist an interior $q_{n+1}^* \in (0, 1)$ that satisfies (8). Moreover, according to (7), $\tilde{q}_i(q_{n+1}^*) \in (0, 1)$ for $i \in \{1, 2, \ldots, n\}$. Therefore, the existence of an interior solution in the case of $n+1$ is proved.

Next, we show that the interior solution $(q_1^*, q_2^*, \ldots, q_n^*)$ is the unique global maximum. In particular, we will prove that the sufficient condition of the local maximum is able to guarantee the unique global maximum, the underlying idea of which was used previously by Petruzzi and Dada (1999) and Aydin and Porteus (2008). First, we show that $(q_1^*, q_2^*, \ldots, q_n^*)$ is a strict local maximum. Denote by $H(\mathbf{q})_{ij}$ the entry in row i and column j of the matrix $H(\mathbf{q})$, where $i, j \in \{1, 2, \ldots, n\}$. Due to Assumption 2, we have $|H(\mathbf{q}^*)_{ii}| = C''_i(q_i^*) > (n-1)(p+l) > \sum_{j=1,j\neq i}^{n}\left[(p+l)\prod_{k=1,k\neq i,j}^{n}q_k^*\right] = \sum_{j=1,j\neq i}^{n}\left|H(\mathbf{q}^*)_{ij}\right|$, for $i \in \{1, 2, \ldots, n\}$, which implies that $H(\mathbf{q}^*)$ is a diagonally dominant matrix. Besides, since $H(\mathbf{q}^*)$ is also a symmetric real matrix with negative diagonal entries, we know that the Hessian of $\Pi_{\mathrm{B}}(q_1, q_2, \ldots, q_n)$ is negative definite in the neighborhood of any $\mathbf{q}^* = (q_1^*, q_2^*, \ldots, q_n^*)$ satisfying (2), where $q_i^* \in (0, 1)$ for $i \in \{1, 2, \ldots, n\}$. Thus, any interior stationary point is a strict local maximum. Then, we show that the interior stationary point is the unique global maximum.

Suppose now that there exist more than one, say two, interior stationary points for the function $\Pi_{\mathrm{B}}(q_1, q_2, \ldots, q_n)$. Because both points need to be local maxima, the function should also have an interior local minimum somewhere in between, which is a contradiction to the result that all interior stationary points are local maxima. Consequently, we can conclude that there exists only one stationary point $(q_1^*, q_2^*, \ldots, q_n^*)$ that satisfies (2), which is the unique local maximum, and thus, the unique global maximum.

Next, suppose suppliers are symmetric. By (2), considering $q_i^*(n)$ as functions of n, we have $(p+l)(q_i^*(n))^{n-1} = C_i'(q_i^*(n))$. Thus we have $\frac{dq_i^*(n)}{dn} = \frac{(p+l)(q_i^*(n))^{n-1}\ln q_i^*(n)}{C_i''(q_i^*(n))-(n-1)(p+l)(q_i^*(n))^{n-2}} < 0$, where the inequality is due to Assumption 2.

Finally, we prove the last part of the lemma. By (2), we have $C_i'(q_i^*)q_i^* = C_j'(q_j^*)q_j^*$. Assuming $C_i'(q) \leqslant C_j'(q)$ for all $q \in (0, 1)$, we can prove $q_i^* \geqslant q_j^*$ by contradiction. Specifically, suppose $q_i^* < q_j^*$, we have $C_i'(q_i^*)q_i^* \leqslant C_j'(q_i^*)q_i^* < C_j'(q_j^*)q_i^* < C_j'(q_j^*)q_j^*$, which is contradictory to $C_i'(q_i^*)q_i^* = C_j'(q_j^*)q_j^*$. Thus, $q_i^* \geqslant q_j^*$ holds. □

Proof of Proposition 1 We first derive the suppliers' optimal quality decisions. Given contract (w_i, t_i), supplier $i \in \{1, 2, \ldots, n\}$ chooses his quality q_i to maximize his expected profit $\Pi_{\mathrm{S}_i}(q_i|w_i, t_i, q_{-i})$. The first-order condition is $\frac{d\Pi_{\mathrm{S}_i}(q_i|w_i,t_i,q_{-i})}{dq_i}\Big|_{q_i=\tilde{q}_i(w_i,t_i,q_{-i})} = (w_i - t_i)\prod_{j=1,j\neq i}^n q_j - C_i'(\tilde{q}_i(w_i, t_i, q_{-i})) = 0.$ Taking the second-order derivative of $\Pi_{\mathrm{S}_i}(q_i|w_i, t_i, q_{-i})$ w.r.t. q_i yields $\frac{d^2\Pi_{\mathrm{S}_i}(q_i|w_i,t_i,q_{-i})}{dq_i^2} = -C_i''(q_i) < 0$. Thereby, the solution of the first-order condition is supplier i's optimal quality, in response to contract (w_i, t_i). The n suppliers' best response functions form a system of equations, solving which yields the suppliers' equilibrium quality decisions, $\tilde{q}_i(\mathbf{w}, \mathbf{t})$, as functions of the buyer's contract decisions $(\mathbf{w}, \mathbf{t})$:

$$(w_i - t_i)\prod_{j=1,j\neq i}^{n} \tilde{q}_j(\mathbf{w}, \mathbf{t}) = C_i'(\tilde{q}_i(\mathbf{w}, \mathbf{t})). \tag{11}$$

Next, consider the buyer's problem. Since the IR_i constraint must be binding in equilibrium,[5] after plugging $\tilde{q}_i(\mathbf{w}, \mathbf{t})$ into $\Pi_{\mathrm{S}_i}(q_i|w_i, t_i, q_{-i})$, we have

$$C_i'(\tilde{q}_i(\mathbf{w}, \mathbf{t}))\tilde{q}_i(\mathbf{w}, \mathbf{t}) + t_i - C_i(\tilde{q}_i(\mathbf{w}, \mathbf{t})) = 0. \tag{12}$$

From the system of equations formed by the n binding IR_i constraints, we can obtain $\tilde{t}_i(\mathbf{w})$ as functions of $\mathbf{w}$. Plugging $\tilde{t}_i(\mathbf{w})$ into (11), $\tilde{q}_i(\mathbf{w})$ reduces to a function of only $\mathbf{w}$ as well. Moreover, from (12), we have

[5]The result that the IR_i constraint is binding in equilibrium can be proved by assigning Lagrangian multipliers λ_i and μ_i to the IR_i and IC_i constraints, respectively. The Lagrangian of (3) is

$$\tilde{t}_i(\mathbf{w}) = C_i(\tilde{q}_i(\mathbf{w})) - C_i'(\tilde{q}_i(\mathbf{w}))\tilde{q}_i(\mathbf{w}). \tag{13}$$

Then, with $\tilde{t}_i(\mathbf{w})$ and $\tilde{q}_i(\mathbf{w})$ plugged into (3), the buyer's problem becomes

$$\begin{aligned}\max_{\mathbf{w}} \ \Pi_{\mathrm{B}}(\mathbf{w}) &= (p+l)\prod_{i=1}^{n}\tilde{q}_i(\mathbf{w}) - l - \sum_{i=1}^{n} C_i'(\tilde{q}_i(\mathbf{w}))\tilde{q}_i(\mathbf{w}) - \sum_{i=1}^{n} t_i \\ &= (p+l)\prod_{i=1}^{n}\tilde{q}_i(\mathbf{w}) - l - \sum_{i=1}^{n} C_i(\tilde{q}_i(\mathbf{w})),\end{aligned} \tag{14}$$

where the first step follows from (11) and the second step follows from (13).

We now analyze the buyer's optimal contract decisions. Since $\mathbf{w}$ affects $\Pi_{\mathrm{B}}(\mathbf{w})$ through $\tilde{q}_i(\mathbf{w})$, optimizing $\mathbf{w}$ is equivalent to optimizing $\mathbf{q}$. Further, notice that after rewriting $\Pi_{\mathrm{B}}(\mathbf{w})$ as $\Pi_{\mathrm{B}}(\mathbf{q})$, $\Pi_{\mathrm{B}}(\mathbf{q})$ is the same as the buyer's profit function in the first-best problem. Therefore, we have $q_i^{\mathrm{N}\dagger} = q_i^*$ and $\Pi_{\mathrm{B}}^{\mathrm{N}\dagger} = \Pi_{\mathrm{B}}^*$, and the supply chain achieves the first-best in equilibrium. Then, plugging $q_i^{\mathrm{N}\dagger}$ into (13), we have

$$t_i^{\mathrm{N}\dagger} = C_i(q_i^{\mathrm{N}\dagger}) - C_i'(q_i^{\mathrm{N}\dagger})q_i^{\mathrm{N}\dagger} = C_i(q_i^{\mathrm{N}\dagger}) - (p+l)\prod_{j=1}^{n} q_j^{\mathrm{N}\dagger}, \tag{15}$$

where the last step follows from (2). Moreover,

$$w_i^{\mathrm{N}\dagger} = \frac{C_i'(q_i^{\mathrm{N}\dagger})}{\prod_{j=1,j\neq i}^{n} q_j^{\mathrm{N}\dagger}} + t_i^{\mathrm{N}\dagger} = C_i(q_i^{\mathrm{N}\dagger}) + (p+l)\left(1 - \prod_{j=1}^{n} q_j^{\mathrm{N}\dagger}\right),$$

where the first step follows from (11), and the second step follows from (15). Finally, it is easy to see that $w_i^{\mathrm{N}\dagger} > 0$. Regarding $t_i^{\mathrm{N}\dagger}$, notice that $\Pi_{\mathrm{B}}^{\mathrm{N}\dagger} = (p+l)\prod_{i=1}^{n} q_i^{\mathrm{N}\dagger} - l - \sum_{i=1}^{n} C_i(q_i^{\mathrm{N}\dagger}) = -l - t_i^{\mathrm{N}\dagger} - \sum_{j=1,j\neq i}^{n} C_j(q_j^{\mathrm{N}\dagger})$, where the last step follows from (15). Since $\Pi_{\mathrm{B}}^{\mathrm{N}\dagger} \geqslant 0$, we have $t_i^{\mathrm{N}\dagger} < 0$ always holds. □

$$\begin{aligned}L(\mathbf{w},\mathbf{t},\mathbf{q}) = {} & p\prod_{i=1}^{n} q_i - l\left(1-\prod_{i=1}^{n} q_i\right) - \left(\sum_{i=1}^{n} w_i\right)\prod_{i=1}^{n} q_i - \left(\sum_{i=1}^{n} t_i\right)\left(1-\prod_{i=1}^{n} q_i\right) \\ & + \sum_{i=1}^{n}\lambda_i\left[w_i q_i \prod_{j=1,j\neq i}^{n} q_j + t_i\left(1 - q_i\prod_{j=1,j\neq i}^{n} q_j\right) - C_i(q_i)\right] \\ & + \sum_{i=1}^{n}\mu_i\left[(w_i - t_i)\prod_{j=1,j\neq i}^{n} q_j - C_i'(q_i)\right].\end{aligned}$$

The first-order conditions w.r.t. w_i and t_i lead to $\lambda_i = 1$ and $\mu_i = 0$, implying that the IR_i constraint is binding in equilibrium.

Proof of Proposition 2 We first derive the suppliers' optimal quality decisions. Given contract (w_i, t_i), supplier $i \in \{1, 2, \ldots, n\}$ chooses his quality q_i to maximize his expected profit $\Pi_{S_i}(q_i|w_i, t_i)$. The first-order condition is $\left.\frac{d\Pi_{S_i}(q_i|w_i,t_i)}{dq_i}\right|_{q_i=\tilde{q}_i(w_i,t_i)} = w_i - t_i - C_i'(\tilde{q}_i(w_i, t_i)) = 0$. Taking the second-order derivative of $\Pi_{S_i}(q_i|w_i, t_i)$ w.r.t. q_i yields $\frac{d^2\Pi_{S_i}(q_i|w_i,t_i)}{dq_i^2} = -C_i''(q_i) < 0$. Thereby, the solution of the first-order condition is supplier i's optimal quality, in response to contract (w_i, t_i). The n suppliers' best response functions form a system of equations, corresponding to the suppliers' equilibrium quality decisions, $\tilde{q}_i(w_i, t_i)$, as functions of the buyer's contract decisions (w_i, t_i):

$$w_i - t_i = C_i'(\tilde{q}_i(w_i, t_i)). \tag{16}$$

Next, consider the buyer's problem. Since the IR_i constraint must be binding in equilibrium, after plugging $\tilde{q}_i(w_i, t_i)$ into $\Pi_{S_i}(q_i|w_i, t_i)$, we have

$$C_i'(\tilde{q}_i(w_i, t_i))\tilde{q}_i(w_i, t_i) + t_i - C_i(\tilde{q}_i(w_i, t_i)) = 0. \tag{17}$$

From the system of equations formed by the n binding IR_i constraints, we can obtain $\tilde{t}_i(w_i)$ as functions of w_i. Plugging $\tilde{t}_i(w_i)$ into (16), $\tilde{q}_i(w_i)$ reduces to a function of only w_i as well. Moreover, from (17), we have

$$\tilde{t}_i(w_i) = C_i(\tilde{q}_i(w_i)) - C_i'(\tilde{q}_i(w_i))\tilde{q}_i(w_i). \tag{18}$$

Then, with $\tilde{t}_i(w_i)$ and $\tilde{q}_i(w_i)$ plugged into (4), the buyer's problem becomes

$$\begin{aligned}\max_{\mathbf{w}}\ \Pi_B(\mathbf{w}) &= (p+l)\prod_{i=1}^{n}\tilde{q}_i(w_i) - l - \sum_{i=1}^{n} C_i'(\tilde{q}_i(w_i))\tilde{q}_i(w_i) - \sum_{i=1}^{n} t_i \\ &= (p+l)\prod_{i=1}^{n}\tilde{q}_i(w_i) - l - \sum_{i=1}^{n} C_i(\tilde{q}_i(w_i)),\end{aligned} \tag{19}$$

where the first step follows from (16) and the second step follows from (18).

We now analyze the buyer's optimal contract decisions. Since $\mathbf{w}$ affects $\Pi_B(\mathbf{w})$ through $\tilde{q}_i(w_i)$, optimizing $\mathbf{w}$ is equivalent to optimizing $\mathbf{q}$. Further, notice that after rewriting $\Pi_B(\mathbf{w})$ as $\Pi_B(\mathbf{q})$, $\Pi_B(\mathbf{q})$ is the same as the buyer's profit function in the first-best problem. Therefore, we have $q_i^{A\dagger} = q_i^*$ and $\Pi_B^{A\dagger} = \Pi_B^*$, and the supply chain achieves the first-best in equilibrium. Then, plugging $q_i^{A\dagger}$ into (18), we have

$$t_i^{A\dagger} = C_i(q_i^{A\dagger}) - C_i'(q_i^{A\dagger})q_i^{A\dagger} = C_i(q_i^{A\dagger}) - (p+l)\prod_{j=1}^{n} q_j^{A\dagger}, \tag{20}$$

where the last step follows from (2). Moreover,

$$w_i^{A\dagger} = C_i'(q_i^{A\dagger}) + t_i^{A\dagger} = C_i(q_i^{A\dagger}) + (p+l) \prod_{j=1, j\neq i}^{n} q_j^{A\dagger}\left(1 - q_i^{A\dagger}\right),$$

where the first step follows from (16), and the second step follows from (20). Finally, it is easy to see that $w_i^{A\dagger} > 0$. Regarding $t_i^{A\dagger}$, notice that $\Pi_B^{A\dagger} = (p+l)\prod_{i=1}^{n} q_i^{A\dagger} - l - \sum_{i=1}^{n} C_i(q_i^{A\dagger}) = -l - t_i^{A\dagger} - \sum_{j=1, j\neq i}^{n} C_j(q_j^{A\dagger})$, where the last step follows from (20). Since $\Pi_B^{A\dagger} \geqslant 0$, we have $t_i^{A\dagger} < 0$ always holds. □

Proof of Corollary 1 Parts (1) and (2) of the corollary follow from comparing the equilibria characterized in Propositions 1 and 2. To show part (3), notice that

$$w_i^{N\dagger} - w_i^{A\dagger} = (p+l)\left(1 - q_j^* \prod_{k=1, k\neq i,j}^{n} q_k^*\right),$$

$$w_j^{N\dagger} - w_j^{A\dagger} = (p+l)\left(1 - q_i^* \prod_{k=1, k\neq i,j}^{n} q_k^*\right).$$

By Lemma 1, if $C_i'(q) \leqslant C_j'(q)$ for all $q \in (0, 1)$, we have $q_i^* \geqslant q_j^*$, and thereby, we have $w_i^{N\dagger} - w_i^{A\dagger} \geqslant w_j^{N\dagger} - w_j^{A\dagger}$. □

Proof of Propositions 3 and 4 First, consider the case without accountability. We view $w_i^{N\dagger}(p)$ as functions of p, and define $G(p) \equiv \sum_{i=1}^{n} w_i^{N\dagger}(p) - p$. Based on the characterization of $w_i^{N\dagger}$ in Proposition 1, when $p \to 0$, we have

$$\lim_{p\to 0} G(p) = \sum_{i=1}^{n} C_i(q_i^{N\dagger}) + n \cdot l\left(1 - \prod_{i=1}^{n} q_i^{N\dagger}\right) > 0.$$

Hence, there must exist a threshold $\bar{p} > 0$ such that $G(p) > 0$ for $0 < p < \bar{p}$. In other words, there exists a threshold $\bar{p} > 0$ such that $\sum_{i=1}^{n} w_i^{N\dagger} > p$ if $0 < p < \bar{p}$. Then, suppose suppliers are symmetric. We view $w_i^{N\dagger}(n)$ as functions of n. Taking the first-order derivatives of $w_i^{N\dagger}(n)$ w.r.t. n yields

$$\frac{dw_i^{N\dagger}(n)}{dn} = -(n-1)(p+l)(q_i^*(n))^{n-1}\frac{dq_i^*(n)}{dn} - (p+l)(q_i^*(n))^n \ln q_i^*(n) > 0,$$

where the inequality follows from $q_i^*(n) \in (0, 1)$ and $\frac{dq_i^*(n)}{dn} < 0$ in Lemma 1. Since $w_i^{N\dagger}(n)$ increases in n, we have $\sum_{i=1}^{n} w_i^{N\dagger}(n)$ increases in n as well. Hence, we know that it is more likely for $\sum_{i=1}^{n} w_i^{N\dagger} > p$ to occur with the increase of n.

Next, consider the case with accountability. Based on the characterization of $w_i^{\mathrm{A}\dagger}$ in Proposition 2, we have

$$\sum_{i=1}^{n} w_i^{\mathrm{A}\dagger} < p \Leftrightarrow \sum_{i=1}^{n} C_i(q_i^{\mathrm{A}\dagger}) + \sum_{i=1}^{n}\left[(p+l)\left(1-q_i^{\mathrm{A}\dagger}\right)\prod_{j=1, j\neq i}^{n} q_j^{\mathrm{A}\dagger}\right] < p$$

$$\Leftrightarrow \Pi_{\mathrm{B}}^{\mathrm{A}\dagger} > -(p+l)\left[1+(n-1)\prod_{i=1}^{n} q_i^{\mathrm{A}\dagger} - \sum_{i=1}^{n}\prod_{j=1, j\neq i}^{n} q_j^{\mathrm{A}\dagger}\right].$$

Define $F(n) \equiv 1+(n-1)\prod_{i=1}^{n} q_i^{\mathrm{A}\dagger} - \sum_{i=1}^{n}\prod_{j=1, j\neq i}^{n} q_j^{\mathrm{A}\dagger}$. If we can show $F(n) > 0$ for any $n \geqslant 2$, then $\sum_{i=1}^{n} w_i^{\mathrm{A}\dagger} < p$ always holds since $\Pi_{\mathrm{B}}^{\mathrm{A}\dagger} = \Pi_{\mathrm{B}}(\mathbf{q}^*) \geqslant 0$. We now prove $F(n) > 0$ by induction. Consider $n = 2$, we have $F(2) = \left(1-q_1^{\mathrm{A}\dagger}\right)\left(1-q_2^{\mathrm{A}\dagger}\right) > 0$. Suppose $F(n) > 0$ holds in the case of $n > 2$. Then, regarding the case of $n+1$, since

$$F(n+1) - F(n) = \left(1-q_{n+1}^{\mathrm{A}\dagger}\right)\sum_{i=1}^{n}\left[\left(1-q_i^{\mathrm{A}\dagger}\right)\prod_{j=1, j\neq i}^{n} q_j^{\mathrm{A}\dagger}\right] > 0,$$

we have $F(n+1) > 0$, and thus, $F(n) > 0$ for any $n \geqslant 2$. Hence, $\sum_{i=1}^{n} w_i^{\mathrm{A}\dagger} < p$ always holds. □

References

Ang, E., Iancu, D., & Swinney, R. (2017). Disruption risk and optimal sourcing in multitier supply networks. *Management Science, 63*(8), 2397–2419.

Aydin, G., & Porteus, E. (2008). Joint inventory and pricing decisions for an assortment. *Operations Research, 56*(5), 1247–1255.

Babich, V., & Hilary, G. (2020). Distributed ledgers and operations: What operations management researchers should know about Blockchain technology. *Manufacturing & Service Operations Management, 22*(2), 223–240.

Babich, V., & Tang, C. (2012). Managing opportunistic supplier product adulteration: Deferred payments, inspection, and combined mechanisms. *Manufacturing & Service Operations Management, 14*(2), 301–314.

Baiman, S., Fischer, P., & Rajan, M. (2000). Information, contracting, and quality costs. *Management Science, 46*(6), 776–789.

Baiman, S., Fischer, P., & Rajan, M. (2001). Performance measurement and design in supply chains. *Management Science, 47*(1), 173–188.

Baiman, S., Netessine, S., & Kunreuther, H. (2004). Procurement in supply chains when the end-product exhibits the "weakest link" property. https://doi.org/10.2139/ssrn.2077640

Bajpai, P. (2019, February 14). IBM and Blockchain: What it did in 2018, and where it's going in 2019. Nasdaq. https://www.nasdaq.com/article/ibm-and-blockchain-what-it-did-in-2018-and-where-its-going-in-2019-cm1100102

Balachandran, K., & Radhakrishnan, S. (2005). Quality implications of warranties in a supply chain. *Management Science, 51*(8), 1266–1277.
Basu, M. (2015, September 21). For first time, company owner faces life sentence for food poisoning outbreak. CNN. https://edition.cnn.com/2015/09/20/us/peanut-butter-salmonella-trial/index.html
Chao, G., Iravani, S., & Savaskan, R. (2009). Quality improvement incentives and product recall cost sharing contracts. *Management Science, 55*(7), 1122–1138.
Chen, J., Qi, A., & Dawande, M. (2020). Supplier centrality and auditing priority in socially-responsible supply chains. *Manufacturing & Service Operations Management, 22*(6), 1199–1214.
Chod, J., Trichakis, N., & Yang, S. (2021). Platform tokenization: Financing, governance, and moral hazard. Management Science, Forthcoming.
Chod, J., Trichakis, N., Tsoukalas, G., Aspegren, H., & Weber, M. (2020). On the financing benefits of supply chain transparency and blockchain adoption. *Management Science, 66*(10), 4378–4396.
Chu, L., & Lai, G. (2013). Salesforce contracting under demand censorship. *Manufacturing & Service Operations Management, 15*(2), 320–334.
Chu, Y. (2016, June 23). Getting smart about smart contracts. *Wall Street Journal*. https://deloitte.wsj.com/cfo/2016/06/23/getting-smart-about-smart-contracts/
Chu, Y. (2017, March 9). Contracts get smarter with blockchains. *Wall Street Journal*. https://deloitte.wsj.com/cio/2017/03/09/contracts-get-smarter-with-blockchains/
Corbett, C., & DeCroix, G. (2001). Shared-savings contracts for indirect materials in supply chains: Channel profits and environmental impacts. *Management Science, 47*(7), 881–893.
Cui, Y., Gaur, V., & Liu. J. (2020). Blockchain collaboration with competing firms in a shared supply chain: Benefits and challenges. https://doi.org/10.2139/ssrn.3626028
Cui, Y., Hu, M., & Liu, J. (2021). Value and design of traceability-driven blockchains. https://doi.org/10.2139/ssrn.3291661
Dong, Y., Xu, K., Xu, Y., & Wan, X. (2016). Quality management in multi-level supply chains with outsourced manufacturing. *Production and Operations Management, 25*(2), 290–305.
Fang, X., Ru, J., & Wang, Y. (2014). Optimal procurement design of an assembly supply chain with information asymmetry. *Production and Operations Management, 23*(12), 2075–2088.
Gan, J., Tsoukalas, G., & Netessine, S. (2021). Initial coin offerings, speculation, and asset tokenization. *Management Science, 67*(2), 914–931.
Goldschmidt, D. (2018, October 5). 6.5 million pounds of beef recalled due to salmonella outbreak. CNN. https://edition.cnn.com/2018/10/04/health/salmonella-outbreak-jbs-tolleson-meat-recall-bn/index.html
Gümüş, M., Ray, S., & Gurnani, H. (2012). Supply-side story: Risks, guarantees, competition, and information asymmetry. *Management Science, 58*(9), 1694–1714.
Hellwig, D., & Huchzermeier, A. (2021). Distributed ledger technology and fully homomorphic encryption: Next-generation information-sharing for supply chain efficiency. In V. Babich, J. Birge, & G. Hilary (eds.), *Innovative technology at the interface of finance and operations*. Springer Series in Supply Chain Management, (Vol. 13, pp. xx–yy) Springer Nature.
Hilary, G. (2021). Blockchain and other distributed ledger technologies, an advanced primer. In V. Babich, J. Birge, & G. Hilary (eds.), *Innovative technology at the interface of finance and operations*. Springer Series in Supply Chain Management, Springer Nature.
Hu, B., & Qi, A. (2018). Optimal procurement mechanisms for assembly. *Manufacturing & Service Operations Management, 20*(4), 655–666.
Hwang, I., Radhakrishnan, S., & Su, L. (2006). Vendor certification and appraisal: Implications for supplier quality. *Management Science, 52*(10), 1472–1482.
Isidore, C. (2001, May 22). Ford recalls 13 million tires. CNN. https://money.cnn.com/2001/05/22/recalls/ford/
Jiang, L., & Wang, Y. (2010). Supplier competition in decentralized assembly systems with price-sensitive and uncertain demand. *Manufacturing & Service Operations Management, 12*(1), 93–101.

Karimi, F., & Goldschmidt, D. (2018, July 26). These foods linked to ongoing outbreaks might make you sick. CNN. https://edition.cnn.com/2018/07/20/health/salmonella-cyclospora-outbreak-food-recall/index.html

Li, L. (2012). *Capacity and Quality Contract Design in Supply Chain*. PhD thesis, Northwestern University

Metcalfe, R. (2019, February 21). The life story of your supermarket chicken: Food tracking goes high-tech. *Wall Street Journal*. https://www.wsj.com/articles/the-life-story-of-your-supermarket-chicken-food-tracking-goes-high-tech-11550761202

Mu, L., Dawande, M., Geng, X., & Mookerjee, V. (2016). Milking the quality test: Improving the milk supply chain under competing collection intermediaries. *Management Science, 62*(5), 1259–1277.

Nash, K. (2016, December 16). Wal-Mart readies Blockchain pilot for tracking U.S produce, China pork. *Wall Street Journal*. https://blogs.wsj.com/cio/2016/12/16/wal-mart-readies-blockchain-pilot-for-tracking-u-s-produce-china-pork/

Nath, T. (2020, March 25). How do recalls affect a company? Investopedia. https://www.investopedia.com/articles/investing/010815/how-do-recalls-affect-company.asp

Nikoofal, M., & Gümüş, M. (2018). Quality at the source or at the end? Managing supplier quality under information asymmetry. *Manufacturing & Service Operations Management, 20*(3), 498–516.

Petruzzi, N., & Dada, M. (1999). Pricing and the newsvendor problem: A review with extensions. *Operations Research, 47*(2), 183–194.

Plambeck, E., & Taylor, T. (2016). Supplier evasion of a buyer's audit: Implications for motivating supplier social and environmental responsibility. *Manufacturing & Service Operations Management, 18*(2), 184–197.

Prevor, J. (2008, March 25). Wal-Mart announces product removal fee. Perishable Pundit. http://www.perishablepundit.com/index.php?date=03/25/08&pundit=2

Pun, H., Swaminathan, J., & Hou, P. (2021). Blockchain adoption for combating deceptive counterfeits. *Production and Operations Management, 30*(4), 864–882.

Rui, H., & Lai, G. (2015). Sourcing with deferred payment and inspection under supplier product adulteration risk. *Production and Operations Management, 24*(6), 934–946.

Tsoukalas, G., & Falk, B. (2020). Token-weighted crowdsourcing. *Management Science, 66*(9), 3843–3859.

Enterprise Payments with Central Bank Digital Currency

An End-to-End Technology Point of View

Martin Fleming, Alan King, and Francis Parr

1 Introduction

In a globalized internet-connected world, there is an increasing need for enterprises to make high-value payments digitally and with near real-time settlement. New technologies using distributed ledgers and blockchains claim to be able to offer this capability with improved reliability, greater efficiency and reduced cost compared with current bank based global payment systems. Financial, banking and monetary system institutions are facing increasing pressure to provide improved facilities for real-time digital payments, and over the next few years will need to pick and choose from the menu of available technology and monetary approaches addressing this. This paper surveys the available approaches, articulates the potential technology advantages of different distributed ledger solutions and makes specific recommendations on the approaches with greatest potential benefit from a technology perspective.

We will focus on the problems faced by enterprises making high-value payments in real-time, both in domestic currencies and across borders. Digital currencies are most often thought of as an innovation for consumer payments, for storage of value,

M. Fleming
The Productivity Institute, Alliance Manchester Business School, Manchester, UK
e-mail: martin@fleming41.com

A. King (✉)
IBM Research, Yorktown Heights, NY, USA
e-mail: kingaj@us.ibm.com

F. Parr (retired)
IBM Research, Croton-on-Hudson, NY, USA
e-mail: fnparr@gmail.com

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_5

for speculation or, on the darker side, for money laundering. But we emphasize that, of all economic agents who suffer high costs for making payments, it is large global enterprises that have the greatest incentives to innovate and are therefore likely to be important influencers in the evolution of future payment systems.

To develop this point, we first review the stresses to enterprise payments and the landscape of potential innovations – especially the cryptoassets. Then, to develop the topic further, we make two assumptions. First, there will be digital currencies in all centers of finance and trade, and the winners in this race will be the ones supported by the respective domestic Central Banks. Second, the monetary implementation of these Central Bank Digital Currencies (CBDC) will follow the outlines published in a Bank of England working paper (Kumhof & Noone, 2018).

The main contributions of this paper are twofold. First, we discuss a model and provide a working demonstration, based on Distributed Ledger Technologies (DLT), of a Digital Payment Provider (DPP) system for trusted, immediate, and final exchange of domestic CBDC. We note that this DPP system can implement the Bank of England framework, and so any risk to the existing domestic commercial banking can be managed by the Central Bank. Second, we describe a novel linkage mechanism that will enable trusted, immediate, and final exchange of CBDC across independently managed sovereign digital currencies.

We conclude that the combination of these ideas offers an architecture for trusted real time enterprise payments with minimized risk: to privacy of enterprise data, to national currency monetary policy, or to Central Bank sovereignty, and with substantial reduction in direct and indirect cost to the global economy.

2 Acknowledgements

We wish to acknowledge extensive conversations with a group of experts in economics, payments, and developments in digital currencies: Professor Rod Garratt, University of California at Santa Barbara; Nitin Gaur, Director of Financial Sciences, IBM Research; Professor Emeritus Charles Kahn, University of Illinois at Urbana-Champaign; and Antoine Martin, Federal Reserve Bank of New York. In addition, we wish to acknowledge the many contributions of our IBM Research colleagues to our work and the maturing of our ideas: Hao Chen, Upendra Chitnis, Prasenjit Dey, Aliza Heching, Keith Houck, and Sachiko Yoshihama; as well as our IBM colleagues from the World Wire program: Malavan Balanavaneethan, and Winston Yong. Finally, we wish to acknowledge the substantial contributions of Matthew LaFontaine, undergraduate at Cornell University, during his summer internship.

3 The Global Payment System – Real Time Gross Settlement

Despite unimaginable stresses and strains in recent years, the US dollar clearing and settlement system has performed remarkably well. Through the 2001 terrorist attacks on the US, the Great Financial Crisis of 2008–2009 and the 2020 global pandemic, the payment system has continued to process record trade volumes. The success has in part been the result of collaborative efforts of central banks and financial sector participants as well as continued innovation by all parties. (See Bech & Hobijn, 2006).

Indeed, nowhere has financial sector innovation been more apparent than in the means of payment and the ability to settle large interbank payments in near real-time. Historically, interbank payments have been settled via end-of-day netting systems in which positions are tallied and funds are exchanged on a net basis between the participants.

However, with rapid financial innovation, integration, and globalization, the volume of interbank payments has increased dramatically. As a result, central bank concern over the potential for contagion or a systemic event due to the unwinding of the net positions has resulted in the implementation of real time gross settlement (RTGS) systems. (See Bech & Hobijn, 2006).

Nonetheless, despite recent innovation, inefficiencies remain. Even though transactions can be settled in real time, funds very often are not posted and available for days. The resulting lag imposes a tax on global economic activity of 1–3% of GDP. However, recent developments in distributed ledger technology (DLT) provides the means and the opportunity to reduce such costs and inefficiencies.[1] The potential offered by DLT to address issues related to ownership, security, and traceability has not only created an interest in realizing possible transaction cost improvement but also in the creation of digital currencies, most notably among central banks. Substantial pressure appears to be building across private sector firms, governments, and central banks for a widely used and accepted CBDC.

The potential for one or more CBDCs to meaningfully reduce transactions costs through the successful launch of a CBDC will lead to mounting pressure for adoption. The People's Bank of China appears to be ahead of many other central banks and has taken a number of steps with a trial reported to be underway in Shenzhen, Suzhou, Chengdu, and Xiong'an. A successful CBDC, used globally and at scale, could have significant implications for the global economy, trade and capital flows, and importantly, the balance of global economic power (see Birch, 2020).

As the global reserve currency, the dollar has conferred on the US an "exorbitant privilege". The broad, deep, and liquid US financial markets along with sizeable

[1]The term "distributed ledger technologies" refers to the constellation of cloud-based services and encryption technologies that can support multiple parties, with appropriate privileges, to record and verify transactions in a consensus system of record. There are different component services making up the DLT "bundle", and different digital payments innovations will benefit from different technology combinations.

outstanding holdings of US Treasury securities makes the dollar a low-cost store of value and medium of payment. Consequently, US borrowers benefit with lower interest costs and the US government is able to attract abundant capital, reducing, or at least postponing, tax collections.

A successful, widely accepted non-US CBDC would reduce the US advantage to some degree and, likely confer on the nation offering the successful CBDC a source of economic advantage. Consequently, the nation best able to innovate and adopt the needed technology, transform business practices and financial sector payment processes, enable consumer and business adoption, and address legal and regulatory requirements could face the possibility of stronger economic growth, increased income, and more rapid wealth accumulation.

Key takeaways from this discussion:

- While the Real Time Gross Settlement (RTGS) systems are an important evolution in payments protecting the stability of monetary and financial systems, they do not by themselves provide enterprises with a low cost or convenient solution for making high value payments
- RTGS systems could use, but in practice have not used, distributed ledger or other emerging technologies in their implementation
- US based financial, banking and monetary institutions may be particularly motivated to innovate in the area of digital currencies and real-time payments to preserve the historically advantageous position of the US dollar in world trade.

4 Pressures for Change in Real Time Payments

The focus of our paper is High Value Payments (HVP) for enterprises. In this section we review some of the pressures for change. Part of the pressure, of course, arises from the daily frustrations experienced by treasury operations.

- Uncertain delays for settlement
- Difficulty viewing global account balances in real time
- Reconciliation of payments with internal and external accounts

Another part of the pressure is the emergence of digital technologies and FinTech companies that seem to promise real time payments, 24 $\times$ 7, with low transaction costs, minimal settlement delays, and information to automate many treasury processes.

We propose to analyze this landscape through the lens of Real Time Payments for enterprises.

4.1 Payments Through the Banking System

The current vehicle for business payments is via transactions in national currencies (USD, EUR, JPY, RMB, etc.), which are ultimately settled through reserve accounts in Central Banks.

- The payor enterprise submits instructions for its bank to make the payment.
- That bank transfers funds to the bank of the payee enterprise, using the Real Time Gross Settlement (RTGS) facilities of the Central Bank.
- Finally, the account of the payee enterprise is credited.

Even though banks settle in real time, the payment will likely not post to the payee account in real time. Banks have discretion when to post payments – enabling banks some degree of flexibility in managing their reserve account balances. When this scheme is extended globally for a payment involving multiple national currencies with correspondent banks, settlement may take several days, a week, or more.

4.2 Electronic Commerce

Since the emergence of electronic commerce services, pioneered by PayPal and Amazon, consumers have become used to anywhere, anytime digital payments fully integrated into merchandize ordering systems, account histories, and delivery logistics. Treasury payment systems have also evolved information-rich applications that support integration with accounting, banking, and Enterprise Resource Planning (ERP) infrastructure and interface directly with payment service providers.

But despite this surge of innovation, almost all payment servicing for electronic commerce in the G8 countries rides on the banking system rails – ultimately these are transactions between banks in the RTGS systems provided by the Central Bank. Message standards such as ISO 200022, while supporting the transmission of rich information to accompany the payment, do not provide the instantaneous settlement and closure of the transaction. Some wire service providers and banks are building infrastructure to track payment progress, such as SWIFT gpi, but there are competing frameworks backed by new technologies on the horizon.

See (Klein, 2020) for a discussion of AliPay and WeChat electronic commerce based payment systems in China.

4.3 Digital Currency Innovation

A dramatic burst of innovation in transactions technology centered around blockchain has appeared and has caused a worldwide ferment in the payments industry. In addition, the cryptoasset explosion has spurred the banking industry to investigate blockchains for a great part of their information infrastructure, and Central Banks have even begun to explore the possibility of issuing Central Bank Digital Currencies.

4.3.1 Cryptoassets – Bitcoin

Bitcoin was the first popularized digital currency and continues to be the best known. It was introduced in a paper posted to a cryptography forum (Nakamoto, 2008). The goal was to provide an anonymous, completely unregulated digital currency using a public distributed ledger, based on a technology of cryptographically linked records called a blockchain.

Bitcoin is a dramatic proof of the feasibility of blockchain for securing a distributed ledger on the open internet. The ledger has not been hacked, even though the quantity of bitcoin in circulation is well over 100 billion USD.

The consistency of the blockchain system of record is maintained by distributed consensus protocols:

- Transactions are grouped into blocks.
- New blocks are submitted for consensus validation.
- Once validated, the block is entered into the system of record.
- Hashing and encryption processes ensure that the sequence of transactions is tamper-proof.

In Bitcoin, and many other cryptoassets, the blockchain extension is validated by a proof-of-work protocol based on solving a cryptographic hashing problem of known difficulty. The difficulty is increased over time to produce bitcoins at a predetermined rate. Bitcoin miners compete to solve the problem and are rewarded with newly created bitcoins. The power consumed in this competition is legendary – in 2019, bitcoin miners used more electricity than the entire nation of Switzerland.

Since Bitcoin appeared, several thousand cryptoassets have been introduced and new entrants continue to appear. Notable among more recent arrivals is Ethereum and the underlying Ether coin. This has a more transparent governance arrangement, has options to use proof of stake rather than proof of work (reducing the energy consumption) and better APIs to develop and launch distributed applications using smart contracts on each DLT node.

The defining property of cryptoassets are:

- Transactions extend a distributed ledger stored on a blockchain validated by a consensus proof;
- The ledger is readable by anyone with access to the blockchain (Bitcoin is public);
- Settlement is instantaneous and irrevocable once a transaction is proved valid;
- Transaction records can contain any amount or type of digital information – even programs implementing “smart contracts”.

The Bitcoin demonstration has led to many innovative products where multiple parties each need to create, share, and verify a permanent record of a chain of transactions between them. Blockchain technology now has widespread acceptance in the shipping industry to create a permissioned, shared, and permanent record of processes, replacing manual and often unreliable paper-based systems. Blockchain technologies are being investigated for many similar kinds of bank transaction recording, especially in custody and clearing.

4.3.2 Stablecoins

A stablecoin is a cryptoasset issued with a promise to ensure a stable value that is guaranteed by a backing portfolio of high-quality assets. Examples are Tether and TrueUSD, which are cryptoassets pegged to the value of one US dollar and backed by US dollar assets. See (Birch, 2020) for additional discussion of stablecoins particularly relating to the need for global digital identities.

The combination of a stablecoin and wallet services, enables at least conceptually:

- Any user can send payments directly to any other user with presumably automated wallet services and stable coin ledger as the only intermediaries
- This service could be real time, 24 × 7 available, and potentially global

Stablecoins have "lit a fire" in the world of banking. The prospect of billions of consumers trading stablecoins internationally through easily accessible applications raised many alarms. A sample of the questions raised:

1. Are the requirements for obtaining a stablecoin wallet the same as those for obtaining a bank account?
2. Who has access to the logs and for what purposes can this information be used?
3. Are there any guarantees on stablecoin claims in cases of financial stress?
4. Are specific answers to such questions business policies, which can be changed over time, or are they legally binding agreements?

Finally, like any national currency backed by portfolios of "reserve assets", stablecoins would be a currency domain with potentially billions of users – giving rise to concerns about its influence on national monetary policies.

Perhaps the best summary and reaction to the evolution of stablecoins comes from Mark Carney, then-Governor of the Bank of England, as reported in Bloomberg News (2019-06-18): "Anything that works in this world will become instantly systemic and will have to be subject to the highest standards of regulation," and on the other hand, "We need to have an open mind."

4.4 Central Bank Digital Currency

According to the Bank of International Settlements (BIS) Innovation Hub: "A 2021 BIS survey of central banks found that 86% are actively researching the potential for CBDCs, 60% were experimenting with the technology and 14% were deploying pilot projects." (BISIH, 2021). Central banks are chartered to protect the use of their currency as a unit of account, store of value, and a medium for payment. They also conduct monetary policy and regulated their banking system. As part of their charter, many central banks are also studying the issuance of a Central Bank Digital Currency (CBDC) that would have the function of a stablecoin but would also be a direct liability of the central bank. See (Barontini & Holden, 2019) for descriptions of pilots in Sweden and Uruguay.

Some examples of different central banks with differing motivations exploring CBDC approaches are reviewed in (Patel, 2018). As argued by (Kumhof & Noone, 2018) any CBDC must be carefully designed to avoid draining deposits from the fractional reserve banking system, and thereby undermining a critical source of credit creation and liquidity provision. In summary:

- Central banks are exploring feasibility and possible benefits of introducing a CBDC with the function and features of a stablecoin.
- It is likely to be easier for Central Banks to establish credibility and trust in its CBDC than for a stablecoin initiated by a private enterprise.
- CBDCs would encourage competition to offer real time digital payment and digital wallet services to end users.
- Careful planning is required to manage the impact on the existing banking system.

A CBDC is a stablecoin that is supervised and monitored by a central bank with payment wallets and end user accounts provided by the private sector. In this way, a Central Bank could leverage private sector competition to improve payment services based on digital technologies, but within a regulated monetary framework.

4.5 High Value Payments Are Different

Real-time high-value payments, and especially cross-border payments, require a more complex operational infrastructure than considered in most digital currency scenarios. Payment operations may involve multiple accounts across multiple institutions, even possibly drawing upon lines of credit along the way.

Existing mechanisms in capital and foreign exchange markets suggest some possible approaches using digital currency. Trillions of USD are raised daily by repurchase agreements (repos) between investment banks and money market mutual funds. Settlement is nearly instantaneous and certainly intra-day, using repo market maker custody and settlement services. Even more value is moved daily through foreign exchange markets via a Continuous Linked Settlement (CLS) system that in effect implements a version of RTGS.

The development of infrastructure and regulation for enterprise real time payments will take time. But the trend is clear. Enterprises have already shifted a substantial percentage of their deposits away from the banking system and into mutual funds and investment banks. Digital currency for high value payments is a natural progression.

5 Basic Principles of CBDC Design

The Bank of England working paper (Kumhof & Noone, 2018) discusses in detail some basic principles for the design of a Central Bank Digital Currency, under the socially and economically important criterion of preserving (and improving) the health of the commercial banking sector. In summary, these principles are:

- CBDC may pay (or charge) an interest rate that is managed by the Central Bank.
- CBDC and Central Bank Reserves are distinct and not convertible.
- No guaranteed conversion of currency-denominated bank deposits into CBDC.
- CBDC is issued only to Distributed Payment Providers chartered by the Central Bank, in exchange for transferred eligible securities, such as government bonds.

Note that these principles do not rule out the use of CBDC as a means of direct "peer-to-peer" payments between businesses or individuals.

Note also that the Central Bank has two tools for managing monetary policy:

- Adjust the amount of CBDC exchanged for eligible bonds;
- Adjust the interest rate on CBDC balances.

5.1 Digital Payment Providers

The Bank of England design principles suggest a hybrid CBDC model. The commercial banking system of lending is essentially unchanged. The Central Bank charters some number of Digital Payment Providers (DPPs). The Central Bank issues CBDC to DPPs in exchange for high quality collateral. Customers create accounts at DPPs and purchase CBDC either by transferring currency or pledging high quality collateral.

Digital Payment Providers offer the following services:

- Enable real time payments via token transfer;
- Charge fees and collect or pay interest on CBDC;
- Buy and sell currency, bonds, and CBDC in appropriate markets.

Some of the key business benefits of this hybrid CBDC approach are:

- Central banks are not organizationally structured or equipped with IT and other resources needed to directly manage large numbers of customer accounts as required for an enterprise payments system.
- The existing account-based systems operated by financial institutions are already equipped to provide the entire payment services bundle. DPPs could be created as bankruptcy remote divisions of banks.

The basic institutional thoughts underlying Digital Payment Providers are to outsource the infrastructure for enterprise real-time payments to independent specialized payments providers. Under the Bank of England model, DPPs would be chartered as financial institutions. Moreover, the IT infrastructure for payment services bundles are readily available. It is quite natural for DPPs to be created as bankruptcy remote divisions of commercial banks.

5.2 DPP Balance Sheet Transitions

We provide some insight into the DPP, by describing the top-level transactional steps in terms of balance sheet transitions.

5.2.1 Purchase of CBDC by a DPP

The Central Bank creates digital currency in exchange for high-quality collateral. The diagram in Fig. 1 shows the incremental change to the balance sheets of a Central Bank and Digital Payment Provider (DPP) D1 after the exchange of a $1M bond for $1M central bank digital currency (CBDDC). Notice the following:

- The bond transferred to the Central Bank by DPP D1 and the CBDC tokens received are both assets of D1 so D1's capital is unchanged.
- CBDC in circulation is considered a liability of the Central Bank, since CB is committed to buy back CBDC from any chartered DPP holding it at par so the CB capital is unchanged.
- By creating the CBDC tokens and transferring them to D1 in return for the $1M bond, CB has expanded its asset base.

The righthand side of figure y shows D1's balance sheet after this account D1.E1 is set up.

5.2.2 Funding a Customer CBDC Account with DPP D1

The following diagram, Fig. 2, illustrates the change in DPP D1's balance sheets before and after a customer – say Enterprise E1 – sets up and funds a 5k CBDC account. We note the following:

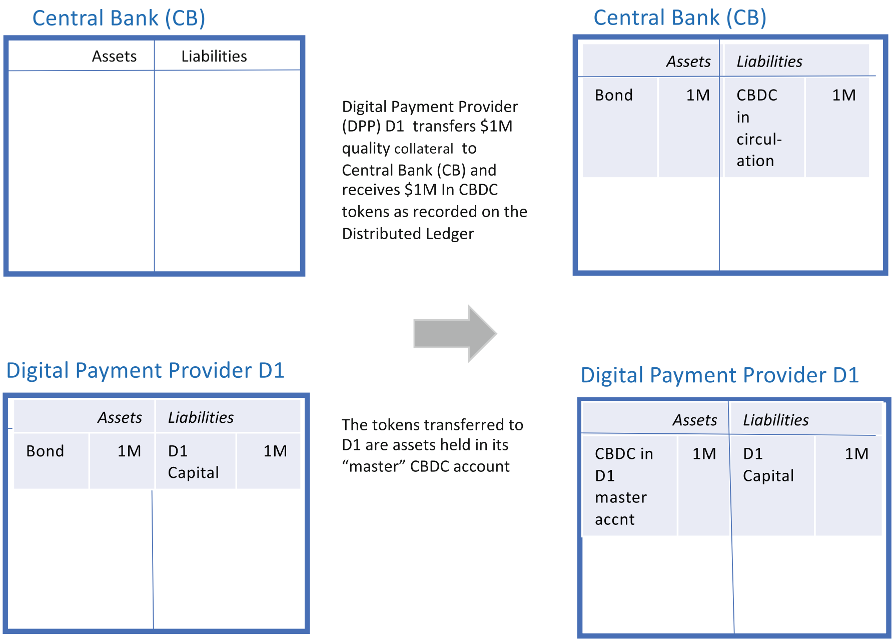

Fig. 1 Balance sheets for DPP D1 transferring collateral and receiving CBDC

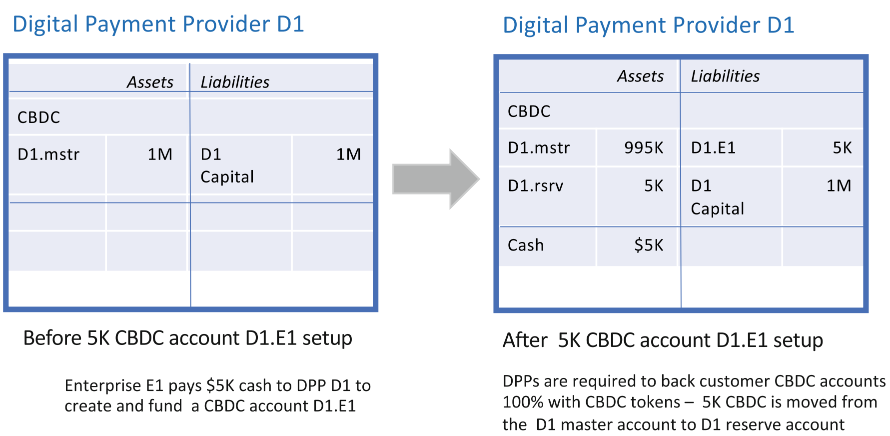

Fig. 2 Balance sheet for D1 before/after Enterprise E1 funds its CBDC account

- The DPP may charge more or less than $5k currency depending on market conditions.
- The enterprise may deliver currency as physical cash, or more likely, through a bank transfer.
- The DPP balance sheet allocates 5K CBDC in a reserve account to balance the 5K CBDC liability to E1.

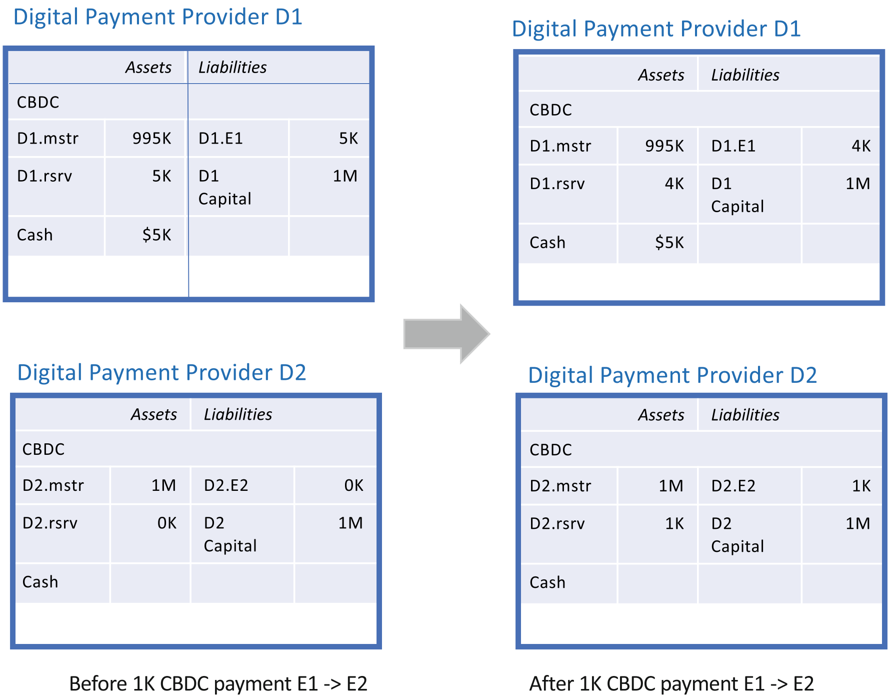

Fig. 3 Balance sheets for DPP's D1, D2 before/after E1 to E2 CBDC payment

Figure 2 highlights the assumption that the DPPs are "narrow" in the sense that the DPP's asset book reserves 100% of the liabilities. This assumption need not be a requirement of the DPP model. But without it DPPs become more like traditional banks, and settlement will require some kind of clearing and netting process. The point of this model is that exchange of CBDC is a final settlement.

5.2.3 Payment Between Enterprise Accounts

Figure 3 shows the balance sheet changes when Enterprise E1 makes a real time payment of 1K CBDC to an Enterprise E2 with a CBDC account at DPP D2.

The payor DPP is D1. Its assets and liabilities are both decreased by the amount of the payment. The payee DPP is D2. Its assets and liabilities are both increased. In both cases the notional capital remains the same, but the distribution of the asset side of the balance sheet has changed. The process is as follows:

- DPP D1 will debit the balance in D1.E1 – checking sufficient funds available in the process.
- DPP D1 transfers $1K CBDC in tokens to D2 from its D1 reserve account.
- D2 receives this CBDC into its D2 reserve account.
- D2 credits the account D2.E1.

6 Using a Distributed Ledger/Blockchain for CBDC

6.1 Distributed vs Centralized Ledgers

The CBDC design proposed in section "Basic principles of CBDC design" could be realized as a centralized ledger implemented in a single database system. DPP and Central Bank agents would connect as clients to this database to view ledger balances and submit update operations. In contrast, a distributed ledger implementation allows for multiple instances of the same ledger maintained and operated by different organizations. The ledger instances share distributed ledger infrastructure and chaincode that provides the access controls, query, and update functions available to the user organizations. This infrastructure and chaincode will have been reviewed and approved by all the user organizations at system startup or when they started participating in the ledger.

The distributed ledger infrastructure coordinates update transactions across different instances of the ledger. Each requested update transaction must be (1) sequenced with respect to other updates, (2) validated by automated access control and chaincode logic, (3) grouped with other updates into a block which in turn is chained cryptographically onto all preceding update blocks and, (4) distributed to all ledger instances to be applied in order to advance the ledger state in that instance.

An example of automated access control and chaincode logic would be a check before allowing debit of a ledger account that the requester is properly authorized to make debit and that there are sufficient funds in the account at this time for the debit to proceed.

The term *blockchain* describes the cryptographically linked sequence of update blocks since the start of database operations or last checkpoint. This chaining of blocks ensures that every instance of the ledger is seeing exactly the same historical sequence of historical update transactions and that no blocks have been modified or improperly inserted into the chain. The different instances are not synchronized but are all advancing on the same evolutionary path as they apply updates in order to their view of the ledger state. Any one instance could be a few update transactions or, in extreme cases a few blocks, ahead of or behind its peers. In particular if one ledger instance fails, its peers can typically proceed and continue submitting and applying update transactions. If the instance normally used by a DPP or Central Bank fails, that organization can temporarily switch over to using an alternate instance until its own primary instance recovers and catches up to a current state. This gives distributed ledger implementations a 24×7 resiliency that would require careful mirroring and duplication for a centralized database ledger implementation.

6.2 Trust in a Ledger Based Payment System

A key issue for any payment system is trust. Payment by exchange of physical coins or bills, engineered to be hard to copy or forge, is surely trustworthy. But is slow and expensive for large payments, or when the parties are remote. A real-time payments system for modern enterprises must be digital. That is to say it must be possible to make a payment by adjusting balances in trusted ledgers, maintained by trusted third parties, using trusted units of account to make the transfer. Conventionally, enterprises establish accounts with a major bank or financial institution. These banks are highly regulated and as a result permitted to hold reserve accounts with their Central Bank. The Central Bank itself is highly trusted and indeed is the lender of last resort for the banking system.

When enterprise E1 banking with B1 makes a payment to enterprise E2 banking with B2:

- the payment is deducted from E1's account in bank B1's ledger.
- Bank B1 records this as part of its netted business with bank B2 for the settlement period.
- The Central Bank records the adjustment in the reserve balances of banks B1 and B2 – reflecting the net transfer to be made between B1 and B2 at the close of the business day, or for an intra-day period if RTGS is used.
- Bank B2's ledger will reflect the receipt of a credit for enterprise E2 in its netting log at settlement.
- Bank B2 will post this credit amount to enterprise E2's account.

There are at least three different ledgers involved in this payment process. In addition to all parties trusting the Central Bank, E1 must trust B1 to remain solvent, to maintain its B1.E1 account, to ensure that requested payments reach the intended payee and to preserve the confidentiality of E1's balance and payment history. E2 must have similar confidence in B2. If there is some question about whether E1's payment completed and was delivered to the correct payee (E2) account, a reconciliation audit will be required to check through each of the involved ledgers to find the relevant transaction step, confirm that it executed correctly and is properly completed.

In a CBDC distributed ledger, banks B1 and B2 are chartered as enterprise Digital Payment Providers. When enterprise E1 makes a real-time payment from its CBDC account at DPP B1 to enterprise E2's CBDC account at DPP B2, the ledger system configuration is significantly altered. There is now just a single distributed ledger, shared between B1, B2 and the Central Bank. As noted in the previous section, this single distributed ledger accomplishes the end-to-end payment from E1 to E2 in a single atomic transaction. Notice that reconciliation for CBDC payments using a distributed ledger does not occur. Inspection of the distributed ledger history will locate the payment transaction showing if when it completed and who the recipient was. All instances of the ledger – let us assume B1, B2 and the Central Bank each have their own instance – will eventually see this transaction and its exact outcome. No cross-referencing of ledgers is required.

Although the ledger configuration is significantly changed when a distributed CBDC ledger is introduced, the organizational trust relationships stay essentially the same. Enterprises E1 and E2 must both trust their banks B1 and B2 respectively. Everyone trusts the Central Bank. And as is current practice, the Central Bank charters and regulate the Digital Payment Providers B1 and B2. The technical details of setting up the DPP distributed ledger are not technically more difficult than the current RTGS system.

While processing of CBDC payments is handled by the distributed CBDC ledger, this does not relieve Digital Payment Provider banks B1 and B2 from having to maintain ledgers. Enterprises E1 and E2 are likely to need regular (non-CBDC) accounts with B1 and B2 to fund and receive regular credit from their CBDC accounts, and to conduct regular non-CBDC banking business. DPP banks B1 and B2 also will be required to maintain Know Your Customer (KYC) information about customer account holders such as E1 and E2 as a CBDC DPP responsibility – and for other enterprise banking interactions.

6.3 Trust in Digital CBDC Tokens

The attributes of trust in digital tokens are:

- Stable value;
- Difficult to forge;
- Prevention of "double-spending".

Stable value is the property that makes a digital token useful as a medium for payment rather than as a vehicle for speculation. Difficulty forging is made possible by the invention of public key infrastructure (PKI). Preventing double spending had been considered a difficult problem until (Nakamoto, 2008) introduced the blockchain scheme for this subsequently used in Bitcoin and other more recent cryptoassets.

Double spending is prevented in these schemes by chaining ordered blocks of update transactions together cryptographically. Checksum information is incorporated into each reference link in the chain in such a way that no individual ledger instance can modify the contents of any "accepted" historical block of updates.

History can be rewritten only if a majority or consensus of ledger instances were to decide to rewrite history in exactly the same way at the same time. This can and has occurred in some crypto assets; it typically takes the form of a split when the consensus method fails – leading to variants of the ledger appearing and competing with each other.

In successful publicly accessible cryptoassets such as Bitcoin, where anyone with access to the internet can compete to add acceptable new blocks of updates into the chain and there may be many thousands of instances of the ledger, the consensus mechanism to determine which new blocks are to be added to the chain is strong enough to prevent this happening. In Bitcoin, consensus is driven by a competition between 'miners' who must (1) assemble a sequenced block of recent valid new transaction requests to be added as latest block in the chain, (2) link this new block with a reference to the preceding block in the chain, and, (3) compete to be the first to solve a computationally intensive numerical problem involving the signature of the new block and its link to the predecessor block. The first miner to succeed in doing this and subsequently has their block used and appended to by others is rewarded with some fraction of a bitcoin. In this design, the rate of creation of new digital assets is determined be a completely decentralized algorithmic rule built in to the crypto asset definition.

Our proposed use of a distributed ledger for a CBDC also uses cryptographically chained blocks of updates to prevent double spending and maintain consistency between ledger instance. But our CBDC distributed ledger has a centralized owner and sponsor, namely the Central Bank, and can be set up as a private permissioned distributed ledger. The permissible payment transactions are implemented in chaincode reviewed and approved by all participants – including the Central Bank. Access controls are such that each DPP is authorized to submit payments for their customers' accounts but to accept payment from any customer CBDC account.

Finally, and most importantly, only authorized agents of the Central Bank can create (or possibly remove) CBDC digital tokens. This is completely decoupled from any consensus mechanism on any blocks of update transactions. For a successful CBDC, the Central Bank will manage CBDC digital token supply so as to maintain parity with the underlying national currency and in accordance with its monetary policy for both the regular and digital money supply.

The digital CBDC tokens issued by the Central Bank have the following properties:

- Each token is uniquely identified with its value (denominated in the underlying national currency) date of issue and a unique serial number issued by the central bank.
- These properties are associated with the token and signed cryptographically by the Central Bank using a public/private key scheme so that the properties of a taken can be read but not modified by DPPs.
- The distributed CBDC ledger records a unique owner/holder (a DPP or Central Bank account) for each CBDC token at any point in time. It is only the current DPP owner who can authorize transfer of the token to another owner.
- The distributed ledger blockchain log records the full provenance – sequence of holders and transfers – of each identified digital CBDC token from time of issue to the present, assuring that no double spending has occurred.

6.4 Other Issues Relating to a Distributed CBDC Ledger

6.4.1 Token Denominations

We make the simplifying assumption that the distributed CBDC ledger uses digital token "denominations" – like D$ 100, D$ 50, and so forth – just as with physical currency. A DPP will usually have a workable combination of tokens in its holdings to support any user payment request. On rare occasion when this is not the case, a DPP may be able to use a change service, possibly fully automated in chaincode, and possibly provided by the Central Bank or authorized agents.

An alternate scheme is to allow base tokens to be subdivided and have the CBDC ledger record the movements of a collection of fractions of base tokens for every payment. Fragmentation of tokens is part of the design of this scheme. Defragmentation services may be authorized to take in collections of partial tokens, remove them from circulation, and replace them with new complete tokens. Since defragmentation involves creation and destruction of tokens it should only be performed by the Central Bank.

Neither case seems much different than the other. For a successful CBDC where we would want each digital CBDC token to be involved in many payment transfers during its lifetime, the multi-denomination approach looks significantly simpler than tracking fragments of individual tokens on the ledger.

6.4.2 Serving Consumers and Enterprises with a Single CBDC Ledger

Our perspective in this paper has been to explore monetary schemes and supporting technologies to facilitate real time payment between enterprises. CBDC is a leading contender for this role. But the Central Banks who must necessarily be the leaders in any CBDC deployments, will be as concerned with its appeal to their general public for digital consumer payments.

There are important differences in requirements for consumer and enterprise digital currencies. Enterprise payments must be account based, with KYC information documenting account holders to enable Anti-Money Laundering, audits to detect fraud or abuse and to allow reversal of error. Authorization of high value payments must be robust. Account identities must be stable and publicized in directories to ensure correct routing of payments and allow evidence that a specific payment was made by or received by an intended enterprise. In contrast, consumers may appreciate the convenience of digital wallets and bearer instruments, the ability to make payments offline when no network is available and, in some circumstances, presumably for payment amounts below some money laundering threshold, to make payments privately and anonymously.

Consumer preferences for digital payments are beyond the scope of this paper. The saving grace from the perspective of CBDC technology is that the distributed ledger used to process and settle CBDC payments is in large measure insensitive to these different usage patterns. It is the Distributed Payment Providers and the additional services they offer outside processing payments through the distributes CBDC ledger where distinctions between enterprise and consumer use become important. The enterprise DPP business model will need to allow for account initiation processes supporting validated owner identification, and multifactor multiparty authorization for large payments. Consumer DPPs will offer digital wallets and offline payments possible even creating a CBDC account per transaction or managing user CBDC payments through the DPP's account. Typical consumer payment amounts would be smaller and the volume correspondingly greater.

This higher transaction volume for consumer payments will present performance and scaling challenges for distributed ledger processing, but in principle a distributed CBDC ledger system could be initially deployed for enterprise payments, then engineered and extended for performance as additional categories of DPP are commissioned with different target customers.

7 Exploratory Chaincode for a CBDC Digital Payment Provider

In this section we present an exploratory model for a Digital Payment Provider system for enterprise real time payments using digital ledger infrastructure technology – specifically Hyper Ledger Fabric. We identify the exact roles played by different components of the technology. The key benefits of this implementation, we shall argue, are due to the decentralized operations afforded by the technology, the automatic audit and reconciliation, the privacy and security guarantees, and the immediate and final settlement.

7.1 A Design for the Digital Payment Provider Model

Our design focusses quite narrowly and specifically on the adjustment of account balances and exchange of CBDC between Digital Payment Providers. This is the component of DPP operational logic where the underlying distributed ledger technology most directly influences payment processing.

We will describe our implementation by stepping through each of the fundamental operations illustrated by the balance sheet changes in the previous section:

- DPP exchanges securities for CBDC with the Central Bank;
- DPP exchanges CBDC for currency with customers;
- DPPs execute real time CBDC payments on behalf of enterprises holding CDBC accounts with them.

Our implementation is based on the HyperLedger Fabric, an open-source infrastructure for Distributed Ledger applications, see https://hyperledger-fabric.readthedocs.io. Hyperledger Fabric source code was donated by IBM to the Linux Foundation and is readily available in open source.

7.2 Distributed Ledger Concepts Used by the Prototype

We begin this overview by introducing some key Distributed Ledger/Hyperledger Fabric technical concepts used in our demonstration.

7.2.1 Peer Nodes

The distributed ledger or fabric has a collection of *peer nodes*, each of which has identical copies of the ledger. Organizations may operate one or more peer nodes. Following our example above, the Central Bank and the two Digital Payment Providers would each operate one or more peer nodes. Multiple peer nodes maintaining identical copies lends resilience to the distributed ledger.

Open DLs permit anyone to create a peer node and submit transactions. Permissioned DLs, as the name implies, maintain a private pool of peer nodes. The HyperLedger Fabric is a permissioned ledger.

7.2.2 Transactions and Consensus

Transactions are approved and added to distributed ledgers by consensus. In the HyperLedger Fabric, consensus involves a stateless "ordering" node. Transactions are submitted to a subset of peers for validation. When a sufficient consensus of the peer subset verifies the result, the validated transaction is sent to the ordering node; the ordering node processes sets of transactions into a (blockchain) update and distributes the updates to the peers. The ordering node ensures the consistency of the transactions across all the peers and eliminates any need for reconciliation.

7.2.3 Chaincode and Smart Contracts

Some distributed ledgers use software modules called "smart contracts" or "chaincode" to create transactions. In the HyperLedger Fabric the installation, modification, and maintenance of chaincode is treated just like a transaction. Chaincode modifications are submitted to a relevant subset of peers, evaluated, and submitted to an ordering node. The evaluation, of course, may require extensive testing before it can be committed to the ordering node.

Key features of chaincode relevant to its use in a CBDC implementation follow from the definition above:

- The chaincode is collectively owned by the "consortium" of user organizations. Typically, no single organization can modify it unilaterally. Hence all (authorized) peers are seeing exactly the same logical capabilities.
- It is fully executable and automated – hence specific obligations such as checking the sufficient funds are available in a payer's account before making a payment can be written into the smart contract and will execute precisely as implemented in all payment transactions.

7.2.4 Channels

A Hyperledger Fabric *channel* is a scoping mechanism. Channels identify specific subsets of peers with access to the channel scope, specific subsets of business objects that can be read or written from or to the peers in the channel, and specific ordering nodes to update the objects in the channel.

Chaincode deployments are also scoped to channels. A chaincode deployment request will specify the target channels for which deployment is requested. Use of multiple channels can enhance the scalability of a blockchain distributed ledger: when different channels use independent ordering services, updates to their data can be sequenced separately.

The fabric analogy is motivated by the following perspective. Transactions modify business objects in specific channels and at specific times. Channels are like parallel sequences (the warp) and updates by ordering nodes link multiple channels at particular timestamps (the weft).

7.3 Transactions and Queries

Chaincode makes available to authorized users a set of transaction and query functions for manipulating the state of business objects.

Queries are read only, i.e. they make no change to the state of the distributed ledger. Hence, in typical simple cases, ledger queries:

- Require no additional approvals beyond the authorization of the requester to execute the query;
- Can execute on any peer node instance – a single node is sufficient;
- Do not have to be processed through a network ordering service- the query will just see the current state of the peer node selected for its execution.

Transactions are updates to the network state and therefore require:

- Multiple approvers possibly from multiple organization may be required to approve an update transaction;

- When approved the transaction must be distributed to all peer instances of the channel and applied locally in network order – in blocks for the case of Hyperledger Fabric.

7.4 Demonstration DPP Implementation Using Hyperledger Fabric

In this section we describe the design of our prototype demonstration the DPP model using Hyperledger Fabric infrastructure. This prototype shows in some detail how permissioned DL concepts can be used to support reliable real time payments.

One important benefit of this prototype is to clarify which processing occurs on the distributed ledger and which processing is realized by some system managed privately by a Digital Payment Provider or the Central Bank.

A second benefit is to make clear the visibility rules – which parts of the distributed ledger are visible to each player.

In our implementation, the business objects on the ledger are:

- A record of all digital CBDC tokens currently in existence – including for each CBDC token:
  - The unique identifier (analogous to paper currency serial numbers) for each CBDC token;
  - The denomination of that token i.e. its declared currency face value;
  - The current holder of that token – either the CB or a specific DPP;
  - If the current holder is a DPP, token state. This is a true/false value indicating whether it is currently freely held by the DPP in a "master account" or whether it is currently allocated as the required 100% CBDC backing for enterprise customer accounts with that DPP.
- A record of all (enterprise) CBDC accounts with each DPP – including for each account:
  - The DPP identifier for the DPP providing the account;
  - The account identifier (within that DPP);
  - The current balance – denominated in CBDC – in the account.

7.4.1 CBDC Transactions – Mixed Chaincode and Manual Steps

The CBDC transactions in our demonstration implement the balance sheet transitions in the example of the previous section. Each balance sheet transition involves both chaincode and private steps.

The table Tx below summarizes the on-distributed ledger and off-distributed ledger aspects of each of the balance sheet transitions of the example from the preceding section.

Balance sheet transition	DLT processing implemented in chaincode	Manual or private IT system steps
T1: DPP transfers collateral and receives CBDC for its use	CB creation of CBDC tokens CB transfer of created CBDC to DPP with master account state	CB approval of the DPP collateral Transfer and acceptance of collateral
T2: Enterprise E1 sets up CBDC account with D1, and provides currency in exchange for CBDC balance in the account	Create D1.E1 account Credit D.E1 account with CBDC Move CBDC from master to reserve account	D1 accepts currency from E1 and records in private D1 ledger D1 saves E1 certificates, passwords, identifying information for AML, KYC in private ledger associated with new account identity
T3: Enterprise E1 requests D1 to make CBDC payment to enterprise E2 with account at D2	Debit D1.E1 account Credit D2.E2 account Transfer ownership of CBDC tokens currently held "reserve" state at D1 to "reserve" state at D2	D1 must ensure properly authorized payment request from E1 D1 and D2 must ensure payee has the identity payer expected
T4: Enterprise E2 converts received CBDC payment to national currency	Debit D2.E2 Change status of corresponding CBDC held at D2 from "reserve" to master	D2 must ensure properly authorized payout request from E2 Transfer cash or national currency balance out using private D2 ledger system and fractional reserve or RTGS settlement

7.4.2 Transition T2 – Establish Account

In balance sheet transition T2, Enterprise E1 establishes a CBDC account with DPP1 and funds it with $5 k cash to be held in CBDC in its account D1.E1. The balance sheet implications of this were shown in Fig. 2 with a $5 k payment in currency from E1 to D1 to set the opening balance in its CBDC account.

The following details are handled manually or within DPP D1's systems:

1. E1 contact information, addresses, login and password information to identify who will be authorized to make payment or "cash out" requests on the account.
2. D1's ledger balance of cash in hand could be of interest to the CB to ensure that the DPP in a financially healthy state, but this balance is not involved in any CBDC payment transactions.

The following details are handled as an atomic operation by the "Credit customer CDBC account" chaincode:

1. Credit D1.E1 account with 5k CBDC;
2. Move 5k CBDC from D1 master account to the reserve account.

7.4.3 Transition T3 – Payment

In balance sheet transition T3, enterprise E1 requests a real-time CBDC payment of 1k from account D1.E1 to enterprise E2's account D2.E2.

The following details are handled manually or within D1's systems:

1. Establish that this request is made by an authorized user;
2. Establish that D2.E2 is an authorized payee.

The following details are handled as an atomic operation by "CBDC Payment" chaincode:

1. Ensure that the account D2.E2 exists;
2. Check that there are sufficient funds in D1.E1;
3. Check that DPP2.E2 is a defined account at DPP2;
4. Identify CBDC tokens in the appropriate amount in D1's reserve account;
5. Transfer identified CBDC tokens from D1 reserves to D2 reserves;
6. Credit D2.E2 with the transferred CBDC.

In the case of a high value payment, the manual step of D1 authenticating the payment request could involve multifactor authentication steps or multiparty authentication.

The Central Bank is not involved directly in this payment scenario (although it may have visibility to the log of executed transactions and be able to conduct audits). This is consistent with the view that real-time CBDC payments must be available 24×7 while this may not be the CB's usual and natural operating / availability policy.

7.4.4 Transition T4 – Cash out

In transition T4, Enterprise E2 "cashes out" the 1K CBDC payment which has arrived in its account D2.E2.

The following details are handled manually, or within D2's systems:

1. The request comes from a properly authenticated E2 officer.
2. The resulting funds will be transferred to an approved account.
3. The D2 ledger of cash in hand is updated accordingly.

The following details are handled atomically by the "Cash out" chaincode:

1. Ensure sufficient funds in D2.E2 for the requested cash out amount;
2. Debit the CBDC account by this amount;
3. Move the corresponding CBDC tokens from reserve account to master account.

This completes our description of each major balance sheet transitions involved in real-time payments with CBDC. We have shown which function are executed as distributed ledger transactions implemented in deployed chaincode and which are provided either as manual steps or automated processing using some private non-shared ledger systems operated by a DPP, CB or commercial Bank. This separation provides significant insight into the role and influence of distributed ledger technology on feasible CBDC systems.

7.5 Trust Principles for the CBDC Ledger Based on Chaincode Checks and Balances

We now use the CBDC demonstration to discuss the ways that a DPP model combined with a CBDC token model, implemented on a distributed ledger that supports chaincode, and augmented with Central Bank policies and rules, can create a real-time payments system with a claim to be "trusted" by DPPs, Enterprise users and the Central Bank itself.

P1: *The Central Bank has clear control of the maximal amount of digital CBDC in circulation*

- The Central Bank is the only organization with authority to create CBDC tokens.
- Any individual DPP is limited in how much inflow into its CBDC customer accounts it can attract by the amount of CBDC it holds in "master" status.
- When this master account is depleted, the DPP will usually have to post additional collateral and get CB approval for more master CBDC before it can accept additional customer inflows. In extreme situations, such as a pandemic, the Central Bank could minimize or even waive the collateral requirement to promote economic growth.
- This mechanism when applied across all DPP's limits the possibility of a run into CBDC as noted in (Kumhof & Noone, 2018).

P2: *An Enterprise with a balance in a CBDC account with a DPP has minimal counterparty risk.*

Chaincode executed on the distributed ledger assures its account balance is backed 100% by CBDC regardless of whether the balance result from deposits by the enterprise or CBDC payments from other parties. We note that legal or policy innovation might be required to fully back this up. In the event of a DPP failure, there would need to be assurance either the CBDC would be depositors or that the Central Bank would guarantee CBDC balances at the time of failure.

P3: *There is no settlement risk to a DPP when accepting a CBDC payment*

- Settlement occurs instantaneously as the transaction updates the distributed ledger.
- Some instances of the distributed ledger may be aware of the change before others, but all instances will eventually see every successful update transaction.
- The chaincode check to ensure that there are sufficient funds available in the payer account will be applied in the same ordering relative to other credits and debits at that account in every instance of the distributed ledger. This is accomplished by the payment network wide transaction ordering service.

Once again there could be a need for legislation to protect DPP's and allow them to rely on the current state of the distributed ledger as returned by a chaincode query on any certified instance as reliable evidence that settled funds are available and payments based on them can be made without further due diligence by the DPP.

P4: *DPPs can treat CBDC transferred to them on the distributed ledger as unforgeable value holding tokens*

- Each CBDC token is "minted" and "signed" with a unique serial number and fixed denomination set by the CB at creation time.
- For every payment transferring CBDC tokens, the chaincode has checked that the source is the current valid "holder" of each specific CBDC token transferred.

Essentially the CBDC tokens have the serial number identifications of paper currency, together with a cryptographically protected provenance record which could be made available with appropriate protections to investigators, courts etc.

Payment confidentiality is a critical concern. Chaincode query capabilities must be designed to assure that this is protected – a topic which we will take up in the next subsection.

P5: *A CBDC Distributed Ledger system in operation will follow the source code logic of its chaincode queries and transactions and provide data protections following the configured security protocols*

This is a claim that the system is not "hackable" and an assertion about the reliability of the blockchain infrastructure, rather than the design of the blockchain itself. For our prototype CBDC implementation, we used Hyperledger Fabric where a measure of confidence is provided by the following:

- Hyperledger Fabric is an Open source Apache Project – open to wide review and use.
- The data on the blockchain itself is cryptographically protected.

- Chaincode is written in standard programming language with standard business object data definitions – source code logic is well understood, and standard compilation, execution, and deployment processes translate these into service endpoints.
- Deployed chain code executes in cryptographically protected closed containers.
- A variety of certification and security protocols are available in HLF and can be configured to provide an appropriate level of cryptographically protected operation for a specific permissioned blockchain network and distributed ledger.

7.6 *Discussion*

Our work in developing a prototype CBDC distributed ledger implementation led to identifying the above principles for establishing the trustworthiness of any CBDC system. We argue that working in detail through the validity of these principles for any specific proposed CBDC blockchain system is a necessary step before roll-out.

7.6.1 CBDC Chaincode Queries – Preserving Privacy

In a distributed ledger CBDC system, each peer has complete information about its balances and transactions. CBDC chaincode queries must ensure that DPP's, Enterprises and the Central Bank can only see information they are entitled to. DPP's will see their customers' balances and CBDC holdings but have no access to balances and holdings at other DPP's. The Central Bank, as owner of the multiparty ledger and manager of all DPP roles and permissions, will usually have audit/review access to all CBDC balances, holdings and payment transactions on its ledger.

Queries are implemented by chaincode. Normally, the chaincode will use the concept of "roles" to define what users can see or do. Checking supplied credentials on every query will ensure that all end user roles are restricted to only receiving CBCD balance and payment information to which they are entitled. Cryptography in the blockchain infrastructure is used to enforce these rules.

7.6.2 Know Your Customer

In our distributed ledger CBDC design, we have left Know Your Customer (KYC) capabilities to be entirely a responsibility of a DPP to handle for its customers and customer account. For the purposes of making a payment it is sufficient to have account identifiers for the source and target accounts. Each DPP is responsible for gathering contact and authentication information about its customers and storing this in some private DPP system with account specific credentials then managed and communicated in the CBDC chaincode API.

7.6.3 Directory Services

DPP's provide directory services that relate CBDC account identifiers with actual end customer information. This is used for payers to confirm that any payment is going into a valid account of the intended payee enterprise, and conversely confirm to the payee who a given payment is from. This information could be communicated in some off-ledger messaging between the end parties say in an invoice or payment notice.

It is likely that identifying payees from a CBDC account identifier and getting directory-based checking of a payment target account are likely to be important services on which DPP's compete to offer the most reliable and convenient service API to their customers.

7.7 Deployment and Scalability

The critical requirements for a distributed ledger based CBDC to operate with the approach we advocate above is that:

- It is highly available and reliable.
- It is in principle scalable allowing use of additional and larger processors to address high transaction volumes.
- That the consortium consisting of the Central Bank, the chartered DPP's and possibly regulatory organizations agree to use the CBDC smart contracts and chaincode, reviewing and approving updates for deployment as required.

This can be set up as a cloud service with an appropriate number of replicas to ensure high availability, processing power for each replica to meet transaction volume needs, and service interfaces for DPPs to integrate the CBDC chaincode operations with their own private ledgers and customer service management.

Scalability of distributed ledger solutions to support high volume payment transaction loads is improving over time but has not yet been fully established. Concerns stem in some degree that all update transactions against the ledger have to be ordered and applied to each instance of the distributed ledger in the blockchain network.

7.8 Slicing and Velocity

The exchange of CBDC between DPP D1 and DPP D2 is between the reserve accounts maintained by the DPPs. The actual disposition of the exchanged CBDC from and to enterprise is recorded on the private ledgers of the respective DPPs but does not need to be made visible outside these ledgers.

This detail is central to one key element of the design of the Project Aber CBDC implementation by the Central Banks of Saudi Arabia and United Arab Emirates [reference]. In Project Aber, reserve account transfers are bundled into "slices" in order to obfuscate the timing and amount of E1 to E2 payments.

We note also that one responsibility of the Central Bank is to manage the *economic* value of CBDC relative to currency. The rate of flow of CBDC – the velocity of money – is a critical metric in this regard. DPP reserve account transfers, perhaps disaggregated by region or industry, may be sufficient information to determine the velocity of money. Details of inter-enterprise flows may only be necessary when investigating fraud or money laundering activity.

7.9 Conclusions

This is not exactly a real-time payment using Central Bank Money with Central Bank taking all counterparty and transfer risk, but it is very close to it. It may indeed be best and most feasible approach for providing real-time enterprise CBDC payments within a single national currency domain.

8 Global Real-Time Payments with Interconnected CBDCs

Global enterprises need global payment systems – so let us explore how one may extend the DPP model to payments between enterprises in different CBDC domains. In this section, we introduce a new type of agent called a Digital Foreign Exchange (DFX) provider, through which end-to-end cross-domain payments can be executed as a series of chained payment legs.

Central Banks and Monetary authorities will require the independence to manage their issuance of CBDC. But we will suppose that there are inter-domain agreements concerning the DFX. The key requirement is that DFX agents must be able to write DFX transactions into the distributed ledgers supporting the various CBDCs. Such agreements will necessarily be quite complex, respecting the various business and legal structures in each domain. This section explores this concept with a simple implementation of a basic service we call *reconcilable linkage* that has much in common with the Continuous Linked Settlement process currently used for the settlement of foreign exchange.

8.1 Digital Foreign Exchange (DFX) Services

The primary strength of CBDCs is that they are denominated in a national currency and maintained at par with that currency – a property called "stablecoins" in the literature. But the relative nominal value of different national currencies, and hence CBDCs, may fluctuate over time, perhaps in response to market forces or perhaps due to policy changes.

The reconcilable linkage inter-CBDC payment process begins with DFX service providers, who publish – on the originating CBDC ledger – a contract offering an exchange rate and volume at which they will digitally process a payment between CBDCs. The DFX contract is a commitment to digitally receive a payment from a payer, and – *as part of this contract* – commits to making a payment of the corresponding DFX-rate amount in the destination CBDC. Each critical step in the chain of events to execute this contract must follow the rules of, and be validated by chaincode, in the originating or destination CBDC. First and foremost, each step must be recorded as a transaction on one of the distributed ledgers supporting the CBDCs.

DFX service providers need to be authorized and chartered by the respective Central Banks to be able to read and write to the CBDC distributed ledgers. DPP's could also be DFX providers, but the evolution of conventional currency FX markets suggests that DFX operations in any volume would require specialized skills and expertise different from the core DPP service of providing enterprises with CBDC accounts. By being authorized in both CBDCs, DFX providers can naturally provide both forward and reverse CBDC conversion services, hence providing crucial arbitrage services matching supply and demand for cross-domain flows. As with regular FX, competition between DFX providers will limit spreads and bound the cost of cross CBDC payments to user enterprises.

8.2 Reconcilable Linkage

Distributed ledger technology enables all participants in a single CBDC to agree on the state of their holdings at any point in time without complex reconciliation processing. Chaincode, with logic visible to all parties, ensures that sufficient funds are available and that balances are adjusted at the atomic moment when a payment is committed to the shared ledger. All the payment commits are serialized by an ordering service provided as part of the distributed ledger. Participants may not see the payment record at the same instant, but all eventually will see exactly the same record with the same sequence number processed by the same chaincode in the shared ledger.

The challenge for end-to-end payments across linked but autonomous CBDCs is to show that the same processing guarantees for single CBDCs can be achieved when there are two independent ledgers. We introduce, and articulate below, the principle of '*reconcilable linkage*' – a chaincode design pattern that can provide irrevocable and immediate evidence in the domestic *ledger* that the leg in the *foreign ledger* has been completed. Reconcilable linkage chaincode uses *cross references* to transactions on the partner ledger to enforce the proper sequencing of each sub-transaction.

End-to-end payments across two CBDCs cannot be atomic – that is not possible for independent ledgers – but they can use a combination of chaincode protocol and encryption technologies to link the associated transactions. Essentially the initiating (domestic) payment leg is held in a reversible state until it receives irrevocable confirmation that the second (foreign) payment in the other CBDC ledger is complete.

8.3 Logic of Reconcilable Linkage

We illustrate the logic of reconcilable linkage in the context of an enterprise E1 making a payment from its CBDC_A account to settle an obligation in CBDC_B to an enterprise E2.

Let us suppose that enterprise E1 has selected the Digital Foreign Exchange service DFX1 for this payment. DFX1's published exchange rate on the CBDC_A ledger determines the amount of CBDC_A needed to settle enterprise E1's CBDC_B obligation to enterprise E2. The critical requirement of DFC is that both CBDC distributed ledgers support *escrow states* for tokens.

The *available* escrow state is created by CBDC_B chaincode on tokens owned by DFX1. In this escrow, the CBDC_B tokens are only usable as part of an end-to-end payment from CBDC_A to CBDC_B. The *assigned* escrow state is created by CBDC_A chaincode on tokens owned by the payor, enterprise E1. In this escrow, the CBDC_A tokens are associated with enterprise E1's payment to enterprise E2.

Tokens in these escrow states are irrevocably committed, but actual ownership is not surrendered until the end-to-end payment completes. The process goes as follows. When DFX1's available escrow in the CBDC_B ledger is passed the details of enterprise E1's creation of the assigned escrow to enterprise E2, the CBDC_B chaincode will transfer tokens to enterprise E2. Up until this point, the first leg of the payment was completely reversible. Then, when the CBDC_A assigned escrow is made aware of the linked end payment completion, the chaincode releases the escrow and transfers the tokens to DFX1.

At this point the end-to-end payment is complete. Enterprise E2 has received the agreed amount of CBDC_B in its account, and enterprise E1 has a record on the CBDC_A ledger, including a CBDC_B transaction reference, to show that the end-to-end payment has been made. DFX1 has free use at this point of the CBDC_A tokens it received from E1's DPP.

The overall end-to-end payment flow is illustrated in Fig. 4.

With the reconcilable linkage protocol both the CBDC_A and CBDC_B ledgers have a complete record documenting its legs of the end-to-end payment from E1 to E2. Reconciliation is provided with minimal increase in complexity over the single CBDC case. Participants in local domains will need to have confidence in chaincode executing in the remote domains: to ensure that remote available escrow balances are being handled correctly and that payment-complete messages have valid CBDC_B transaction references. The reason for holding local CBDC_A tokens in escrow is

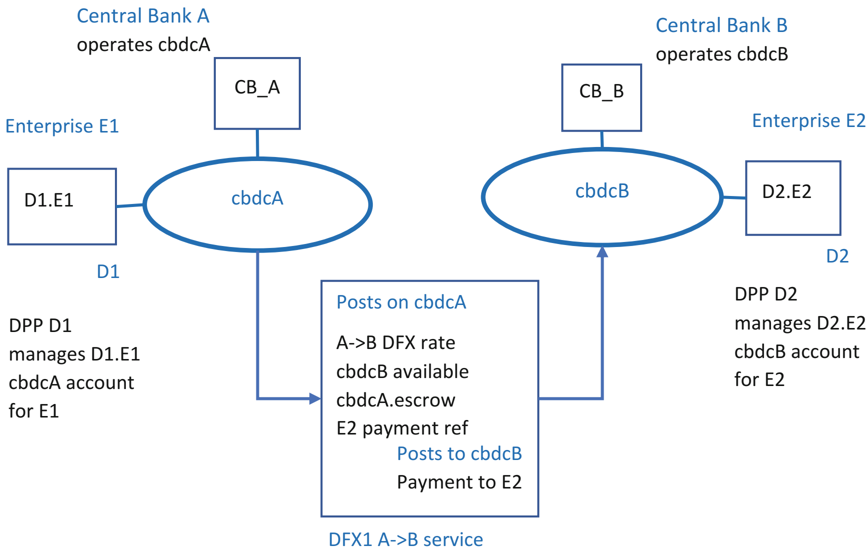

Fig. 4 Cross CBDC payments with reconcilable linkage

to ensure that these tokens – with value guaranteed at par by Central Bank A – can always be returned to E1 should the CBDC_B payment leg fail. Reasons which could possible cause a failure could include (1) failure of DFX1 before the payment request was submitted, (2) intervention by Central Bank B to prevent the payment (as owner of CBDC_B) or, (3) failure of E2's DPP provider.

The chaincode logic for reconcilable linkage builds on and extends that for a single CBDC payments system that we implemented in our demonstration prototype.

9 Summary and Conclusion

In this paper we applied a technology perspective to explore monetary innovations that could facilitate enterprise digital payments being settled in near real-time with low overhead cost.

Inter-bank payments are moving towards real time settlement with RTGS, but such developments apply only within single currency domains, do not prevent delays in posting credits to an end user account, and are not likely to be extended to large inter-enterprise payments. Emergence of Bitcoin, Ethereum, R3 and other cryptocurrencies established the feasibility of multiparty ledgers. But value swings in these digital assets make them speculative investments rather than useful vehicles for payment. Stable coin proposals such as Facebook's Libra are designed as payment systems but lack a clear explanation of how they would be regulated and governed so as to establish themselves as a trusted public good.

Our examination of these trends from the point of view of inter-enterprise payments led us to several conclusions:

- Central Bank currencies are the only medium of exchange that can provide the scale and stability needed for enterprise payments.
- Payment with digital currencies introduces novel and desirable features, but these are perhaps not relevant nor cost-competitive for all payment use cases – hence payment with digital currencies will very likely co-exist with payments supported by RTGS.

Correspondingly, we noted the key structural proposals of the Bank of England working paper (Kumhof & Noone, 2018) that enable Central Banks to manage the introduction of CBDC to minimize the risk to commercial banking services, namely:

- CBDC pays an adjustable interest rate.
- CBDC and bank reserves are distinct and not convertible into each other.
- No guaranteed, on-demand conversion of bank deposits into CBDC.
- CBDC is issued only against collateral consisting of eligible securities.

Finally, we noted that payment services provided by the commercial banking system are actually a bundle of several services of which ledger entry is only one part. Additional, off-ledger services, such as financial crimes monitoring and wholesale liquidity provision are also part of the payment service bundle. The most natural means to provision such off-ledger services would likely be based on and adapted from the existing banking infrastructure.

These considerations lead us to the proposal that a Digital Payment Provider (DPP) model for CBDC payment services – narrow banks that provide account services for payments in CBDC – is the most natural arrangement for enterprises to receive the benefits of real time payments and the associated innovations of distributed ledger technologies.

We explain how the use of a permissioned distributed CBDC ledger puts the Central Bank in a position to (1) establish the ledger, (2) commission DPPs and provide them with access rights to process CBDC payments while, (3) retaining complete control of the supply of CBDC tokens and hence the ability to make CBDC digital money part of its monetary policy. We illustrate the DPP model by working through the before and after balance sheets for each required step: the acquisition of CBDC, establishing and funding an enterprise CBDC account, and making CBDC real time payments to another enterprise. Taken together, these balance sheet transitions are a specification for a simple implementation of CBDC operations in a multi-party ledger.

Distributed ledger technology enables payments end-to-end from payer to payee as a single atomic transaction in a shared ledger, hence removing many of the duplications and inefficiencies of the existing payment infrastructure, including the need for mirrored system of record for reconciliation. Chaincode is the automation mechanism that enables end-to-end payments to settle in near real time. Chaincode executes on all instances of the distributed ledger, validates requested payments before recording them in the ledger and also enforces access control on ledger queries. This ensures that enterprise end user account balances are private – visible only to the specific DPP providing the account and possibly in audit to the Central Bank and its regulators. We used Hyperledger Fabric infrastructure to develop a prototype demonstration of DPP chaincode. Our explanation of how chaincode works to establish trust in the system and deliver real time payments is based on this prototype.

Global enterprises need cross-currency payments which settle in real time. We introduce and describe a linkage protocol, easily implemented by extending our chaincode, that ensures either both legs of an end-to-end cross-domain payment commit on their respective ledgers or neither does – a technique analogous to two phase commit in transaction systems. This simple extension demonstrates the potential for global atomic peer-to-peer payments, even though each Central Bank manages its own CBDC on its own autonomous distributed ledger. The required cooperation is minimal: authorizing Digital Foreign exchange providers and supervising the chaincode that implements the exchange services.

One relevant topic we have not addressed in this paper, is the ultimate performance and scalability achievable with distributed ledger systems. Engineering and gathering benchmarks to understand this is beyond our scope. Permissioned networks can reduce the overhead of consensus for the distributed ledger but so far to our knowledge there are no deployed systems operating with transaction rates of the order of tens of transactions per second.

Based on these arguments we claim that if adequate performance can be demonstrated, CBDC systems implemented with permissioned distributed ledgers are technically capable of providing a trusted, low-cost real time payment service for enterprises.

References

Bank of International Settlements Innovation Hub (2021). Available at https://www.bis.org/about/bisih/topics/cbdc.htm. Downloaded 2 December 2021

Barontini, C., & Holden, H. (2019, January). *Proceeding with Caution: A survey on Central Bank Digital Currency*. Bank for International Settlements(BIS) Papers 101.

Bech, M. L., & Hobijn, B. (2006, September). *Technology diffusion within central banking: The case of real-time gross settlement*. FRB of New York Staff Report No. 260. Available at SSRN: https://ssrn.com/abstract=932596 or https://doi.org/10.2139/ssrn.932596

Birch, D. (2020, April). *The edigital currency revolution*. Center for Study of Financial Innovation (CSFI) Number 134. Available at http://www.csfi.org

Klein, A. (2020, April). *China's digital payments revolution*. Brookings Institution report. Available at https://www.brookings.edu/research/chinas-digital-payments-revolution/

Kumhof, M., & Noone, C. (2018, May). *Central bank digital currencies —Design principles and balance sheet implications*. Bank of England Staff Working Paper No. 725.

Nakamoto, Satoshi (pseudonym). (2008, October). *Bitcoin: A peer-to-peer electronic cash system*. Available at https://bitcoin.org/bitcoin.pdf

Patel, B. (September 2018). *Central bank digital currencies*. Official Monetary and Financial Institutions Forum and IBM Blockchain joint report. Available by request from https://www.omfif.org

Integrated Framework for Financial Risk Management, Operational Modeling, and IoT-Driven Execution

Stephan Biller and Bahar Biller

1 Introduction

It is critical for companies to use system-level thinking effectively to support the decisions for their production system design and the activities surrounding the achievement of a unit cost which is low enough to be competitive in the market. The goal of this chapter is to describe the business analytics effort industrial companies should follow as the best practice to enable informed decisioning. The initial focus of the analytical modeling in this chapter is on supporting strategic investment decision making for a manufacturing facility. This is followed by addressing strategic-operational transition as the focus of analytical models moves towards daily management of operations. The representative timeline under consideration is illustrated in Fig. 1.

More specifically, we consider a hypothetical situation of having developed a new product for a competitive market and assume high exposure to both price and demand risks. The first step is to build a production facility to manufacture this new product for which there does not exist a verified process flow yet. The primary strategic investment decision is to select an equipment portfolio for the new facility. The equipment portfolio decision is followed by the arrival of the purchased equipment to the new facility. Unpacking each equipment presents an opportunity to collect data to verify the process assumptions made during the phase of solving the equipment portfolio selection problem. This is followed by observing the yield

S. Biller (✉)
Advanced Manufacturing International Inc., Clearwater, FL, USA
e-mail: stephan.biller@advmfg.org

B. Biller
SAS Institute Inc., Cary, NC, USA
e-mail: bahar.biller@sas.com

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_6

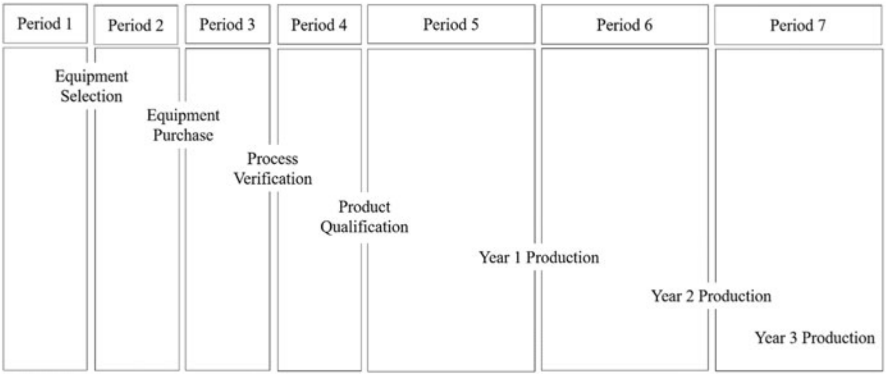

Fig. 1 A representative investment decision timeline spanning from strategic to operational phases

realizations, pushing the engineers to investigate how to improve the existing yield which will be critical to meet the production targets in the following years. Often, the target volume of production is relatively low in the first year, i.e., year 1 serves as a pilot production period. It is important to plan and optimize operational decisions involving equipment, process, and people management to increase the volume of production over the years to reach a unit cost of production and product quality to successfully compete in the market.

The objective of this chapter is to describe how to conduct a project with the timeline presented in Fig. 1 by using an integrated framework of financial risk management, operational modeling, and IoT-driven execution. The equipment portfolio selection problem is the strategic decision we choose to focus on during this description. However, its solution requires the consideration of the underlying supply chain including the portfolio selection of both suppliers and end products to be customized to meet the demand which is highly uncertain at the time of equipment portfolio selection (see Fig. 1). A high-level view of the supply chain is illustrated in Fig. 2 where types of all strategic decisions and their characteristics are given.

The remainder of the chapter is organized as follows. Section "Strategic Equipment Portfolio Selection" provides a detailed description for the equipment portfolio selection problem and presents a high-level process map for managing the risk and the value involved in the decisioning. Section "Market-Operations-Finance Integration" discusses market-factory-finance integration together with introduction of key performance indicators and a generic example of financial proforma that applies to manufacturing settings. This is followed by section "Real Options: Managing Risk in an Uncertain World" discussing the role of real options in effective management of risk in an uncertain world. Section "Process Digital Twin: Operations Optimization Decision Support" describes the process digital twin technology enabling operational modeling and IoT-driven execution to support both long-term scenario planning and near-term prediction effort. Section "Conclusion"

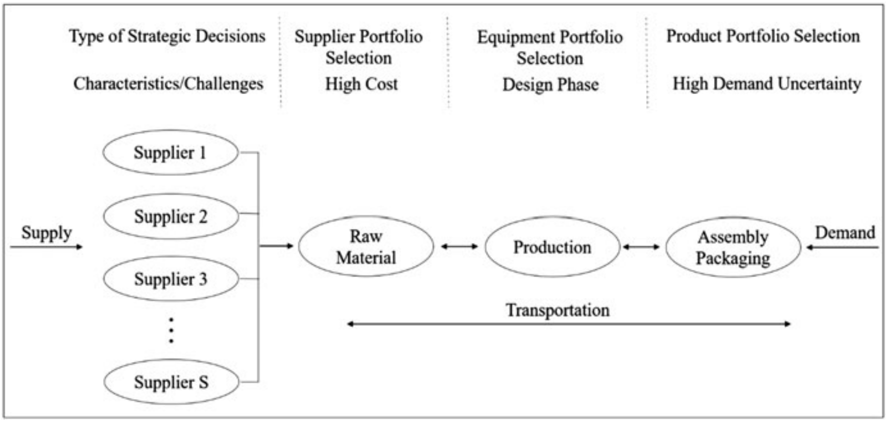

Fig. 2 A high-level view of the supply chain

concludes with our view of the challenges that arise in the integration of financial risk management with operations modeling and IoT-driven execution.

2 Strategic Equipment Portfolio Selection

The solution to the equipment portfolio selection problem is expected to consist of the number of equipment of each type to purchase to reach the business production target with minimum shortfall and capital expenditure (CapEx). Therefore, the equipment portfolio selection problem is generally formulated as the minimization of the expected production shortfall (i.e., the amount by which annual production falls short of the target specified in the business model) while meeting the CapEx budgetary constraints. It is also critical to quantify the financial risk associated with the selection of the equipment portfolio. At first glance, the decision variables of this equipment portfolio optimization problem appear only as the number of equipment of each type to purchase. However, the solution is to be accompanied both by the identification of management policies for raw material release and control of lead-time and inventory and by the determination of the minimum operator staffing needs as a function of the number of shifts on a production day and the number of production days in a week. It is also important to have a good understanding of the performance of the proposed equipment portfolio with respect to the manufacturing system bottlenecks and to propose tactical-level solutions for bottleneck management. Such visibility into the future operations of a production facility — which does not exist yet — would be ideally enabled by its process digital twin. Challenges of developing process digital twins and maintaining them are discussed in detail in section "Process Digital Twin: Operations Optimization Decision Support".

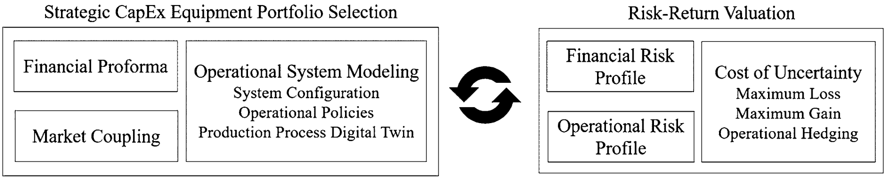

Fig. 3 A high-level process map for risk and value management

Furthermore, the financial risk associated with the selection of an equipment portfolio should be quantified. It is because of the need for such an integrated solution to the strategic equipment portfolio selection that the ideal practice is to develop three different modules, namely, operational system modeling, market coupling and financial proforma, to solve the problem in several steps, each with its own objective:

1. To build the production capacity risk profile of the manufacturing facility.
2. To identify a robust solution to the equipment portfolio selection problem.
3. To make operational recommendations using chosen equipment portfolio to achieve production target.
4. To derive the financial risk profile of the manufacturing investment by (i) characterizing the distributions of the unit cost of production and the net present value and (ii) studying their sensitivities to the business model assumptions on raw material cost and aggregate yield, both of which are highly uncertain at the time of initial investment.

The first three steps fall under the umbrella of operational system modeling while the last step requires integration of market, operations, and finance components of the project under consideration. Figure 3 provides an illustration of this modular approach to the solution of the equipment portfolio selection problem and demonstrates how the operational system modeling, market coupling and financial proforma modules come together to achieve the project objectives via risk and value management.

3 Market-Operations-Finance Integration

Figure 4 presents an alternative view of the market-operations-finance integration, which is critical both for solving the equipment portfolio selection problem and for seamless transition into the next phases of the project lifecycle. At the center of this integration lies the operational system modeling module. It is the core of the business analytics effort to support production system design. It is jointly defined by (i) the system configuration characterized in terms of equipment, people, and inventory; (ii) the operating policies for the management of lead-times, inventory,

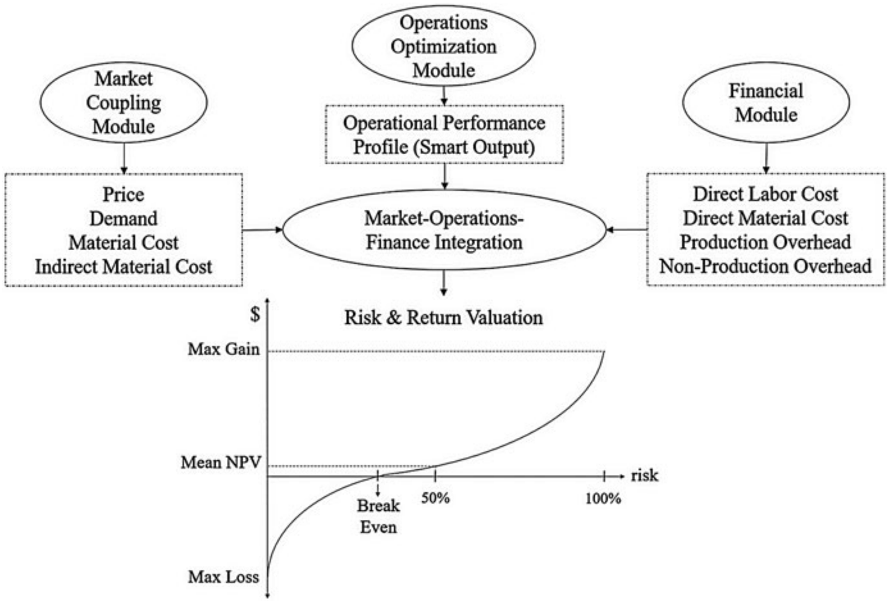

Fig. 4 Risk and return valuation

and operator staffing; and (iii) the production process digital twin. The system configuration and the operating policies are among the inputs to drive the process digital twin from which the production capacity risk profile is obtained. Combination of production process digital twin and optimization with the principles of operations management enables long-term scenario planning. This capability leads to the identification of a robust solution to the equipment portfolio selection problem and the recommendation of effective operational actions to achieve business production targets while using the chosen equipment portfolio.

Market coupling is the result of accounting for price risk and raw material cost. It is customary to further utilize a previously conducted market analysis presenting future demand and competition projections. The financial proforma module takes the price risk quantified by the market coupling module and the production capacity risk profile obtained from the operational system modeling module as inputs; uses Monte Carlo simulation for risk analysis; and presents financial risk profile as the output. The resulting integration creates smart outputs for populating the financial module and couples the module with the market drivers. The risk-and-return valuation is the outcome of this market-operations-finance integration. The analysis of the simulation output data leads to the quantification of the cost of uncertainty and the financial impact of operational decisions.

Prior to the detailed discussion of the risk-and-return valuation illustrated in Fig. 4, it is important to introduce the key performance indicators and the generic financial proforma whose population with production and market risk profiles is critical to quantify risk and return. Section "Key Performance Indicators" discusses the key performance indicators of interest and section "Financial Proforma" describes a financial proforma that is relevant to manufacturing operations management.

3.1 Key Performance Indicators

It is the common practice to conduct system performance analysis with focus both on operational performance standards and on financial performance standards. The operational performance standards include the annual throughput (and hence, the production shortfall), lead-time, utilization, and inventory (queue sizes and queue times). The production shortfall captures the amount by which annual production falls short of the given annual production target. The lead-time represents the number of calendar days – accounting for both production days and non-production days – needed to finish the production of a single stock keeping unit. The equipment utilization measures the fraction of time during which the equipment is busy. In a similar manner, the fraction of time an operator is busy is presented with the operator utilization. The work-in-process (WIP) provides a quantification for the amount of inventory (i.e., the total number of lots) in the system, including those waiting for processing and those that are being processed by the equipment. Figure 3 refers to the probability distributions of these performance indicators as "Operational Risk Profile".

The financial performance standards are, on the other hand, the break-even probability, the minimum and maximum losses, the unit product cost and the net present value of the manufacturing investment. The computation of their probability distributions is displayed under "Financial Risk Profile" in Fig. 3. Specifically, the unit product cost is compared to the future projections of the unit product selling price via the coupling of the market module with the operational system modeling and financial proforma modules. Consequently, we obtain the net present value distribution which enables the quantification of expected profit, break-even probability, and minimum and maximum losses. Utilizing this information gained from the strategic phase of the project, the purchasing decision is made for the proposed equipment portfolio. As the ordered equipment arrives at the facility, the plan moves into the phase of collecting data from the purchased equipment as part of the qualification process and learning from data to improve the process digital twin and risk predictions.

3.2 Financial Proforma

The financial proforma module calculates the cash flows and forms the core of the net present value computation. This is where the cost framework is established to estimate the unit cost for a single stock keeping unit. The cost components are

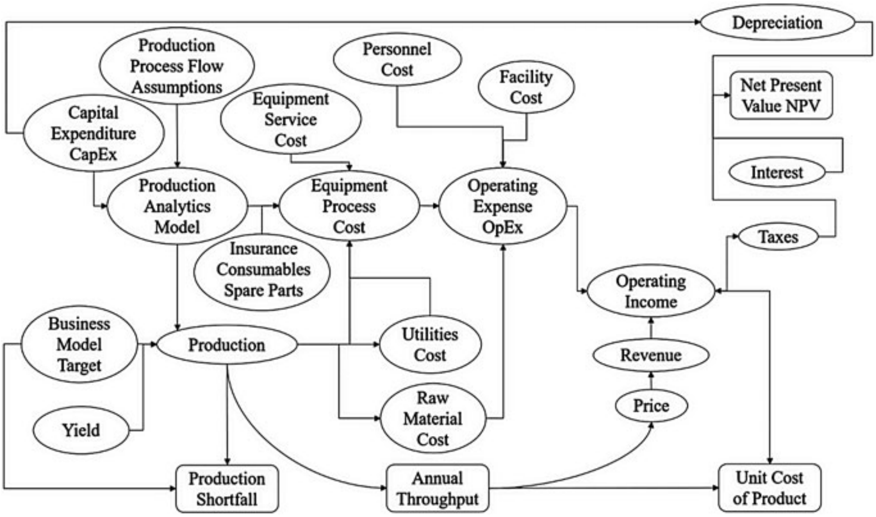

Fig. 5 A generic financial proforma to support product manufacturing

often divided into four different categories as illustrated in the financial module of Fig. 4: (1) direct material cost; (2) direct labor cost; (3) production overhead; and (4) non-production overhead. The sum of the direct material cost, the direct labor cost and the production overhead provides the total production cost. Dividing the total production cost by the product of the number of finished products and the yield results in the unit product cost. We study the contribution of each cost category to the unit cost of production to determine the key cost drivers, which require immediate attention.

Biller et al. (2019) provides details of a financial proforma developed to support semiconductor manufacturing investment decisions. Figure 5 provides a higher-level illustration of that financial proforma to relate to generic manufacturing investment projects. Fig. 5 is also combined with the outputs of the operational system modeling module (i.e., annual throughput and production shortfall) and the market coupling module to present unit product cost and net present value. Furthermore, the direction of the arrows indicates the order of the calculations that eventually lead to the quantification of the risk surrounding the operational system design.

3.3 Risk and Return Valuation

The final component of Fig. 4 is the illustration of risk and return on a two-dimensional plot. Specifically, the x-axis represents the risk (i.e., a probability taking values between 0 and 1) while the y-axis provides the net present value ($). The two-dimensional plot of Fig. 4 is an alternative view of the cumulative distribution function of the net present value. A traditional cumulative-distribution-function plot places the possible values of the net present value on the x-axis while the probability of attaining that net present value or less appears on the y-axis. It has been our experience that the form of the risk-and-return illustration in Fig. 4 is intuitive and often grasped quickly by the stakeholders struggling with the quantification of the risk they are exposed to. In particular, observing minimum gain, maximum gain, expected net present value and break-even probability are the key statistics that summarize the financial risk profile.

However, there are two main challenges to overcome in order to accurately quantify risk and return for strategic investment decisions under high uncertainty:

1. The first challenge is due to the lack of full information about the business process flow and its assumptions (and parameters): this is known as the input uncertainty. There exists a well-established body of methodological work, which should be utilized to account for this additional layer of input uncertainty in risk-and-return valuation. We refer the interested reader to Biller et al. (2017a, b, 2019) as examples of the industrial applications where input uncertainty is factored into models of strategic decision making in manufacturing. We also caution that the probabilistic models of input uncertainty should be revised over the course of the project lifecycle, especially as the uncertainty is realized over time.
2. The second challenge is the calculation of the net presented value with discounted-cash-flow view.

Next, we discuss how the use of real options would help overcome the second challenge.

4 Real Options: Managing Risk in an Uncertain World

Real options define ways to lead in a dynamic environment. Investing today in R&D or in a new marketing program, or in capital expenditures such as phased plant expansions generates the possibilities of new products or new markets tomorrow. When evaluating corporate investment opportunities (such as the equipment portfolio selection), management should embed real options into actions. The goal of this section is two-fold:

1. To reshape thinking about strategic investments with real options approach to learn and adapt to win
2. To enable delivery of results in an uncertain world when faced with complex, risky strategic decisions.

Real option is the right, but not the obligation, to take further strategic action at a future date with respect to the underlying investment (Luehrman, 1998). Option to change decisions later, based on the actual outcome, forms an integral

Table 1 An Illustrative Example: Impact of Real-Option Valuation

	Option-free valuation	Real option valuation
Maximum loss	\$29M	\$6M
Expected NPV	−\$14M	\$9M
Maximum gain	−\$2M	\$31M
Chance of losing money	100%	10%
Expected cost of uncertainty	\$34M	\$11M

component of projections. It paves the path to success by learning and adapting (Ries, 2017). Thinking of the investments in terms of embedded options accounts for uncertainty and manages risk, increasing the likelihood of delivering results in an uncertain world through experimentation, learning and iteration. More specifically, decisioning is exposed to high uncertainty, requiring waiting for more information. While conventional tools fail to capture upside potential in strategic decisions, value lies in growth options. Thus, the net present value of the project is written as the sum of the present value of expected future cash flows and present value of real options. The traditional net-present-value calculations would miss the real-option component and underestimate the return of an investment, which may be significantly profitable.

The next step is the identification of the types of real options that would apply to strategic investment decisions. While the following list is not exhaustive, it includes those that frequently arise in manufacturing settings:

- The option to expand production capacity;
- The option to manufacture raw materials in house;
- The option to manufacture a flexible mix of end products;
- The option to switch equipment and/or process technology;
- The option to abandon the entire new-product manufacturing project.

Table 1 presents a numerical example, which is representative of our past industrial decisioning experience when we were faced with a complex and risky strategic decision and using the real options approach (as described above) enabled us to correctly quantify the net present value. Figure 6, on the other hand, illustrates the tabulated impact of the real-option valuation. Specifically, the left-hand-side of Fig. 6 describes the case of option-free valuation while the right-hand-side represents real option valuation case.

In each plot, the \$20M on the y-axis represents the net present value that is obtained from the financial proforma but without any operational system modeling and market coupling. Since this is a deterministic valuation, it is free of any uncertainty modeling. The risk-and-return profile, on the other hand, accounts for uncertainty, independent of whether the valuation is option free or real-option based. Thus, the difference between the net present value obtained from the financial proforma and the expected net present value obtained from the market-operations-finance integration quantifies the expected cost of uncertainty. In the case of option-free valuation, the expected cost of uncertainty is \$34M. However, the real-

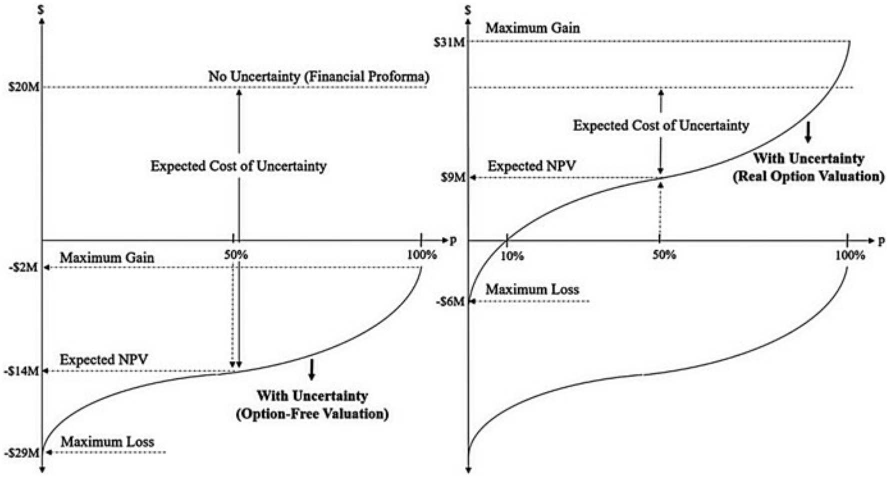

Fig. 6 Illustration of representative impact of real-option valuation

option valuation reduces the expected cost of uncertainty to $11M. Thus, the real options valuation would prevent us from overlooking the project which may result in long-term growth and profitability. By following this approach, we would be equipped to better manage the project by using learnings to adapt to win and deliver in an uncertain world.

5 Process Digital Twin: Operations Optimization Decision Support

Adoption of emerging technologies such as IoT and cloud has accelerated the digital twin development in a wide variety of industries. Today, digital twins have expanded to go beyond assets and include processes. Specifically, a process digital twin refers to a digital representation of a business process flow and equips its users with the power to predict the future performance of their operations. At the heart of this solution lies a flexible, data-driven, and scalable computer simulation mimicking the journey of thousands of objects flowing through the system and predicting future key performance indicators. The key advantages of any process digital twin are two-fold:

1. Providing enhanced visibility into the future of the system performance.
2. Enabling analysis of thousands of scenarios to perform risk-and-return tradeoff, enhancing resilience.

The need for the development of a process digital twin often arises due to the existence of a complex process flow and the input variability, which invalidates the

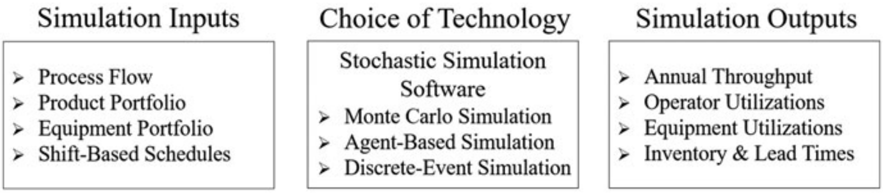

Fig. 7 Using stochastic simulation for building process digital twins

use of deterministic solution techniques to provide decision support. The resulting digital representation serves as a virtual lab to assess the operational policy impact and vast number of strategic investment decisions.

The production process digital twin corresponds to the digital representation of the manufacturing facility under consideration. The simulation of the factory operations would capture all relevant aspects of the process flow, equipment portfolio, product portfolio, and shift-based schedules. The discrete-event simulation methodology is often chosen to develop the simulation, which is driven by these inputs to predict a wide variety of key performance indicators, including throughput, inventory and lead-times, and equipment and operator utilizations. As illustrated in Fig. 7, the choice of simulation technology may range from Monte Carlo simulation to agent-based simulation and discrete-event simulation. Today, there exists a wide variety of commercial software products that enable the development of data-driven, scalable, and flexible simulations.

Despite its modeling flexibility, a simulation-based solution methodology provides random output, requiring the use of statistical methods for error control. However, the error in the simulation outputs can be controlled by setting the number of simulation replications to produce a standard error which is less than a prespecified fraction of a mean performance measure. We refer the reader to Law (2015) for the theory underlying the experimental design and analysis and to Johnson et al. (2016) for a detailed presentation of the application of this analytics approach to risk and value management at General Electric.

It is important to design a process digital twin to align with the speed of decision making. In any case, the technology impact of manufacturing plant digital twin development would be expected to be three-fold: (1) reducing development time; (2) learning in the design stage; and (3) understanding the system. It has been our experience from past industry projects that a digital twin can effectively reduce the development time from weeks to days. Any digital twin must be flexible, scalable, and driven by data. Despite initial focus on using the digital twin for design, any digital twin development effort provides support from design to real time. The best practice is to use the tool to improve manufacturing line performance continuously, for example, via real-time bottleneck detection.

An example of a process digital twin output dashboard is illustrated in Fig. 8. Such a dashboard would play a critical role in learning and understanding the system under consideration. It provides the benefits of (i) measuring the capability

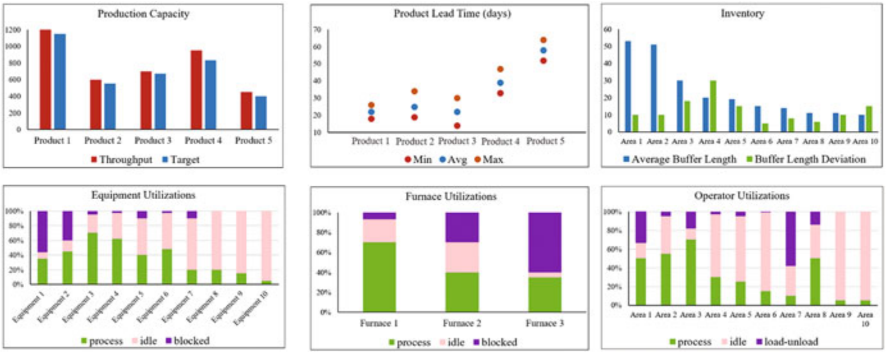

Fig. 8 An illustrative digital twin KPI dashboard

of the manufacturing lines to meet annual production targets for different products; (ii) predicting the manufacturing plant lead time; (iii) quantifying equipment utilizations; (iv) identifying the production areas to focus; (v) showing where inventory accumulates to contribute to space discussions; and (vi) determining minimum staffing needs and value of cross-training the operator. Using the process digital twin, we can further stress test the manufacturing facility and identify the best courses of action to take when low-probability, but high-consequence events would occur.

If the process digital twin is developed to support strategic decision making, then the underlying simulation model is often customized to predict steady-state performance. It is plausible to expect historical data and experts' opinions to be the two key sources of information in this case. However, there may be cases for which there would exist no reliable source of historical data to use, for example, when the product to be manufactured is new and the process flow to be followed has not been verified yet. In such a situation, we would rely on experts' opinions solely, and it becomes critical to quantify the impact of assumption uncertainty on operational and financial risk profiles. As the project moves from equipment selection to the arrival of equipment to the manufacturing facility, process verification and product qualification start, creating other sources of data that can be collected in real time. It is important to revise the previously developed stochastic models with the new data, combining experts' opinions with real-time data. It is also critical to periodically update these risk models as both process details and yield are continuously improved as the learnings of the engineers continue. The risk modeling assumptions would likely stabilize into the second year of production (Fig. 1). Still, it would be of value to account for different sources of uncertainty inherent in most manufacturing settings as well as for the disruptions that we would not expect to happen on any regular day.

The success of the strategic – operational transition is heavily dependent on the alignment of data collection with the analytical models associated with the different phases of the project; see Fig. 9. The model validation effort involves

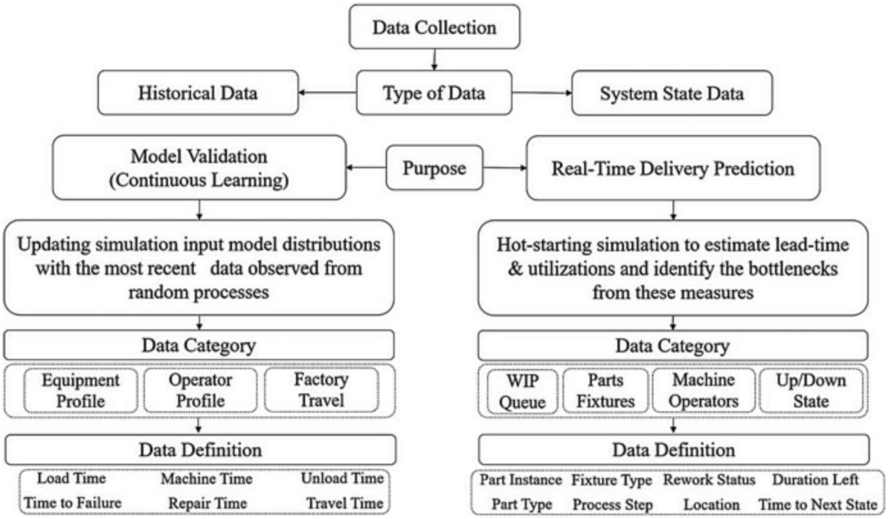

Fig. 9 Alignment of data collection with analytical modeling

updating simulation input distributions with the most recent data observed from random input processes to improve delivery prediction accuracy. The validation is followed by building capability to hot-start the simulation to enable near real-time prediction. In other words, the additional type of input data for the simulation would be the system state data, i.e., the IoT data collected from the sensors presenting a snapshot reflecting the state the system is in at that point in time. Finally, link is to be established between MES (Manufacturing Execution Systems) data and plant simulation. The expected impact includes the elimination of assumption uncertainty and the improvement of prediction accuracy with the use of IoT and market data; see Fig. 10.

It is important to note that the area of application for process digital twins goes beyond the walls of the manufacturing facilities and extends to supply chains and service operations; see Fig. 11. This enables the creation of scenarios to identify corrective actions in near real time and optimize planning and lifecycle costs while improving the reliability of the entire ecosystem. Each scenario may present a potential solution to a problem that may arise at a certain level, and the optimal action may be identified in near real time by conducting scenario-based simulations and comparing their effectiveness in terms of the key performance indicator predictions. The combined use of simulation, machine learning, and optimization in a parallel computing environment is expected to accelerate both speed and accuracy of the problem-solving effort. Finally, the digital thread is expected to make it possible to provide system-level information to assets, people, and process managers in real time, allowing continuous learning and improvement.

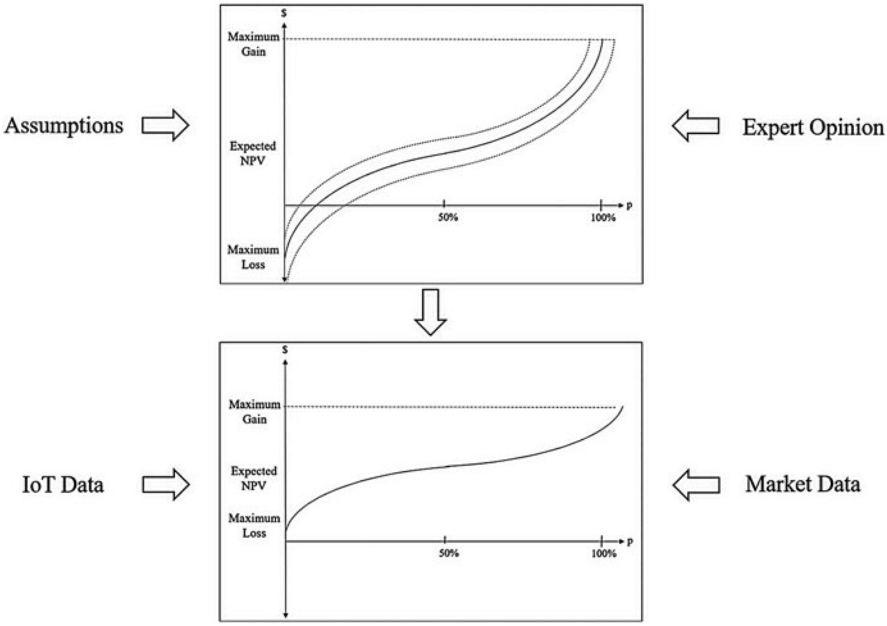

Fig. 10 Impact of IoT and market data on eliminating assumption uncertainty and improving accuracy

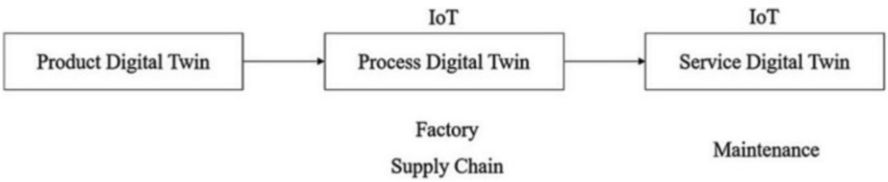

Fig. 11 Digital twin lifecycle

6 Conclusion

As an example of an industrial application of the integrated framework for financial risk management, operational modeling and IoT driven execution, we refer the reader to Biller et al. (2019) in which the use of process digital twin and risk analysis is discussed to support silicon carbide production system design at General Electric. Using the integrated framework, strategic-level equipment-portfolio selection and tactical-level operations-management recommendations are made for the silicon carbide manufacturing facility to ensure the attainment of the business production target with confidence; to minimize the expected production shortfall under the available CapEx budget; to determine the minimum operator staffing needs for the facility; and to provide management improved visibility into silicon carbide manufacturing. The resulting equipment portfolio is shown to eliminate the bottlenecks of the original portfolio, increase expected production capacity of the manufacturing facility by 67% with additional investment corresponding to less than 1% of the original CapEx. It is possible to achieve similar benefits by following an integrated framework of financial risk management, operational modeling and IoT driven execution.

References

Biller, B., Biller, S. R., Dulgeroglu, O., & Corlu, C. G. (2017a). The role of learning on industrial simulation design and analysis. In W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, & E. Page (Eds.), *Proceedings of the 2017 Winter simulation conference*. Institute of Electrical and Electronics Engineers.

Biller, B., Dulgeroglu, O., Hartig, M., Olson, R. J., Sandvik, P., & Trant, G. (2017b). Semiconductor manufacturing simulation design and analysis with limited data. In *The 28th annual SEMI advanced semiconductor manufacturing conference proceedings* (pp. 298–304). Institute of Electrical and Electronics Engineers.

Biller, B., Hartig, M., Minnick, A., Olson, R. J., Sandvik, P., Trant, G., & Sui, Y. (2019). General Electric uses simulation and risk analysis for silicon carbide production system design. *INFORMS Journal on Applied Analytics, 49*(2), 117–128.

Johnson, C., Biller, B., Wang, S., & Biller, S. R. (2016). Analytics approach for incorporating market demand into production design and operations optimization. In J. Li, S. Zhou, & Y. Han (Eds.), *Advances in battery manufacturing, service, and management systems* (pp. 99–126). Wiley.

Luehrman, T. A. (1998). Investment opportunities as real options: Getting started on the numbers. *Harvard Business Review*. Print 98404.

Ries, E. (2017). *The startup way: How modern companies use entrepreneurial management to transform culture and drive long-term growth*. Currency.

Market Equilibrium Models in Large-Scale Internet Markets

Christian Kroer and Nicolas E. Stier-Moses

1 Introduction

Firms in the technology industry often face situations where they must allocate goods to buyers, either literally or figuratively. Among them, internet companies have routinely employed mechanisms centered around auctions because they are robust and allow for changing market conditions and successful price discovery. Some of these mechanisms are 'static' in the sense that the whole market is cleared at once, while others are 'dynamic' meaning that decisions are made on a rolling horizon basis. In financial markets, to sell a newly released bond, potential buyers submit a supply function which specifies how many bonds they are willing to buy at what price. Then, the issuer computes a market-clearing price and use those functions to allocate bonds to buyers. In spectrum auctions, buyers and sellers submit combinatorial bids, and the market maker solves a large mixed integer program (MIP) to find the optimal allocation of spectrum to firms. In electricity markets, market equilibrium is used for pricing electricity in a way that incentivizes suppliers to generate the right amount of electricity. These prices are hard to compute due to non-convexities in the electricity production cost of a supplier (e.g., due to fixed costs of starting production), and integer programming is often used to compute these equilibria. In the technology industry, the volume of transactions and the dynamic nature of its markets make it hard to solve the whole allocation centrally and in one shot. Usually, firms resort to dynamic versions of the market that can be solved in a repeated way. For example, in the internet advertising use

C. Kroer (✉)
Columbia University, New York, NY, USA
e-mail: christian.kroer@columbia.edu

N. E. Stier-Moses
Facebook Core Data Science, Menlo Park, CA, USA

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_7

cases, an individual auction is run for the ad slots generated when a user interacts with the system. This may be triggered by a search query with keywords, by loading a news article, or by refreshing the feed of a social network. Similarly, there are applications to recommender systems used in the technology industry. In that setting, no explicit market and real money is exchanged. Nonetheless, the allocation of recommendation slots to different content creators can be modeled as a market with 'funny money,' where content creators use their funny-money budgets to optimize their allocation in recommendation slots to users. Another example of a market without money is robust content review problems, where different categories of sensitive content need to be reviewed, and the allocation of reviewing capacity towards the content categories can be modeled as a market allocation problem.

We will focus on the technology industry with its variety of use cases of market-equilibrium based allocations for divisible goods. These kind of market models can be used for a variety of purposes. The most immediate is to find a solution to these markets when the solution is needed and use it to perform the actual allocation. This would be comparable to the bond and spectrum markets mentioned earlier where the transactions are based on the solution to the market model. There are situations in which solving the problem in real-time is not feasible. In those cases, the solution may be computed offline and used as a benchmark. In an ex-post analysis, the firm can judge the merits of the allocation used in practice vis-a-vis the market approach and decide if the online solution approach should or should not be improved in terms of solution quality or computational efficiency. Yet another alternative is to use these models to compute features that can useful to forecast outcomes at the right level of granularity. If one would like to forecast relevant market metrics for next year—e.g., welfare, prices, revenues—running each ad auction individually, given that there might be millions of them per day, does not seem to be the right granularity. Instead, viewing the situation as recurrent realizations of a market for which we can predict the input parameters can provide a better handle to make the right forecast. Finally, another important use case is to evaluate counterfactuals. Having a market model that can deliver predictions allows us to change some of the underlying premises or interactions and find how the solution depends on those changes. An example of this could be to understand how a marketing promotion can provide incentives to advertisers and transform the resulting situation to another equilibrium.

An important factor in common in the use cases above is the need to *compute* solutions in those market models. It is not enough to know that a solution exist, one actually needs the solution itself to operate the system, to forecast it, or to make strategic decisions. To that end, we will discuss algorithmic approaches to solving these models, with a focus on large-scale methods.

To set the stage, we offer a few more details on these ideas by discussing how to use *fair recommender systems* on a job recommendations site. Such a site is a two-sided market. On one side are the users, who view job posts. On the other side, there are the employers creating job posts. Naively, a system might try to simply maximize the number of job posts that users click on, or apply to. This can lead to extremely imbalanced allocations, where a few job posts get a huge number of views and applicants, which is bad both for users and employers. Instead, the system may wish to fairly distribute user views across the many different job posts. To achieve a balance between fair distribution and market efficiency, market-equilibrium-based allocation can be used. In this setting the buyers are the job posts, and the goods are slots in the ranked list of job posts shown to job seekers.

In the next sections, we describe the various models that relate to the main use cases, including ad auctions, recommender systems, and fair division. Then we focus on algorithms and present several ideas that permit solving large-scale models as required by the use cases in the technology industry. This is an expository piece that exhibits existing theoretical research and computational studies done in the setting of internet-scale market applications.

2 Introduction to Fisher Markets and Market Equilibrium

This section introduces the *market equilibrium problem*, the basic modeling element of this chapter. We focus on a particular type of market, usually referred to as *Fisher market*, where there is a set of n buyers that are interested in buying goods from a distinct seller. Every buyer has a budget of $B_i > 0$ dollars. There is a set of m infinitely-divisible goods and each good j has a supply of $s_j > 0$ units that can be divided and sold arbitrarily. We refer to the full supply vector by s.

We will use $x \in \mathbb{R}_+^{n\times m}$ to denote an allocation of goods to buyers, where $x_{ij} \geq 0$ is the amount of good j that is allocated to buyer i. We also denote the bundle of goods given to buyer i as $x_i \in \mathbb{R}^m$ (the i'th row of x). Each buyer has a utility function $u_i(x_i) \mapsto \mathbb{R}_+$ that captures how much they like the bundle x_i. We make the following assumption to avoid degeneracy issues: there exists an allocation x such that $u_i(x_i) > 0$ for all buyers i. This means that it is possible to find an allocation such that all buyers get strictly-positive utilities.

When solving for *market equilibria*, the goal is to find a price $p \in \mathbb{R}_+^m$ for each of the m goods such that the market clears. Clearing the market means that there should exist a feasible allocation x such that every buyer is assigned an optimal allocation given their budgets and the prevailing prices. Formally, the *demand set* of a buyer i with budget B_i finds an optimal bundle under a set of prices by solving

$$D_i(p) = \operatorname{argmax}_{x_i \geq 0} \{u_i(x_i) : \langle p, x_i \rangle \leq B_i\}.$$

A market equilibrium is an allocation-price pair (x, p) such that every buyer gets an optimal bundle and goods are not oversold. Mathematically, that corresponds to $x_i \in D_i(p)$ for all buyers i, and $\sum_i x_{ij} \leq s_j$ for all goods j, where the inequality has to be attained if $p_j > 0$.

Market equilibria have been thoroughly studied and found to have many attractive theoretical properties. One of the most celebrated properties is their *Pareto optimality*: a market equilibrium allocation x satisfies that, for *every* other allocation x', if a buyer is better off under x', then some other buyer is strictly worse off.

In other words, x is such that no other allocation can simultaneously (weakly) improve all individuals' utilities. Either all utilities stay the same in other solutions or improving one buyer comes at the expense of another buyer. This is known as *the first fundamental theorem of welfare economics*.

Beyond Pareto optimality, there are several other interesting properties that are verified by market equilibria. They include envy-freeness, where every buyer prefers their own allocation to that of any other after correcting for budget sizes, and proportionality, where every buyer is at least as happy as if they were allocated a fraction of each good proportionally to their budget. These properties will be discussed in more detail when applying market equilibrium to fair division.

2.1 Convex Programming and Utility Functions

A very attractive feature of Fisher markets that make them particularly appealing for modeling purposes is that one can characterize equilibria in computationally efficient ways. Not only this implies that they are guaranteed to exist, but also that they are eminently *computable*, both in theory and in practice. Indeed, there is a nice convex program whose solutions satisfy the market equilibrium properties. Before writing the convex program, let us consider some properties that we would like an optimal allocation x to satisfy. As mentioned before, a feasible allocation necessitates the supply constraints to be respected: $\sum_i x_{ij} \leq s_j$ for all j.

Secondly, since a buyer's demand does not change even if we rescale their valuation by a constant, we require the optimal solution to the convex program to also remain unchanged. Similarly, splitting the budget of a buyer into two separate buyers with the same valuation function should leave the allocation unchanged. These conditions are satisfied by the budget-weighted geometric mean of the utilities:

$$\left(\prod_i u_i(x_i)^{B_i}\right)^{1/\sum_i B_i}.$$

Since taking roots and logs of the objective does not change the set of optimal solutions, we simplify the objective and include the supply constraints to get the so-called EG optimization problem:

$$\begin{aligned} \max_{x \geq 0} \ & \sum_i B_i \log u_i(x_i) \\ s.t. \ & \sum_i x_{ij} \leq s_j, \quad j = 1, \ldots, m. \end{aligned} \tag{EG}$$

We denote the dual variables corresponding to each of the supply constraints by p_j. If the utilities in EG are concave and non-negative then this yields a convex program, since composing a concave and nondecreasing function (the log) with a concave function (u_i) yields a concave function. Moreover, if the utilities are concave, continuous, non-negative, and homogeneous (CCNH) then an optimal solution x to EG satisfies the market equilibrium allocation conditions, and the dual variables p provide the equilibrium prices. Formulating the EG program was a seminal idea in the field of market equilibrium computation. It was originally done for linear utilities (which are CCNH) by Eisenberg and Gale (1959). The general CCNH case was shown by Eisenberg (1961) a few years later. A more modern derivation for differentiable CCNH utilities can be found in Nisan et al. (2007). For a derivation of the fully general CCNH statement with the more modern formulation of EG, see Gao and Kroer (2020).

Let us review the implications of having the EG formulation. First, it gives us an immediate proof of market equilibrium existence for the CCNH Fisher market setting: the feasible set is clearly non-empty, and the max is guaranteed to be achieved. Second, it allows us to show Pareto optimality directly. A maximizer of EG is indeed Pareto optimal since another solution that simultaneously improves all utilities would be feasible and have a strictly higher objective, contradicting optimality. Third, the optimality of a solution to the EG formulation can also be used to show from first principles that the equilibrium utilities and prices must be unique. If there were more than one allocation at equilibrium, then by the strict concavity of the log function we would get that there is a strictly better solution, which is a contradiction. Thus, the set of equilibrium utilities must be unique. From there it can be seen that equilibrium prices are unique as well, which follows from the EG optimality conditions.

2.2 *Classes of Utility Functions*

In the previous section we saw that the EG formulation can be used to compute a market equilibrium as long as the utility functions belong to the fairly abstract class CCNH. To understand this class and to provide more context on what is used in practice, we present some concrete examples of its most common types of utilities. To get intuition on the generality of the class, one should primarily consider the homogeneity constraint. Imposing homogeneity disallows many potential utility functions but we will see that it is still a fairly rich class.

The easiest example of a utility function is a *linear utility* $u_i(x_i) = \langle v_i, x_i \rangle$ where $v_i \in \mathbb{R}^m_+$ is a vector of per-good utility rates. It is immediate that linear utilities are CCNH. They can be useful for modeling internet markets—in particular, both ad auctions and fair recommender systems—so they are of special interest to this chapter. More concretely, ad auction models rely on *quasilinear utilities*, a slight variation of linear utilities, where buyers subtract the price that they pay: $u_i(x_i, p) = \langle v_i - p, x_i \rangle$. Technically, this does not fall under the EG framework, since the utility now depends on the prices p. However, it was shown independently by Chen et al. (2007) and Cole et al. (2017) that a small modification to EG can handle quasilinear utilities.

Beyond linear utilities, the next list enumerates the most famous utility classes within CCNH. Let us consider i to be an arbitrary buyer and a_{ij} to be calibration parameters for every good j.

1. Leontief utilities: $u_i(x_i) = \min_j \frac{x_{ij}}{a_{ij}}$,
2. Cobb-Douglas utilities: $u_i(x_i) = \prod_j (x_{ij})^{a_{ij}/(\sum_j a_{ij})}$,
3. Constant elasticity of substitution (CES) utilities: $u_i(x_i) = \left(\sum_j a_{ij} x_{ij}^{\rho}\right)^{1/\rho}$, where ρ is another calibration parameter, with $-\infty < \rho \leq 1$ and $\rho \neq 0$.

CES utilities turn out to be the most general so far: Leontief utilities are obtained as ρ approaches $-\infty$, Cobb-Douglas utilities as ρ approaches 0, and linear utilities when $\rho = 1$. More generally, $\rho < 0$ implies that goods are complements, whereas $\rho > 0$ implies that goods are substitutes.

An interesting consequence of the existence of the EG formulation is that various natural iterative economic processes converge to a Fisher market equilibrium. This is because many such processes are formally equivalent to some form of iterative first-order optimization on the EG program. For example, various *tâtonnement* algorithms converge to a Fisher market equilibrium. A tâtonnement process is an iterative dynamic where a market operator repeatedly announces prices p_t at each time t, each buyer i reports their demand x_i^t under the given prices, and the market operator increases the price of over-demanded goods and decreases those of under-demanded ones. This can be reinterpreted as subgradient descent on the dual convex program of EG.

Other interesting dynamics based on the EG formulation also exist. Perhaps the most important one is the *proportional response* process, where buyers iteratively specify how much they wish to spend on each good, and the market operator sets the prices to the sum of these spends. This dynamic turns out to perform extremely well in practice, and we will review it in detail later. This was discovered by Wu and Zhang (2007) when analyzing bit-torrent sharing dynamics, and (Birnbaum et al., 2011) later gave a surprising convergence guarantee based on a mirror-descent equivalence.

3 Auction Markets and Budget Management Systems

Advertisers participate in internet ad markets to get impressions, clicks, or conversions of ads that are placed in content shown to users by the platform. To accomplish this, advertisers set up ad campaigns that indicate how much they are willing to bid in exchange for those events. Since the values per conversion are unknown to the platform, in the last decade, platforms turned to computing allocations and prices by running an auction every time a user shows up, and the competition in these auctions gave rise to ad auction markets. In addition to values, ad campaigns usually specify budget or ROI (return-on-investment) constraints. This allows them to control their total spend and to maximize the value they get out of the system while guaranteeing that they do not exceed the maximum amount of money that they are willing to spend. Bidding is typically performed by a *proxy bidder*, operated by the platform but acting on behalf of a given advertiser. This proxy bidder attempts to maximize the utility derived by the advertiser, while taking into account the specified constraints.

When designing the market mechanism and the corresponding proxy bidders, the platform needs to provide tools to allow advertisers to run ad campaigns that are effective. One of the issues that arises is that budgets and bids in a campaign may not necessarily be in agreement with each other. In light of that problem, a platform may offer ways to compute alternative campaign parameters that align budgets and bids. The two fundamental budget management systems that are dominant in practice and in the literature are (a) throttling, which uses a feedback loop to limit the number of auctions an ad participates in, and (b) pacing, which uses a feedback loop to shade bids. We provide more details about these mechanisms below.

The purpose of this section is to illustrate how market models and their equilibria can be used as a tool to understand tradeoffs in auction markets and budget management systems. We will focus on systems based on pacing mechanisms, since that is one of the dominant budget management methods used in practice, and these systems are particularly amenable to analysis via Fisher market models. As a simplification, we assume that each individual auction allocates a single good, usually referred to as a 'slot.' This is a simplification that allows us to model the repeated auctions as a quasi-linear Fisher market, and hence make available all the theory and results that apply to their equilibria. In practice, it is common for platforms to simultaneously auction several *impression opportunities* (slots to be filled with ads) in real time when they display a page or refresh a feed.

It is important to highlight that in the real-time operation of a platform, instead of relying on market equilibria as considered here, they typically rely on control algorithms which tune the parameters used by the proxy bidder to align the advertisers' campaign parameters. (The parameters may include the pacing multiplier which is relevant to our model, but in other implementations they may include participation probabilities for throttling campaigns.) The market equilibria that we describe in this section can be thought of as the desired steady state of the system. In practice, the control algorithms need to learn these parameters in an online fashion. Budget constraints and other pacing aspects invalidate traditional guarantees such as the strategyproofness of second price auctions.

We analyze the pacing equilibrium problem that results from the pacing system when the underlying allocation is produced by either a second or first price auction, in that order. By reinterpreting this problem as a game where players choose pacing parameters, we connect the equilibria of those games to solutions to suitable Fisher markets. After the static analysis, we also discuss the effects of adding temporal considerations to the model to get a dynamic auction market. This more closely parallels how campaigns are tuned in practice. We will see that the static game representation provides a good approximation that can be used as a starting point of dynamic procedures.

3.1 Market Models for Pacing Systems

We define an *auction market* similarly to a Fisher market (Sect. 2). We consider a set N of n buyers and a set M of m goods. Buyer i has value $v_{ij} \geq 0$ for good j, and each buyer has a budget $B_i > 0$. Each good j will be sold by itself in a sealed-bid auction, using either the first or second price as a payment rule. To disregard trivial cases, we assume that for all buyers i, there exists some good j such that $v_{ij} > 0$, and for all goods j there exists i such that $v_{ij} > 0$. Let $x \in \mathbb{R}_+^{n\times m}$ be an allocation of goods to buyers, with associated prices $p \in \mathbb{R}_+^m$. The utility that a buyer i derives from this allocation is

$$u_i(x_i, p) = \begin{cases} \langle v_i, x_i\rangle - \langle p, x_i\rangle & \text{if} \langle p, x_i\rangle \leq B_i \ , \\ -\infty & \text{otherwise} \ . \end{cases} \tag{1}$$

We will use the abbreviations SP and FP for second and first price auctions markets, respectively.

Although auctions have several appealing properties when considered individually, budgets add a coupling constraint across auctions that influences those properties. For instance, it is well known that second price auctions in isolation are strategyproof, but the following example shows that second price auction markets with budgets are not: Consider an instance with two buyers and two goods, with valuations $v_1 = (100, 100)$, $v_2 = (1, 1)$ and budgets $B_1 = B_2 = 1$. If both buyers submit their true valuations then buyer 1 wins both goods, pays 2, and gets $-\infty$ utility. To fix this problem, each buyer needs to smooth out their spending across auctions to make sure that they remain within budget.

For large-scale internet auctions the smoothing is frequently achieved via *budget management systems* as mentioned at the beginning of this section. The following two mechanisms are widely used in practice by platforms. In both, each buyer i (or proxy bidder, as it may be) has to tune a parameter $\alpha_i \in [0, 1]$.

1. *Probabilistic throttling*: The parameter α_i encodes the probability that the buyer participates in each auction. For each auction j, an independent coin is flipped for buyer i. If it comes up heads (with probability α_i) then the buyer participates in the auction with a bid $b_{ij} = v_{ij}$. Otherwise the buyer is excluded from that particular auction.
2. *Multiplicative pacing*: The parameter α_i acts as a scalar multiplier on the reported bids from the advertiser. For each auction j, buyer i submits a bid $b_{ij} = \alpha_i v_{ij}$.

Fig. 1 Comparison of budget management systems. *Left*: no budget management, *middle*: probabilistic throttling, *right*: multiplicative pacing

Figure 1 illustrates these options under second price auctions, in a simplified setting. For ease of presentation, the figures plot the opportunities in terms of time along the x-axis, even though these market abstractions are static. Time is inconsequential in this section, but we will revisit time in depth when we address dynamics in Sect. "Dynamic Budget Management Systems". We consider a focal buyer whose value is constant in all auctions and hence the bids are constant across them. The buyer participates in auctions as long as some budget remains, and then participation stops. Competition arising from other buyers present in the auctions cause resulting prices, plotted in the y-axis, to vary in different auctions. Since the figure represents second price auctions, the price in each auction is not necessarily the same, even though the focal buyer bids a single fixed amount. The circles represent the participation opportunities of the focal buyer and the shaded ones represent the auctions in which the focal buyer won.

The left panel shows the outcome if the budgets are not managed and buyers bid naively: the focal buyer spends the budget too fast, and ends up running out of money prematurely. There are many low-price and high-value goods to the right of the budget exhaustion line that the buyer cannot get. This is a lost opportunity for the buyer. Furthermore, in practice, buyers tend to prefer to smoothly spend their budget throughout the day as opposed to running out of money long before the end of the planning window. The middle panel shows the effect of probabilistic throttling for an appropriately chosen parameter α_i. Buyers only participate in some auctions, allowing them to continue to have a remaining budget until the end of the planning horizon. As before, buyers end up winning some expensive auctions, while missing out on cheaper ones. From the buyer's perspective this is still sub-optimal in terms of utility, since all goods have the same value to the buyer. Finally, the right panel shows the effect of multiplicative pacing for an appropriate value of a pacing multiplier α_i. In this case, the buyer bids optimally in the many auctions, and is able to extract maximum value from their budget by buying the right set of goods. Note that the buyer ends up buying all goods over a certain bang-per-buck threshold (this holds in general for second price markets, if we allow the buyer to get a fraction of a good to reach their budget constraint exactly).

Other budget management systems discussed by Balseiro et al. (2017, see Table 1) include thresholding, reserve pricing, and multiplicative boosting. All these mechanisms work by modifying the participation, bidding or payment rules. For example, thresholding requires the buyer's bids to pass a given threshold to participate, thus forcing buyers to only bid on high-value goods. Reserve pricing is similar, except that the threshold is also used to compute the resulting winning price.

In this section we focus on static models of budget management systems, where the set of goods and values are known ahead of time. One advantage of this perspective is that we can model highly structured valuations across goods. On the other hand, it ignores the stochastic nature associated with impressions that arrive across a day. Several related papers consider goods that arrive stochastically and valuations are then drawn independently. For instance, Balseiro et al. (2015) show that when buyers get to select their bids in each individual auction, a multiplicative pacing equilibrium arises naturally via Lagrangian duality on the budget constraint, under a fluid-based mean-field market model. Balseiro et al. (2017) show the existence of pacing equilibrium for multiplicative pacing as well as the other pacing rules mentioned earlier in a stochastic model with independent valuations. They also give a very interesting comparison of revenue and social welfare properties of the various pacing mechanisms in the unique symmetric equilibrium of their setting. One of the main insights they provide is that multiplicative pacing achieves strong social welfare properties, while probabilistic throttling achieves higher revenue properties.

3.1.1 Second Price Auction Markets

We now explore the case of multiplicative pacing in a static market model with second price auctions. We follow the treatment in Conitzer et al. (2018), and direct the reader there for details, proofs and additional references. In the historical notes at the end of the section, we include additional references to papers that discuss probabilistic throttling.

In the right panel of Fig. 1 we have seen that the focal buyer can optimize its utility by selecting a fixed shaded bid that depends on the total budget. The intuition that a buyer in a repeated auction setting should bid according to a single scalar times their valuations can be shown to hold more generally, even when goods have different values. Specifically, for a given set of bids by all the other buyers, a buyer can always specify a best response by choosing an optimal, constant pacing multiplier. The resulting bid for the buyer on a particular good would be the value of the good in that auction times the fixed pacing multiplier.

Theorem 1 *For arbitrary but fixed bids in each auction for buyers $k \neq i$, buyer i has a best response that consists of multiplicatively-paced bids. This assumes that if a buyer is tied for winning an auction, they can choose the fraction that they want to win. This holds even if the buyer (or proxy bidder) can dispose of some goods that they win, in order to avoid exceeding their budget.*

The previous result takes the perspective of a best response for an individual buyer. The main question we now wish to address is what happens at equilibrium when all buyers play best responses to each others' bids. We refer to such an outcome as a *pacing equilibrium*.

Definition 1 A second price pacing equilibrium (SPPE) is a vector of pacing multipliers $\alpha \in [0, 1]^n$, a fractional allocation x_{ij}, and a price vector p that satisfies the following properties.

(Goods go to highest bidders) If $x_{ij} > 0$, then $\alpha_i v_{ij} = \max_{i' \in N} \alpha_{i'} v_{i'j}$ for each buyer $i \in N$ and good $j \in M$.
(Prices) The unit price of good $j \in M$ is $p_j = \max_{k \neq i} \alpha_k v_{kj}$ for any buyer $i \in N$ such that $x_{ij} > 0$.
(Budget-feasible) $\sum_{j \in M} x_{ij} p_j \leq B_i$ for each buyer $i \in N$.
(No unnecessary pacing) Additionally, if the budget inequality is strict then $\alpha_i = 1$.
(Demanded goods sold completely) $\sum_{i \in N} x_{ij} = 1$ for each good $j \in M$.

The conditions above enforce that winning bids get the goods and buyers pay the second price. The *no unnecessary pacing* condition comes from the practical consideration that buyer should only be paced if their budget constraint is binding. It is basically a complementarity condition that specifies that the mechanism does not want to pace buyers unless it has to. It follows (almost) immediately from Theorem 1 that in an SPPE every buyer is best responding.

Notice that the equilibrium not only includes the pacing multiplier but also the allocations. This is because there may be multiple winning bids for a given good j, and in that case the good may be split among the winning bids, such that each buyer hits their budget constraint exactly. This inclusion of the allocations as part of an SPPE makes it slightly different from a game-theoretic Nash equilibrium. More concretely, we can almost view the problem of finding an SPPE as a pure Nash equilibrium problem in terms of a pacing game that can be defined by the set of pacing multipliers. However, because we must specify the allocation as well, the resulting problem becomes more akin to a market equilibrium (in fact there are strong equivalences between SPPE and market equilibrium, as we shall see later). Nonetheless, it is also possible to formulate a static game with full information such that its pure Nash equilibria and the pacing multipliers α of an SPPE are in one-to-one correspondence. We refer the interested reader to Conitzer et al. (2018) for the details of this.

Importantly SPPE, as defined above, are always guaranteed to exist. This does not follow immediately from previous results such as the existence of Nash equilibria in a standard game. SPPE correspond to a specific type of pure-strategy Nash equilibria and the existence must be explicitly proved.

Theorem 2 *An SPPE of a pacing game is always guaranteed to exist.*

To illustrate this result, we include a quick sketch of the main elements of the proof. First, one constructs a smoothed pacing game, where the allocation is smoothly shared among all bids that are within ϵ of the maximum bid. This makes the allocation a deterministic function of the pacing multipliers α. Several other smooth approximations are also introduced to deal with other discontinuities. In the end, one gets a game in which each player simply has the interval [0, 1] of pacing multipliers as the action space and utilities are well-behaved continuous and quasi-concave functions. For this smoothed game, one can then appeal to a fixed-point theorem to guarantee the existence of a pure-strategy Nash equilibrium in the smoothed game. Finally, the limit point of smoothed games as the smoothing factor ϵ tends to zero yields an equilibrium in the original pacing equilibrium problem.

Unfortunately, while an SPPE is guaranteed to exist, there may be multiple solutions. Moreover, they can have large differences in revenue, social welfare, and other relevant statistics of interest. Figure 2 shows an example of this where the total platform revenue can be orders of magnitude different in two different SPPE. In practice this means that we might need to select the equilibrium that suits ours needs, instead of just solving for one. Although multiplicity of equilibria is a possibility, through simulations one can see that it is not a very common occurrence when looking at instances inspired by real-world ad markets.

Given the practical motivation of the use of market equilibria to understand, manage and forecast ad markets, one may want to actually compute SPPE for a given instance. For instance, the resulting pacing multipliers may be used to shade buyers' bids and drive the system to an operating point in which buyers do not have an incentive to adjust bids further. Although the computational complexity of finding an arbitrary SPPE is open, finding an extremal one (e.g., minimizing/maximizing revenue/social welfare) can be proved to be NP complete. Nevertheless, all the equilibrium conditions can be written as linear constraints with mixed-integer variables, leading to a mixed integer programming (MIP) formulation in which

Problem instance:

i	v_{i1}	v_{i2}	v_{i3}	v_{i4}	B_i
1	100	1	99	100	1
2	1	100	99		1
3				100	100

Equilibrium 1: Revenue = 102

α_i	b_{i1}	b_{i2}	b_{i3}	b_{i4}	spend
1	100	1	99	100	1
0.01	0.01	1	0.99		1
1				100	100

Equilibrium 2: Revenue = 3

α_i	b_{i1}	b_{i2}	b_{i3}	b_{i4}	spend
0.01	1	0.01	0.99	1	1
1	1	100	99		1
1				100	1

Fig. 2 Multiplicity of equilibria. *Left*: a problem instance. *Right*: two possible and very different SPPE

feasibility is equivalent to being at equilibrium. The formulation can be augmented with an objective function of interest to optimize among equilibria.

This formulation can be used to compute equilibria for modestly-sized instances, but as a method it is not very scalable. Instead, we can map SPPE to more general market equilibria to unlock more efficient methods.

To put SPPE in perspective, they can be seen as market equilibria, considering a market setting where each buyer has a quasi-linear demand function $D_i(p) = \text{argmax}_{0 \le x_i \le 1} u_i(x_i, p)$, where u_i was defined in (1). This characterization follows immediately by simply using the allocation x and prices p from the SPPE as a market equilibrium. Theorem 1 tells us that $x_i \in D_i(p)$, and the market clears by definition of SPPE. This implies that SPPE have several useful properties including no envy and Pareto optimality (if one considers the seller as a participant too). This yields the interesting guarantee that, in a budget-adjusted sense, no buyer prefers the allocation of any other buyer, given the prices.

3.1.2 First Price Auction Markets

We now switch to first price auctions in the context of pacing equilibria. Every other aspect of the definition of the market is the same as for SPPE. First price auctions were used initially in internet ad auctions in the 1990s, for example by Yahoo and others. But incentive and stability issues caused a shift to second price auctions. However, first price auctions have seen a recent resurgence of interest for these markets. Notably, several major ad exchanges switched to first price in recent years. For instance, Google Ad Manager switched in September 2019, while Twitter's MoPub exachange switched in August 2020. A major motivation cited by both exchanges is that a first price mechanism will increase transparency and reduce complexity. The incentive and stability issues observed in the 1990s are likely to be less of an issue in today's thicker and much larger-scale markets. We will see that, in the context of repeated auctions, a mechanism that relies on first price repeated auctions has several desirable properties. See also Paes Leme et al. (2020) for an interesting analysis in which firms endogenously choose first price. Our treatment here follows the work of Conitzer et al. (2019); we refer the reader to that article for additional insights, results and proofs.

To build towards market equilibria, we start by defining *budget-feasible pacing multipliers*, which guarantee that buyers stay within budget for goods that are allocated according to first price auction rules.

Definition 2 A set of *budget-feasible first price pacing multipliers* (BFPM) is a vector of pacing multipliers $\alpha \in [0, 1]^n$ and a fractional allocation $x_{ij} \in [0, 1]$ that satisfies the following properties:

(Goods go to highest bidders) If $x_{ij} > 0$, then $\alpha_i v_{ij} = \max_{i' \in N} \alpha_{i'} v_{i'j}$ for each buyer $i \in N$ and good $j \in M$.

(Prices) The unit price of good j is $p_j = \max_{i \in N} \alpha_i v_{ij}$ for each good $j \in M$.

(Budget-feasible) $\sum_{j \in M} x_{ij} p_j \le B_i$ for each buyer $i \in N$.

(Demanded goods sold completely) If $p_j > 0$, then $\sum_{i \in N} x_{ij} = 1$ for each good $j \in M$.
(No overselling) $\sum_{i \in N} x_{ij} \leq 1$ for each good $j \in M$.

To define a pacing equilibrium in the case of first price auctions, we take a BFPM and also impose a complementarity condition between the budget constraint and the pacing multiplier. This guarantees that buyers cannot be paced unless they spend their entire budget.

Definition 3 A *first price pacing equilibrium* (FPPE) is a BFPM (α, x) that also verifies the *no unnecessary pacing* condition, which means that if $\sum_{j \in M} x_{ij} p_j < B_i$, then $\alpha_i = 1$ for each buyer $i \in N$.

The only difference between an FPPE and an SPPE is the pricing condition, which now uses first price.

A very nice property of the first price setting is that BFPMs satisfy a monotonicity condition: if (α', x') and (α'', x'') are both BFPM, then the pacing vector $\alpha = \max(\alpha', \alpha'')$, where the max is taken componentwise, is also a BFPM. The associated allocation is that for each good j, we first identify whether the highest bid comes from α' or α'', and use the corresponding allocation of j (breaking ties towards α').

Intuitively, the reason that (α, x) is also BFPM is that for every buyer i, their bids are the same as in one of the two previous BFPMs (say (α', x') WLOG.), and so the prices they pay are the same as in (α', x'). Furthermore, since every other buyer is bidding at least as much as in (α', x'), they win weakly less of each good (using the tie-breaking scheme described above). Since (α', x') satisfied budgets, (α, x) must also satisfy budgets. The remaining conditions are easily checked.

In addition to componentwise maximality, there is also a *maximal* BFPM (α, x) (there could be multiple x compatible with α) such that $\alpha \geq \alpha'$ for all α' that are part of any BFPM. Consider $\alpha_i^* = \sup\{\alpha_i | \alpha \mathrm{ispartofaBFPM}\}$. For any ϵ and i, we know that there must exist a BFPM such that $\alpha_i > \alpha_i^* - \epsilon$. For a fixed ϵ we can take componentwise maxima to conclude that there exists $(\alpha^\epsilon, x^\epsilon)$ that is a BFPM. This yields a sequence $\{(\alpha^\epsilon, x^\epsilon)\}$ as $\epsilon \to 0$. Since the space of both α and x is compact, the sequence has a limit point (α^*, x^*). By continuity (α^*, x^*) is a BFPM.

We can use this maximality to show existence and uniqueness (of multipliers) of FPPE:

Theorem 3 *An FPPE always exists and the set of pacing multipliers* $\{\alpha\}$ *that are part of an FPPE is a singleton.*

To prove this one can consider the component-wise maximal α and an associated allocation x such that they form a BFPM and show that it has no unnecessarily paced buyers. This follows from supposing that some buyer i is spending strictly less than B_i and $\alpha_i < 1$ and deriving a contradiction to the maximality of the pacing multipliers. Uniqueness follows from the component-wise maximality and the *no unnecessary pacing* condition.

3.1.3 Sensitivity Analysis

FPPE enjoy several nice monotonicity and sensitivity properties that SPPE do not. Several of these follow from the maximality property that we have seen earlier: the unique FPPE multipliers α are such that $\alpha \geq \alpha'$ for any other BFPM (α', x'). The following actions are all guaranteed to weakly increase the revenue at equilibrium.

Adding a buyer $n+1$. The original solution (α, x) together with $\alpha_{n+1} = 0, x_{n+1} = 0$ is a BFPM of the expanded instance. By the monotonicity property, prices must weakly increase.

Adding a good. The FPPE of the expanded instance α' satisfies $\alpha' \leq \alpha$. (To see this, consider the set of buyers whose multipliers increased, since they win more goods and prices are up, some buyer must strictly exceed their budget, a contradiction). The set of buyers $i \in N$ such that $\alpha'_i < \alpha_i$ must be spending their whole budget by the *no unnecessary pacing* condition. For buyers such that $\alpha'_i = \alpha_i$, they pay the same as before, and win weakly more goods.

Increasing a buyer i's budget. The original solution (α, x) is a BFPM in the updated instance. By the maximality of the FPPE solution, its multipliers must be larger.

It is also possible to show that revenue enjoys a Lipschitz property: increasing a single buyer's budget by Δ increases revenue by at most Δ. Similarly, social welfare can be bounded in terms of Δ, though multiplicatively, and it does not satisfy monotonicity.

3.1.4 Convex Program to Compute FPPE

As discussed earlier, besides theory, the motivation of formulating pacing systems as markets is to provide algorithms to compute them. Computing an FPPE turns out to be easier than an SPPE since we do not need to rely on an integer program. This is due to a direct relationship between pacing and market equilibria. FPPE are given exactly by the set of solutions to the *quasi-linear* variant of the Eisenberg-Gale convex program for computing a market equilibrium:

$$\max_{x \geq 0, \delta \geq 0, u} \quad \sum_i B_i \log(u_i) - \delta_i \qquad\qquad \min_{p \geq 0, \beta \geq 0} \quad \sum_j p_j - \sum_i B_i \log(\beta_i)$$

$$s.t. \quad u_i \leq \sum_j x_{ij} v_{ij} + \delta_i, \quad i \in N \qquad\qquad s.t. \quad p_j \geq v_{ij} \beta_i, \quad i \in N, j \in M \tag{2}$$

$$\sum_i x_{ij} \leq 1, \quad j \in M \qquad\qquad \beta_i \leq 1, \quad i \in N \tag{3}$$

We show the primal convex program on the left and its corresponding dual convex program on the right. The variables x_{ij} denote the fractional amount of good j that buyer i wins. The leftover budget is captured by δ_i, which is the primal variable corresponding to the dual constraint $\beta_i \leq 1$.

The dual variables β_i and p_j correspond to constraints (2) and (3), respectively. The variable p_j is the price of good j and $\beta_i = \min_{j:x_{ij}>0}\{p_j/v_{ij}\}$ can be interpreted as the inverse bang-per-buck for buyer i. With this definition of β_i, the constraint $\beta_i \leq 1$ is intuitively clear: a quasi-linear buyer only wishes to spend money if their price-per-utility is at most 1.

One can show via Karush–Kuhn–Tucker conditions that FPPE and EG are equivalent. Informally, the correspondence between them follows because β_i specifies a single price-per-utility rate per buyer which exactly yields the pacing multiplier $\alpha_i = \beta_i$. Complementary slackness then guarantees that if $p_j > v_{ij}\beta_i$ then $x_{ij} = 0$, so any good allocated to i has rate β_i exactly. Similarly, complementary slackness on $\beta_i \leq 1$ and the associated primal variable δ_i guarantee that buyer i is only paced if they spend their whole budget.

Theorem 4 *An optimal solution to the quasi-linear Eisenberg-Gale convex program corresponds to an FPPE with pacing multiplier $\alpha_i = \beta_i$ and allocation x_{ij}, and vice versa.*

It follows that an FPPE can be computed in polynomial time, and that we can apply various first-order methods to compute large-scale FPPE. Such first-order methods will be discussed in Sect. "First-Order Methods".

3.1.5 Comparison Between SPPE and FPPE

The SPPE and FPPE properties have interesting differences, which we summarize in Table 1. For additional details, see the literature in the historical notes at the end of the section. FPPE are unique (this can be shown from the convex program, or directly from the monotonicity property of BFPM) while SPPE are not. In practice SPPE instances admitting multiple equilibria seem rare. FPPE can be computed in polynomial time. While the complexity of SPPE is unknown, it is NP-hard to

Table 1 A comparison of FPPE and SPPE

	SPPE	FPPE
Exists?	Yes	Yes
Is unique?	No	Yes, up to buyer utilities
Is efficiently computable?	NP-hard	Convex program
Is welfare monotone?	No	Yes, in goods
Is revenue monotone?	No	Yes, in goods/buyers/budgets
Is shill proof?	No	Yes
Pacing eq. is best response?	Yes	No
Simulated revenue	SPPE $\leq$ FPPE	
Simulated welfare	Ambiguous	

maximize revenue or social welfare. FPPE are robust to perturbations (e.g., revenue increases smoothly as budgets are increased). Both equilibrium concepts correspond to (different) market equilibria but SPPE requires buyer demands to be "supply aware." SPPE correspond to a pure-strategy Nash equilibria, and thus buyers are best responding to each other. Neither FPPE nor SPPE are strategyproof, but the market equilibrium connection can be used to show strategyproofness in an appropriate "large market" sense.

As we will discuss in Sect. "Numerical Experiments", FPPE and SPPE have also been studied experimentally, using instances generated from real ad auction data. To complete the comparison, we report the most interesting takeaways from those experiments here:

- Manipulation is hard in both SPPE and FPPE if you can only lie about your value-per-click.
- FPPE dominates SPPE on revenue.
- Social welfare can be higher in either FPPE or SPPE. Experimentally, there was not a clear winner of which of the two solution concepts provides a higher social welfare.

3.2 Dynamic Budget Management Systems

The previous section explored a market with repeated auctions viewed as a static game between advertisers that set pacing multipliers. Since that view ignores time, this section presents a dynamic view, where a buyer or a proxy bidder has to sequentially tune its pacing multiplier to manage their bids over time. The goal is to hit the 'right' pacing multiplier as before and match the spend and the budget, but each buyer has to learn that multiplier as the market plays out. We will see how to approach this problem using ideas from regret minimization. The exposition closely follows the work of Balseiro and Gur (2019).

3.2.1 Dynamic Auctions Markets

In this section we consider second-price mechanisms with n buyers who participate in auctions sequentially at times $t = 1, \dots, T$. At time t an auction takes place and each buyer samples a valuation v_{it} independently from a cumulative distribution (CDF) function F_i supported in $[0, \bar{v}_i]$ which is assumed to be absolutely continuous and with bounded density f_i. We use the vector notation v_i to denote the sequence of realized valuations across all auctions. As earlier, we assume that each buyer has some budget B_i to be spent across all auctions. We define by $\rho_i = B_i/T$ the per-period target expenditure, which we assume to be bounded by $\bar{v}_i$. Each buyer is characterized by a type $\theta_i = (F_i, \rho_i)$.

After realizing the valuation v_{it}, buyer i submits a bid b_{it}. We let $d_{it} = \max_{k \neq i} b_{kt}$ denote the highest bid other than that of i, and we use the vector notation d_i to refer to the sequence across all auctions. The buyers' utilities continue to be quasi-linear: they receive a utility of $u_{it} = \mathbb{1}\{d_{it} \leq b_{it}\}(v_{it} - d_{it})$, where the first term is an indicator function that equals one if buyer i wins auction t, and pay $z_{it} = \mathbb{1}\{d_{it} \leq b_{it}\}d_{it}$.[1]

We assume that each buyer has no information on the valuation distributions, including their own. Instead, they just know their own target spend rate ρ_i (i.e., spend per time period) and the total number of time periods T. Buyers also do not know how many other buyers are in the market. At time t, buyer i knows the *history* $(v_{i\tau}, b_{i\tau}, z_{i\tau}, u_{i\tau})_{\tau=1}^{t-1}$ of their own values, bids, payments, and utilities. Furthermore, they know their current value v_{it}. Based on this history, they choose a bid b_{it}. We will say that a bidding strategy for buyer i is a sequence of mappings $\beta = \beta_1, \ldots$ where β_t maps the current history to a bid (potentially in randomized fashion). The strategy β is budget feasible if the bids b_{it}^{β} generated by β are such that $\sum_{t=1}^{T} \mathbb{1}\{d_{it} \leq b_{it}^{\beta}\}d_{it} \leq B_i$ under any vector of highest competitor bids d_i. For given d_i and valuation vectors v_i, we denote the expected value of a strategy β as

$$\pi_i^{\beta}(v_i, d_i) = \mathbb{E}_{\beta}\left[\sum_{t=1}^{T} \mathbb{1}\{d_{it} \leq b_{it}^{\beta}\}(v_{it} - d_{it})\right],$$

where the expectation is taken with respect to randomness in β.

We would like to compare this outcome to the *hindsight optimal* strategy. We denote the expected value of that strategy as

$$\begin{aligned} \pi_i^{H}(v_i, d_i) \coloneqq \max_{x_i \in \{0,1\}^T} & \sum_{t=1}^{T} x_{it}(v_{it} - d_{it}) \\ s.t. & \sum_{t=1}^{T} x_{it} d_{it} \leq B_i \,. \end{aligned} \tag{4}$$

The hindsight optimal strategy has a simple structure: a buyer simply chooses the optimal subset of goods to win while satisfying the budget constraint. In the case where the budget constraint is binding, this is a knapsack problem.

Ideally we would like to consider a strategy π_i^{β} that approaches π_i^{H}. However, this turns out not to be possible. We will use the idea of asymptotic γ-competitiveness to see this. Formally, β is asymptotically γ-competitive if

[1] In this case with continuous distributions, the probability of ties is zero.

$$\limsup_{\substack{T\to\infty,\\ B_i=\rho_i T}} \sup_{\substack{v_i\in[0,\bar{v}_i]^T,\\ d_i\in\mathbb{R}_+^T}} \frac{1}{T}\left(\pi_i^H(v_i,d_i)-\gamma\pi_i^\beta(v_i,d_i)\right)\le 0\,.$$

Intuitively, the condition says that asymptotically, β should achieve at least $1/\gamma$ of the hindsight-optimal expected value.

For any $\gamma < \bar{v}_i/\rho_i$, asymptotic γ-competitiveness turns out to be impossible to achieve. Thus, if our target expenditure ρ_i is much smaller than our maximum possible valuation, we cannot expect to perform anywhere near as well as the hindsight optimal strategy. The general proof of this fact is quite involved, but the high-level idea is not too complicated. We show the construction for $\bar{v}_i = 1$, $\rho_i = 1/2$, and thus the claim is that $\gamma < \bar{v}_i/\rho_i = 2$ is unachievable. The impossibility is via a worst-case instance. In this instance, the highest other bid comes from one of the two following sequences:

$$d^1 = \left(d_{high},\ldots,d_{high},\bar{v}_i,\ \ldots\ldots\ ,\bar{v}_i\right)$$
$$d^2 = \left(d_{high},\ldots,d_{high},d_{low},\ldots,d_{low}\right),$$

for $\bar{v}_i \ge d_{high} > d_{low} > 0$. The general idea behind this construction is that in the sequence d^1, buyer i must buy many of the expensive goods to maximize their utility, since they receive zero utility for winning goods with price $\bar{v}_i$. However, in the sequence d^2, buyer i must save money so that they can buy the cheaper goods priced at d_{low}.

For the case we consider here, there are $T/2$ of each type of highest other bid (assuming that T is even for convenience). Now, we may set $d_{high} = 2\rho_i - \epsilon$ and $d_{low} = 2\rho_i - k\epsilon$, where ϵ and k are constants that can be tuned. For sufficiently small ϵ, buyer i can only afford to buy a total of $T/2$ goods, no matter the combination they get. Furthermore, buying a good at price d_{low} yields k times as much utility as buying a good at d_{high}.

To achieve at least half of the optimal utility under d^1, buyer i must purchase at least $T/4$ of the goods priced at d_{high}. Since they do not know whether d^1 or d^2 occurred until after deciding whether to buy at least $T/4$ of the d_{high} goods, this must also occur under d^2. But then buyer i can at most afford to buy $T/4$ of the goods priced at d_{low} when they find themselves in the d^2 case. Finally, for any $\gamma < 2$, we can pick k and ϵ such that achieving $\gamma\pi_i^H$ requires buying at least $T/4+1$ of the d_{low} goods.

It follows that we cannot hope to design an online algorithm that competes with $\gamma\pi_i^H$ for $\gamma < \bar{v}_i/\rho_i$. However, it turns out that a subgradient descent algorithm can achieve exactly $\gamma = \bar{v}_i/\rho_i$.

3.2.2 An Adaptive Pacing Strategy

In this section, we present a pacing strategy that optimizes the pacing multipliers by adjusting them over time. We consider a focal buyer $i \in N$ for whom we set $\alpha_i = \frac{1}{1+\mu}$ and iteratively tune it by running a subgradient descent scheme on the value for μ, which will allow the buyer to smoothly spend the budget across the T time periods.

The algorithm takes as input a step size $\epsilon_i > 0$ and some initial value $\mu_1 \in [0, \bar{\mu}_i]$ where $\bar{\mu}_i$ is an upper bound on μ. We use $P_{[0,\bar{\mu}_i]}$ to denote projection onto the interval $[0, \bar{\mu}_i]$. The algorithm APS, proposed by Balseiro and Gur (2019) and motivated by Lagrangian duality, proceeds as follows:

Algorithm 1: APS (Balseiro and Gur, 2019)

```
Initialize the pacing parameter μ_1 and the remaining budget B̃_i1 = B_i.
for every time period t = 1, ..., T do
    Observe v_it, construct a paced bid b_it = min(v_it/(1+μ_t), B̃_it).
    Observe spend z_it, and refine the pacing multiplier using the update rule
    μ_{t+1} = P_[0,μ̄_i](μ_t − ε_i(ρ_i − z_it)).
    Update remaining budget B̃_{i,t+1} = B̃_it − z_it.
```

The problem $\max_{x \in \{0,1\}^T} \sum_{t=1}^{T} [x_{it}(v_{it} - (1-\mu)d_{it}) + \mu\rho_i]$ is the Lagrangian relaxation of the hindsight optimal optimization problem (4). The optimal solution for the relaxed problem is easy to characterize: we set $x_{it} = 1$ for all t such that $v_{it} \geq (1-\mu)d_{it}$. Importantly, this is achieved by the bid $b_{it} = \frac{v_{it}}{1+\mu}$ that we use in APS.

The Lagrangian dual is the minimization problem

$$\inf_{\mu \geq 0} \sum_{t=1}^{T} \left[(v_{it} - (1-\mu)d_{it})^+ + \mu\rho_i \right], \tag{5}$$

where $(\cdot)^+$ denotes thresholding at 0. This dual problem upper bounds π_i^H (but we do not necessarily have strong duality since we did not even start out with a convex primal program). The minimizer of the dual problem yields the strongest possible upper bound on ϕ_i^H. However, solving this requires us to know the entire sequences of v_i and d_i. APS approximates this optimal μ by taking a subgradient step on the t'th term of the dual:

$$\partial_\mu \left[(v_{it} - (1-\mu)d_{it})^+ + \mu\rho_i \right] \ni \rho_i - d_{it}\mathbb{1}\{b_{it} \geq d_{it}\} = \rho_i - z_{it}.$$

Thus APS is taking subgradient steps based on the subdifferential of the t'th term of the Lagrangian dual of the hindsight optimal optimization problem.

The APS algorithm achieves exactly the lower bound we derived earlier, and is thus asymptotically optimal.

Theorem 5 APS *with step size* $\epsilon_i = O(T^{-1/2})$ *is asymptotically* $\frac{v_{it}}{\rho_i}$*-competitive, and converges at a rate of* $O(T^{-1/2})$.

This result holds under adversarial conditions: for example, the sequence of highest other bids may be as d^1, d^2 in the lower bound. However, in practice we do not necessarily expect the world to be quite this adversarial. In a large-scale auction market, we would typically expect the sequences v_i, d_i to be more stochastic in nature. In a fully stochastic setting with independence, APS turns out to achieve π_i^H asymptotically:

Theorem 6 *Suppose* $(v_{it,d_{it}})$ *are sampled independently from stationary, absolutely continuous CDFs with differentiable and bounded densities. Then the expected payoff from* APS *with step size* $\epsilon_i = O(T^{-1/2})$ *approaches* π_i^H *asymptotically at a rate of* $T^{-1/2}$.

Theorem 6 shows that if the environment is well-behaved then we can expect much better performance from APS. It can also be shown that when all buyers use APS with appropriate step sizes, then each buyer converges to a solution that achieves the optimal dual value (5) (note that since we do not have strong duality this does not imply that π_i^H is achieved).

3.3 Numerical Experiments

In previous sections, we have illustrated how pairing auction markets and market equilibria can allow us to derive theoretical properties and can give us a tool to effectively compute equilibria in auction markets. Recalling the focus on equilibrium computation, in this section we present an empirical study, drawing from the material in Conitzer et al. (2018, 2019). Relying on SPPE and FPPE computed for a set of stylized and realistic instances, we discuss how these equilibria compare in terms of revenue and welfare, provide evidence that incentives in FPPE arising from the first price auction are not problematic, and show how the static FPPE can be used to effectively seed the dynamic pacing algorithm.

The experiments are mainly based on realistic instances derived from the real-world auction markets at Facebook and Instagram. The instances were constructed in two steps, as explained in Conitzer et al. (2019). The first step is to take bidding data for a region during a period and use it to create n buyers and m goods. The buyers are the top n advertisers that participate in the most auctions in that period in that region. The set of goods is constructed by applying a k-means algorithm to the auctions in which the advertisers participated. The features used for this are the n-dimensional vector of bids of each advertiser in each auction. The valuation of a buyer to a good is set to the average valuation of auctions in the cluster.

The budgets are set equal to the expected value that the buyer would receive in a uniform random allocation of goods to buyers, i.e., $B_i = \frac{1}{n}\sum_j v_{ij}$. The motivation for this is that it leads to a good mixture of budget-constrained and unconstrained buyers, since in aggregate this constrains the sum of prices to be the sum of average valuations, whereas it would be the sum of maximum valuations if every buyer were unconstrained. The set of constructed instances combines different days, platforms, number of buyers ($n \in \{6, 8, 10, 12, 14\}$), and number of goods ($m \in \{10, 20, 30\}$) for a total of 210 instances for FPPE. Instances for SPPE require to be slightly smaller to be able to solve the MIP. The numerical study includes a set of instances with $\{3, \ldots, 8\}$ buyers and $\{4, \ldots, 8\}$ goods, for a total of $2 \times 7 \times 6 \times 5 = 420$ instances.

3.3.1 Computational Comparison Between SPPE and FPPE

This section compares revenue and social welfare under FPPE and SPPE, as shown in Fig. 3. The left panel shows the CDF of the ratio of FPPE revenue to that of SPPE. The right panel is similar but with social welfare. We see that FPPE revenue is always higher than SPPE revenue, though both coincide for about 30% of instances, and almost never more than 4.5 times as high. For social welfare, perhaps surprisingly, neither solution concept is dominating, with most instances having relatively similar social welfare under either solution concept, though FPPE does slightly better. There are two caveats to keep in mind for these results: (a) the numerical study did not compute the social-welfare-maximizing SPPE so it is possible that there is a better equilibrium (although this is highly unlikely given that most instances admit a single equilibrium); (b) many buyers are budget constrained in the FPPE of our setting, and so these insights might not translate to cases where many buyers are not budget constrained. These experiments show that an FPPE is not necessarily worse than an SPPE with respect to social welfare (at least with nonstrategic buyers), while potentially having a significantly higher revenue.

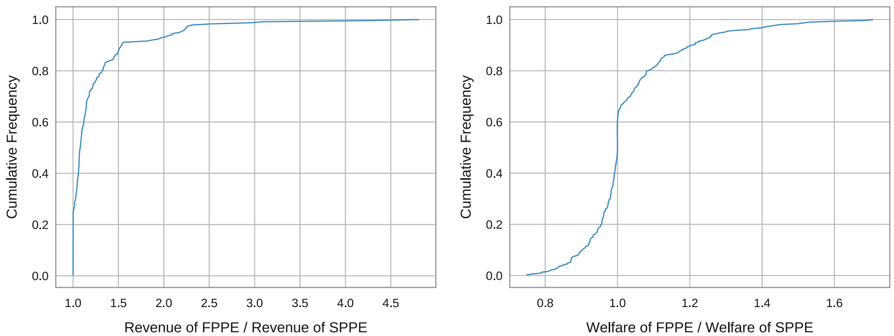

Fig. 3 CDF of the FPPE / SPPE ratio of revenue (left) and social welfare (right)

3.3.2 Incentives in FPPE

This section summarizes why incentives for advertisers are less of a problem than expected when using first price auctions. The incentive to deviate is quantified through the ex-post regret of buyers at FPPE, which capture what can be achieved when they unilaterally deviate to a different pacing multiplier while keeping the FPPE multipliers fixed for all other buyers. Figure 4 displays those ex-post relative regrets as the fraction of utility that is lost if the buyer uses the best-response pacing multiplier instead of reporting truthfully. The median regrets are very close to zero for instances of all sizes and the third quartile is below 0.02. The conclusion is that ex-post incentives to shade bids for individual advertisers when they can report a lower value per conversion or budget is very small in almost all cases. Furthermore, the incentive for misreporting the value per conversion or budget as inputs to the mechanism is vanishingly small. In the unrealistic case when advertisers have the power to shade bids at the auction level, the level of ex-post regret depends on the market thickness. Even in this extreme case, the average relative regret is never above 0.2.

As a hypothesis, this conclusion has to do with the coarseness at which manipulations can be performed when buyers do not have the ability to shade bids in individual auctions. Even if there is a large gap between the first and second price in a given auction, the winning buyer may not be able to exploit this, because once they start lowering their value per conversion, they might start losing some other auction much closer to their first price bid. Thus, a buyer need not face a "thick" market in every auction as one would conclude with first price auctions. It is enough for the incentive to deviate to be small if just a fraction of the auctions targeted by each advertiser is competitive.

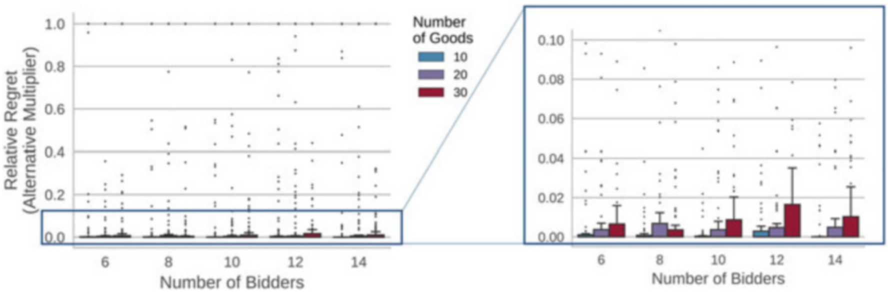

Fig. 4 Summary statistics of relative ex-post regret at an FPPE (ratio of best-response utility keeping competitors' bids fixed to utility under the FPPE). There is a data point for each buyer at the FPPE of each instance. The lower and upper edges of the boxes represent the first and third quartiles; the lines extending from the box show outliers within 1.5 times the inter-quartile range; and the dots represent individual outliers outside that range. The plot on the right is a zoomed-in version of the plot on the left

3.3.3 Seeding Dynamics with SPPE

As we described in Sect. “Dynamic Budget Management Systems”, real-world pacing heuristics rely on tractable adaptive algorithms that update buyers’ pacing multipliers over time. This section looks at the rate at which the APS algorithm converges since the longer it takes to converge, the worse it is at optimizing the buyer’s utility. In the evaluation, the algorithm is seeded with the solutions to the static SPPE and the resulting regret compared to other starting solutions such as constant or random pacing multipliers. The seed is computed from a static instance that abstracts away the dynamics but captures the market structure. The MIP mentioned earlier is used to solve the problem.

For each set of initial pacing multipliers, the runs are done with parameters ϵ and $\alpha^{\mathtt{min}}$, determined through grid search by choosing those that minimize the *average ex-post relative regret* (i.e., the average amount that a buyer could have improved its utility by playing a single best-response multiplier, given the other bids are fixed). As shown on the left of Fig. 5, running APS on stylized random instances with MIP-based initial multipliers produces a lower regret than with other choices of initial multipliers. The performance of the MIP-based solution degrades as the noise parameter σ on the input data grows, but even at the highest levels we considered, this solution outperforms the others. For the fixed initial multipliers, the resulting regret is highly sensitive to choices in the step size: low initial multipliers would often not reach the MIP’s equilibrium multipliers by the time the algorithm terminated. For realistic instances, the right plot of Fig. 5 shows that the regret experienced by buyers when starting from the MIP-based initial multipliers was lower than in the other cases, for every learning rate ϵ that was considered. The worst learning rate for the MIP was better than the best learning rate for any other set of starting points. These findings are robust to different number of clusters m when producing the realistic instances.

In conclusion, using an SPPE of a static representation of an instance to warm-start an adaptive algorithm on the dynamic instance resulted in better convergence,

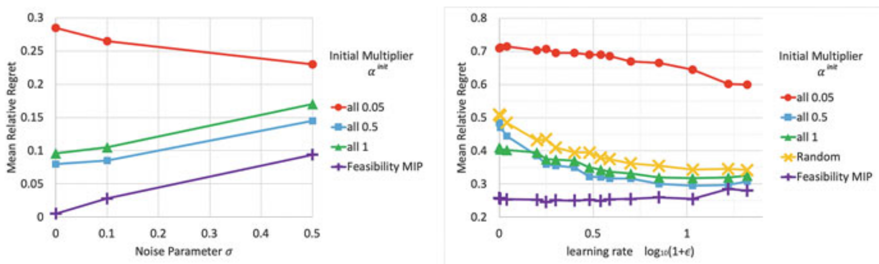

Fig. 5 Mean relative regret from running APS. Each curve plots different initial pacing multipliers α_i^{init}. *Left*: Stylized instances with random perturbations. Regret as a function of the noise parameter σ. *Right*: Realistic instances with 8 clusters (no noise). Regret as a function of the learning rate ϵ (shown in log scale as $\log 10(1+\epsilon)$)

and these improvements were robust to noise in the input data. This robustness provides evidence that the MIP does not need the exact valuation distribution or exact market structure to be useful. In Sect. "Market Abstractions", we discuss how to compress a large instance to create a smaller, approximate representative instance that could be tractably solved by the MIP.

3.4 Historical Notes

Borgs et al. (2007) study a dynamic bid optimization scheme based on first and second price auctions with perturbed allocation rules. While they do not discuss pacing as an equilibrium, their perturbation scheme in the first-price case leads to FPPE. Balseiro et al. (2015) started the study of pacing equilibria and showed that when buyers get to select their bids individually, multiplicative pacing equilibrium arises naturally via Lagrangian duality on the budget constraint, under a fluid-based mean-field market model. The literature has generally studied models where goods arrive stochastically and valuations are then drawn independently. Balseiro et al. (2017) show existence of pacing equilibrium for multiplicative pacing as well as several other pacing rules for such a setting; they also give a very interesting comparison of revenue and social welfare properties of the various pacing options in the unique symmetric equilibrium of their setting. Most notably, multiplicative pacing achieves strong social welfare properties, while probabilistic throttling achieves higher revenue properties. The static multiplicative pacing equilibrium results that we presented in this chapter were developed by Conitzer et al. (2018) for second price auction markets, and by Conitzer et al. (2019) for first price ones. The fixed-point theorem that is invoked to guarantee existence of a pure-strategy Nash equilibrium in the smoothed game is by Debreu (1952), Glicksberg (1952), and Fan (1952).

The quasi-linear variant of Eisenberg-Gale was given by Chen et al. (2007) and later rediscovered by Cole et al. (2017). For discussion on strong duality and optimality conditions for these problems, see Bertsekas et al. (2003, Proposition 6.4.4). The KKT conditions can be significantly generalized beyond convex programming.

The dynamic model of budget management was developed by Balseiro and Gur (2019). Beyond auction markets, the idea of using paced bids based on the Lagrange multiplier μ has been studied in the revenue management literature, see e.g., Talluri and Van Ryzin (1998), where it is shown that this scheme is asymptotically optimal as T tends to infinity. There is also recent work on the adaptive bidding problem using multi-armed bandits (Flajolet and Jaillet, 2017).

The numerical study that we presented was performed by Conitzer et al. (2018) for second price auction markets and the comparison of dynamic pacing under the various starting points, and by Conitzer et al. (2019) for first price auction markets and the comparison between them.

4 Fair Division Problems and Applications at Scale

Market equilibrium is also intimately related to the problem of fairly dividing goods among agents. In *fair division problems* the setup is completely analogous to the Fisher market setting: we have m divisible goods to allocate to n individuals. The preferences of individuals are captured by utility functions $u_i(x_i)$. The goal is to find a "good" assignment x of goods to buyers. However, what is considered "good" turns out to be complicated in the setting of fair division, as there are many possible desiderata we may wish to satisfy.

First, we would like the allocation to be efficient, meaning that it should lead to high utilities for the individuals. One option would be to try to maximize the *social welfare* $\sum_i u_i(x_i)$. However, this turns out to be incompatible with the fairness notions that we will introduce later. An easy criticism of social welfare in the context of fair division is that it favors *utility monsters*: individuals with much greater capacity for utility are given more goods. Since social welfare maximization is typically incompatible with fairness, fair division mechanisms typically opt for the less stringent notion of *Pareto optimality* of the resulting allocation x. That requires that for *every* other allocation x', if one individual i is better off under x' than under x, then some other individual i' is strictly worse off. In other words, x should be such that no other allocation weakly improves all individuals' utilities, unless all utilities stay the same.

In addition to Pareto optimality, fair division mechanisms typically strive for allocations that satisfy various fairness conditions. We will be concerned with the following two desiderata:

Envy free: An allocation x has no envy if $u_i(x_i) \geq u_i(x_{i'})$ for every pair of individuals i and i'. In other words, every individual prefers their own bundle at least as much as that of anyone else.

Proportionality: An allocation x satisfies proportionality if $u_i(x_i) \geq u_i\left(\frac{1}{n} \cdot s\right)$. In other words, every individual prefers their own bundle at least as much as receiving a bundle composed of an nth of every good.

An allocation that satisfies Pareto optimality, no envy, and proportionality turns out to be achievable using the so-called *competitive equilibrium from equal incomes* (CEEI), a classic economic solution concept based on market equilibrium. In CEEI, a fair allocation is obtained as follows. First, we give each individual a unit budget of *funny money* that represents a fake currency to operate the content recommendation system. Second, we compute a market equilibrium for the Fisher market consisting of the individuals and their utility functions, along with the unit budgets of funny money. Finally, we take the corresponding market equilibrium allocation x, call it our fair division, and forget about the funny money.

CEEI is an appealing solution with respect to the previous desiderata. It is Pareto optimal since every market equilibrium satisfies it as discussed in Sect. "Introduction to Fisher Markets and Market Equilibrium". It has no envy since each buyer has the same budget. That means that each buyer can afford the bundle of any other buyer, and every buyer buys an optimal allocation given the prices and budgets. Finally, proportionality is satisfied since each buyer can afford the bundle where they get s_j/n of each good. This is easily shown by noting that the sum of prices must equal the sum of budgets.

In the divisible setting, CEEI is guaranteed to exist, and it is computable both using convex programming (via the Eisenberg-Gale convex program), and at scale via first-order methods as we shall see in Sect. "First-Order Methods". In contrast, for the indivisible setting, CEEI is not necessarily guaranteed to exist. In that case, one could rely on approximate-CEEI, a relaxed solution concept where buyers get slightly unequal budgets (Budish, 2011).

Let us now discuss the main applications of CEEI. The most obvious fair division settings that come to mind as practical examples involve indivisible objects, e.g., housing assignment, school choice, fair estate division, and so on. There have been several interesting applications of market-equilibrium ideas to the indivisible case. In spite of the possible non-existence of market equilibria, Budish et al. (2016) apply an approximate market equilibrium notion to the problem of fairly assigning course seats to MBA students. While market equilibria are not guaranteed to exist in that setting, Budish (2011) shows that approximate equilibria exist, and that they have appealing fairness and incentive properties. This approach is currently applied at several business schools. Another interesting application is that of fairly dividing goods such as items in an estate. A publicly available implementation of this can be found at www.spliddit.org. That webpage allows users to set up fair division instances of moderate size, and a fair allocation is offered by computing the discrete allocation that maximizes the geometric mean of utilities. This is a direct extension of EG to the discrete setting. A market equilibrium approximation is not guaranteed, but approximations to discrete variants of envy freeness and proportionality are guaranteed, as shown by Caragiannis et al. (2019).

The indivisible approaches mentioned above are, however, far from scalable to the size of allocation problems faced in auction markets and content ranking applications. Indivisible fair division problems can be converted into divisible ones by allowing randomized allocation. But this may not lead to particularly fair or acceptable solutions (for example, flipping a coin to decide who inherits a house would not be acceptable to many people). Although sometimes this randomization can be resolved successfully (e.g., Budish et al., 2013), generally speaking the randomization can lead to large ex-post regret. However, internet firms can usually circumvent this issue because of the scale of their markets. For instance, if they allocate content to users attempting to provide an balanced distribution from creators, then a randomized allocation may be enough. Content gets shown a large number of times, which smooths out randomization issues that can crop up in settings where each individual is assigned only a few goods. In the following sections we will describe applications of the divisible CEEI solution concept to such internet-scale problems.

4.1 Fair Recommender Systems and Diversity in Ranking

As a concrete application 'at scale,' in this section we go deeper to describe how Fisher markets and their equilibria can be used to model large-scale content recommendation systems. The main aspect of these recommendations is that we want to explicitly consider fairness and diversity goals. The problem can be summarized as follows: we have a set of n pieces of *content* (e.g., songs we could recommend to users, or job posts that we could show to users). We have m opportunities to show content (e.g., each job seeker is shown a ranked list of five job posts, which would generate five opportunities). The goal is to allocate the content to the different opportunities in a way that maximizes a relevant efficiency metric (e.g., likes of recommended songs, or actual job applications). At the same time, we also wish to treat each piece of content fairly. For example, in the jobs setting, we may wish to avoid showing a small set of job posts over and over, even if most job seekers were likely to apply to them. Not only this is not conducive to improving the ecosystem in this two sided market but also it is unlikely that all applications can be accepted by the company offering the job.

For the content recommendation problem, the analogy to market equilibrium is that the content creators are the buyers, and the content recommendation slots are the goods (e.g., if there are five job posts shown per page then there would be five goods to 'sell' in the market). By reducing the content recommendation problem to a market equilibrium problem, we can guarantee Pareto-efficient content recommendation, where every content creator will have budget-adjusted envy-freeness, and receive at least their budget-proportional utility. The budget of 'funny money' given to each content creator can be chosen according to what fraction of the recommendations we want each content creator to receive. As an example, consider the music recommendation setting and two musicians where one is up-and-coming while the other is a superstar. While the goal is to fairly give some exposure to the up-and-coming musician, we likely would not want to give equal budgets to the two musicians. Instead, the superstar might get a larger budget of 'funny money', while the up-and-coming musician might get a smaller budget that is large enough to ensure exposure.

Similarly to the discussion about proxy bidders in Sect. "Auction Markets and Budget Management Systems", this vision of a recommender system may be implemented by the platform in its entirety. A proxy bidder submits the value of each opportunity acting on behalf of the buyer. In this setup the funny money and budgets are merely abstractions used in the market, and are entirely controlled by the platform. There is no actual money changing hands in this mechanism. The whole point is to make the recommendations fair, and not that anybody can buy their way into more appealing slots.

More concretely, suppose that we have n content creators (say job posts), and m possible recommendation slots to fill (e.g., every time a job seeker shows up we show them a single job post). Furthermore, suppose that we measure the quality of recommending job post j to job seeker i by the probability that the job seeker clicks on (or applies to) the job post. This probability v_{ij} will represent the value in the market model. We wish for a given job post i to get allocated roughly to some percentage of all content slots. We can set the budget B_i for job post i equal to that percentage. Having defined buyers, goods, values and budgets, now we have a Fisher market model that admits an equilibrium. Allocating the job posts according to that equilibrium guarantees several desirable properties. For any job post i, proportionality guarantees that the expected number of job seekers that will click on the job post is at least as high as if that post were shown to each job seeker with probability B_i. Secondly, the no-envy property guarantees that every job post has at least as high of an expected number of clicks as if they receive the allocation for any other job post k, after adjusting the allocation to k by the factor $\frac{B_i}{B_k}$.

An interesting extension of the CEEI application to fair allocation of content would be to more directly guarantee fairness on both sides of the market. In particular, CEEI gives a one-sided fairness guarantee: it only requires that content creators receive a fair share of the set of recommendations. Ideally, we would also like users being shown recommendations to also get explicit fairness guarantees. Mehrotra et al. (2018) describe this problem in the context of *music recommendations* motivated by Spotify. A naive content recommendation approach based on relevance prediction will typically allocate the majority of recommendations to a small set of superstar musicians. Instead, the platform would like to recommend music in a way that more equitably recommends songs by less famous artists. At the same time, there is a clear tradeoff in that user satisfaction is extremely important. Mehrotra et al. (2018) study a group fairness notion where musicians are each assigned one out of K "popularity bins," and they measure fairness as $\sum_{k=1}^{K} \sqrt{|A_k|}$ where $|A_k|$ is the number of artists from bin k assigned a recommendation. This yields a form of regularization towards fair treatment of each artist bin, though it does not yield the kind of per-artist fairness that market equilibrium guarantees.

Let us briefly mention that a market-equilibrium-based allocation is not the only possible approach towards achieving more diverse recommendations. A very related approach is to use a linear programming (LP) approach that maximizes total social welfare, but ensures that each content creator is allocated at least some minimum amount of utility. Such an approach can be adapted to an online setting by adding a Lagrange multiplier on each utility lower bound. This multiplier can be tuned over time using a control algorithm, based on whether the creator is receiving the right amount of utility. In a certain sense, these two approaches can be viewed as equivalent: by the second welfare theorem we know that any Pareto-efficient allocation can be implemented in market equilibrium by an appropriate redistribution of budgets. However, in practice the two approaches are calibrated very differently. The LP approach requires us to specify an exact utility lower bound for each content creator. On the one hand this gives us very concrete utility guarantees, but on the other hand it may be hard to pick appropriate utility lower bounds, especially if we do not know the exact market composition ahead of time. Some choices of utility lower bounds may even lead to infeasibility. By using a market-equilibrium approach one can instead specify the budgets of funny money for each content creator. This is always guaranteed to be feasible, and is akin to instead specifying a fractional share of the market that we would like to allocate to each content creator.

4.2 More Connections to Large-Scale Internet Applications

Besides fair content recommendation systems, allocations based on market equilibria apply to several other problems that are related to large-scale internet applications where no real money is involved. In the *robust content review problem*, we are faced with the task of filtering several types of harmful social media content (e.g., fake news, impersonation, hate speech, ...). Let us consider n categories of harmful content, each with some forecasted amount of content to review in each time period. We also have *m review groups*, which are groups of reviewers (typically in different geographic locations) that have been trained to handle a certain subset of harmful content categories. Each review group has some total amount of reviewing capacity that can be provided during each time period. The goal is to allocate review time to the content categories in a way that satisfies all forecasted review amounts, and then allocate the excess reviewing capacity across the content types to be robust to variations from the forecast. Allouah et al. (2022) show that this problem can be formulated as a variation of the Fisher market equilibrium problem.

Another fair allocation problem that is a component of large-scale internet applications is recommending donation opportunities to people who previously registered as being interested in donating blood (McElfresh et al., 2020). Opportunities arise from requests by blood centers, temporary events such as blood drives, and emergency situations where blood is needed to save lives. The opportunities are submitted to a social network or a donation-specific app, and the recommendation system has to decide what set of users to offer the donation suggestion to. The goal is to optimally allocate suggestions in a way that maximizes the total amount of blood donated, while also treating each donation opportunity equitably.

4.3 Historical Notes

The CEEI solution concept was introduced by Varian et al. (1974). Assigning course seats to students fairly via market equilibrium was studied by Budish (2011). Goldman and Procaccia (2015) created an online service called spliddit.org which has a user-friendly interface for fairly dividing many things such as estates, rent, fares, and others. The motivating example of fair recommender systems, in which we fairly divide impressions among content creators via CEEI was suggested in Kroer et al. (2019); Kroer and Peysakhovich (2019); Murray et al. (2020a).[2] The robust content review problem was introduced by Allouah et al. (2022), where they also show that extensions of CEEI lead to desirable properties. The blood donations problem was introduced by McElfresh et al. (2020). A comprehensive overview of recent fair division work was given by Freeman and Shah (2020). There are also

[2] See also Murray et al. (2020b).

interesting fair division problems related to large-scale internet settings where CEEI is not the preferred solution method. An interesting example of this is the sharing a large set of compute resources such as cloud computing infrastructure or a large compute cluster (Parkes et al., 2015; Ghodsi et al., 2018).

5 Computing Large-Scale Market Equilibria

In Sect. "Introduction to Fisher Markets and Market Equilibrium" we pointed out that one can compute equilibria of Fisher markets solving the Eisenberg-Gale convex program (EG). In this section we discuss the practical considerations of solving this problem for large instances.

This EG formulation can be solved with off-the-shelf software, though it requires the ability to solve convex programs involving exponential cones. For small instances these can be handled with open-source solvers such as SCS (O'Donoghue et al., 2016), and are readily expressed using the CVXPY interface (Diamond and Boyd, 2016). However, in our experience, open-source solvers quickly run into numerical issues for solving EG (around 120 buyers and goods). For moderate-to-large problems, we can rely on the commercial conic solver *Mosek* (Mosek, 2010). If the model is such that the interior-point method in Mosek is able to perform iterations then this is typically the best approach. Mosek is very fast, and has industrial-grade capacity for dealing with numerical issues.

However, for extremely large-scale problems such as fair recommender systems or ad markets, interior-point methods encounter difficulties. This is because the linear system solved at each iteration becomes too slow since the solver does not take sufficient advantage of sparsity, or due to memory usage. To address this and keeping in mind that the purpose of the market formulations discussed throughout this chapter was being able to find equilibria, we now discuss methods suitable for large-scale problems. The approach we describe consists of two complementary elements. First, we will discuss *first-order methods* (FOMs), also known as *gradient-based methods*, for computing market equilibrium. The key selling point of such family of methods is that each iteration of the algorithm can be computed in roughly linear time, and storage costs are low. Second and as a complement to FOMs, we will discuss *abstraction methods*. The goal is to abstract a large Fisher market a smaller one that can be solved efficiently without a big loss in accuracy. Indeed, we propose a way to do this so the equilibria resulting from a reduced-size instance are at approximate equilibrium in the original market. This is crucial in cases where the original market instance is so large that we cannot efficiently store even the explicit iterates. Furthermore, abstraction methods can be used to deal with missing data by leveraging *low-rank models*.

5.1 Convex Programming Formulations for Fisher Markets

We will cover two different algorithms for computing market equilibria of a standard Fisher market model. We will then describe how these can be extended to handle quasilinear utilities. The methods will be based on the EG convex program, as well as its dual convex program:

$$\begin{aligned} \max_{x\geq 0,u} \ & g_{EG}(u) := \sum_{i=1}^{n} B_i \log(u_i) \\ s.t. \quad & u_i \leq \sum_{j=1}^{m} x_{ij} v_{ij}, \quad i \in N, \\ & \sum_{i=1}^{n} x_{ij} \leq 1, \quad j \in M, \end{aligned} \qquad \begin{aligned} \min_{p\geq 0,\beta\geq 0} \ & \sum_{j=1}^{m} p_j - \sum_{i=1}^{n} B_i \log(\beta_i) \\ s.t. \quad & p_j \geq v_{ij}\beta_i, \quad i \in N, j \in M\,. \end{aligned}$$

The convex program on the left is the linear-utility version of the EG program (Eisenberg and Gale, 1959); see Sect. "Convex Programming and Utility Functions" for the general case and Sect. "First Price Auction Markets" for the quasi-linear utility case in the context of FPPE. The dual of EG is also of interest so we included it on the right. An interesting note on the dual of EG is that it optimizes over prices of goods p_j, and per-buyer *utility prices* β_i. At market equilibrium, β_i is exactly the rate at which buyer i derives utility, i.e., $\beta_i = \frac{B_i}{u_i}$.

The convex program below is called the *Shmyrev* formulation and it looks very different from EG. It optimizes over the *spends* b_{ij} (the amount of money that buyer i spends on good j) instead of over allocations directly.

$$\begin{aligned} \max_{b\geq 0,p} \ & f_{sh}(b,p) := \sum_{i=1}^{n}\sum_{j=1}^{m} b_{ij} \log v_{ij} - \sum_{j=1}^{m} p_j \log p_j \\ s.t. \quad & \sum_{i=1}^{n} b_{ij} = p_j, \quad j \in M\,, \\ & \sum_{j=1}^{m} b_{ij} = B_i, \quad i \in N\,. \end{aligned}$$

The first constraint ensures that spends sum to the price of each good, while the second constraint ensures that each buyer spends their budget exactly. The objective is a value-weighted linear combination of spends plus an unscaled entropy regularizer on prices. While the Shmyrev convex program was introduced by Shmyrev (2009) as a new formulation for computing equilibria in Fisher markets, it turns out to be intimately related to EG. Cole et al. (2017) show that the Shmyrev program can be recovered from EG by first taking the dual of EG, applying a change of variables, and then taking the dual again. Despite this equivalence, we shall see that interesting and different algorithms result from solving each convex program.

5.2 First-Order Methods

We now describe two simple and scalable algorithms that arise from the convex programs mentioned above. The first algorithm we will describe is the *proportional response* (PR) algorithm. The PR algorithm is an iterative algorithm that can be viewed as a dynamic updating scheme between buyers and the seller. The buyers see current prices on goods and update their bids, while the seller sees these bids and in turn update the price. This can be summarized as follows:

Algorithm 2: Proportional response (PR)

1. Each buyer i submits a bid $b_{ij}^{t=1} \in \mathbb{R}_+^m$ for each good j.
2. **for** *every time step* $t = 1, \ldots$ **do**
3. Given the bids, a price $p_j^t = \sum_i b_{ij}^t$ is computed for each good.
4. Each buyer is assigned an allocation $x_{ij}^t = \frac{b_{ij}^t}{p_j^t}$ of each good.
5. Each buyer submits a next bid on good j proportionally to the utility they received from good j in round t:

$$b_{ij}^{t+1} = B_i \frac{x_{ij}^t v_{ij}}{\sum_{j'} x_{ij'}^t v_{ij'}} .$$

As it is evident from the price and bid updating schemes, these updates are designed such that they alternatively correspond to each of the constraints in the Shmyrev program. The next theorem provides a convergence rate for this algorithm as a function of the size of the instance and the time period. It shows that PR has a reasonably attractive $1/T$ rate of convergence.

Theorem 7 *The iterates of the PR algorithm converge at the rate of $f_{sh}(b^*, p^*) - f_{sh}(b^t, p^t) \leq (\log nm)/t$, where b^* and p^* denote any optimal solution to the Shmyrev convex program.*

From a practical perspective, the PR algorithm converges very rapidly to a medium-accuracy solution for most numerical examples. Thus, it is a very useful method in practice, since it has a very simple and lightweight implementation, requires no parameter tuning, and can be used for very large instances. This is especially the case if the valuations are *sparse*: we only need a variable b_{ij} corresponding to a buyer-good pair (i, j) if $v_{ij} > 0$.

While we have not described the mirror descent algorithm in this chapter, it is worth making a few comments on the equivalence between the PR and mirror descent algorithms. While this is not immediately apparent from the description of the algorithm, the PR algorithm is a first-order method because it is an application of mirror descent to a convex program. Even though there is no step size in PR (whereas FOMs, including mirror descent, typically have step sizes), the PR dynamics correspond to choosing a step size of one in mirror descent. This turns out to be a valid choice due to a strong connection between the unscaled entropy on prices in the Shmyrev objective, and the way distances are measured when using the negative entropy on the bids as a distance measure. A reader familiar with the typical convergence rates achieved by mirror descent would expect that one must average the iterates across all time steps in order to get convergence, and then this would typically converge at a rate of $1/\sqrt{T}$. Theorem 7 gives a guarantee on the *last iterates* b^t, p^t of PR without averaging, and the rate guarantee is of the order of $1/T$. From a theoretical perspective a rate improvement of $1/\sqrt{T}$ is very strong, and from a practical perspective the last iterate convergence is quite attractive. Both of these properties are again a consequence of the strong connection between the unscaled entropy on prices in the Shmyrev objective and the negative entropy distance measure.

As stated above, PR is a great algorithm for converging to medium-accuracy solutions. However, if one wants higher-accuracy solutions, then a method with a faster asymptotic rate of convergence is necessary. We now describe how this can be achieved via a *projected gradient descent* (PGD) algorithm. PGD operates on the original EG problem. It iterates the following update:

$$x^{t+1} = \Pi_{\mathcal{X}}(x^{t-1} - \gamma_t \nabla g_{EG}(x^t)),$$

where $\mathcal{X} = \{x \in \mathbb{R}_+^{n\times m} : \sum_i x_i = s\}$ is the set of feasible allocations that fully utilize every good,[3] $\Pi_{\mathcal{X}}$ is the projection operator onto $\mathcal{X}$, and γ_t is the step size. Gao and Kroer (2020) show that PGD on EG converges at a linear rate $(1-\delta)^t$ for a small constant δ. Thus, theoretically, the PGD algorithm should be preferred to PR when higher accuracy is needed. This result does come with some caveats however: first, the base of the exponent, $1-\delta$, can be close to 1 as the term δ depends on some hard-to-compute constants, such as the *Hoffman* constant which makes δ very small. Second, the cost per iteration for PGD is slightly higher than for PR: it requires projecting onto $\mathcal{X}$, which can be done via sorting but leads to a per-iteration cost of $nm \log n$, as opposed to nm for PR. Here too the projection can take advantage of sparsity, thereby becoming much faster in the case where only a few buyers are interested in each good. Gao and Kroer (2020) also show numerically that PR is faster than PGD for medium-accuracy solutions, whereas PGD is faster when higher levels of accuracy are desired.

Both PR and PGD algorithms can also be extended to the case of quasilinear utilities, such as those used when modeling budget-smoothing in first price auctions as a market equilibrium problem. For PGD, this relies on extending EG to the quasilinear case, which was shown by Chen et al. (2007) and Cole et al. (2017). Gao and Kroer (2020) show how to apply PGD to achieve a linear rate.

[3]Every optimal solution to EG must lie in this set, assuming that every good j has some i such that $v_{ij} > 0$; if this does not hold then that good can simply be preprocessed away.

In this chapter, we focused on simple and scalable first-order methods for computing market equilibria. We believe these algorithms are very suitable for practical use, due to their easy implementation requirements and cheap per-iteration complexity. Beyond the algorithms covered here, there is a long history of theory-oriented polynomial-time algorithms for computing equilibria of Fisher markets. This line of work started with Devanur et al. (2008) who give a primal-dual algorithm. Later work offered simpler algorithms, e.g., Bei et al. (2019) put forward an algorithm with a linear rate that can be interpreted as a form of ascending-price auctions. However, these methods typically do not have cheap per-iteration complexity, and are thus less suitable for the market sizes considered in this chapter. Another interesting line of extensions studies algorithms for new utility classes such as, e.g., *spending constraint utilities*, which are additively separable utilities that have piecewise-linear utility per good (Vazirani, 2010; Birnbaum et al., 2011).

5.3 Market Abstractions

So far we have described scalable first-order methods for computing market equilibria. Still, these algorithms make a number of assumptions that may not hold in practice. To use the PR or PGD algorithms, one must be able to store the iterates which take nm space. If both the number of buyers and goods are of the order of 100,000, writing down an iterate using 64-bit floats requires about 80 GB of memory (assuming no sparsity). For applications in large internet companies such as ad markets, we might expect n, and especially m, to be even larger than that. Thus efficient computation is important but not enough. We may need to find a way to compress the instances to be solved down to some manageable size where we can at least hope to store iterates efficiently. Furthermore, we may not have access to all valuations v_{ij}. For instance, we may only have some samples and those values may be noisy. This means that we also need to somehow infer the remaining valuations. In a setting where we do not know all the true valuations, or we only have noisy estimates, it is important to understand how these misspecifications degrade the quality of computed equilibria.

The issues mentioned earlier motivated Kroer et al. (2019) to consider abstraction methods to solve those problems. For the purposes of abstraction, it will be useful to think of the set of valuations v_{ij} as a matrix V, where the i'th row corresponds to the valuation vector of buyer i. We will be interested in the outcome of computing a market equilibrium using a valuation matrix $\tilde{V} \approx V$, where $\tilde{V}$ would typically be obtained from an abstraction method. The goal is establishing that the market equilibria corresponding to V and $\tilde{V}$ are similar, which could be quantified by $\|\tilde{V} - V\|_F$. That would enable us to compute one equilibrium to approximate the other. Let us enumerate a couple of reasons why one might prefer to compute a market equilibrium for $\tilde{V}$ rather than V.

Low-rank markets: When there are missing valuations, we need to impute the missing values. Of course, if there is no relationship between the entries of V that are observed and those that are missing, then we have no hope of recovering V. However, in practice this is typically not the case. Valuations are often assumed to (approximately) belong to some low-dimensional space. A popular model is to assume that the valuations are *low rank*, meaning that every buyer i can be represented by some d-dimensional vector ϕ_i, every good j can also be represented by some d-dimensional vector ψ_j, and the valuation of buyer i for good j is $\tilde{v}_{ij} = \langle \phi_i, \psi_j \rangle$. One may interpret this model as every good having an associated set of *d features*, with ψ_j describing the value for each feature, and ϕ_i describing the value that i places on each feature. In a low-rank model, d is expected to be much smaller than $\min(n, m)$, meaning that V is far from full rank. If the real valuations are approximately of rank d (meaning that the remaining spectrum of V is very small), then $\tilde{V}$ will be close to V.

This model can also be motivated via the *singular-value decomposition* (SVD). Assume that we wish to find the matrix of rank d that is closest to V:

$$\begin{aligned} \min_{\tilde{V}} \quad & \|V - \tilde{V}\|_F^2 \quad := \sum_{ij} (v_{ij} - \tilde{v}_{ij})^2 \\ s.t. \quad & \operatorname{rank}(\tilde{V}) \le d\,. \end{aligned}$$

The optimal solution to this problem can be found easily via SVD. Letting $\sigma_1, \ldots, \sigma_d$ be the first d singular values of V, $\bar{u}_1, \ldots, \bar{u}_d$ the first left singular vectors, and $\bar{v}_1, \ldots, \bar{v}_d$ the first right singular vectors, the optimal solution is $\tilde{V} = \sum_{k=1}^{d} \sigma_k \bar{u}_k \bar{v}_k^T$. If the remaining singular values $\sigma_{k+1}, \ldots$ are small relative to the first k singular values, then this model captures most of the valuation structure.

In practice, since the matrix V might not be known exactly, we cannot solve this problem to get $\tilde{V}$. Instead, we search for a low-rank model that minimizes some loss on the observed entries OBS, e.g., $\sum_{i,j \in \mathrm{OBS}} (v_{ij} - \langle \phi_i, \psi_j \rangle)^2$ (this objective is typically also regularized by the Frobenius norm of the low-rank matrices). Under the assumption that V is generated from a true low-rank model via some simple distribution, it is possible to recover the original matrix with only samples of entries by minimizing the loss on observed entries. In practice this approach is also known to perform extremely well, and it is used extensively at tech companies. The hypothesis is that in practice data is approximately low rank, so one does not lose much accuracy from a rank-d model.

Representative Markets: It is also convenient to generate a smaller set of representative buyers, where each original buyer i maps to some particular representative buyer $r(i)$. Similarly, we may generate representative goods that correspond to many non-identical but similar goods from the original market. These representative buyers and goods may be generated using clustering techniques. In this case, our approximate valuation matrix $\tilde{V}$ has as row i the valuation vector of the representative buyer $r(i)$. This means that all i, i' such that $r(i) = r(i')$ have the same valuation vector in $\tilde{V}$, and thus they can be treated as a single buyer for equilibrium-computation purposes. The same grouping can also be applied to the goods. If the number of buyers and goods is reduced by a factor of q, then the resulting problem size is reduced by a factor of q^2, since we have $n \times m$ variables.

We now analyze what happens when we compute a market equilibrium under $\tilde{V}$ rather than V. Throughout this subsection we will let $(\tilde{x}, \tilde{p})$ be a market equilibrium for $\tilde{V}$. We use the error matrix $\Delta V = V - \tilde{V}$ to quantify the solution quality, and we measure the size of ΔV using the ℓ_1-ℓ_∞ matrix norm $\|\Delta V\|_{1,\infty} = \max_i \|\Delta v_i\|_1$. We will also use the norm of the error vector for an individual buyer $\|\Delta v_i\|_1 = \|v_i - \tilde{v}_i\|_1$.

The next proposition turns out to be useful in proving guarantees on the approximate equilibrium. It establishes that under linear utility functions the change in utility when going from v_i to $\tilde{v}_i$ is linear in Δv_i.

Proposition 1 *If* $\langle \tilde{v}_i, x_i \rangle + \epsilon \geq \langle \tilde{v}_i, x_i' \rangle$ *then* $\langle v_i, x_i \rangle + \epsilon + \|\Delta v_i\|_1 \geq \langle v_i, x_i' \rangle$.

This proposition can be used to immediately derive bounds on envy, proportionality, and regret (how far each buyer is from achieving the utility of their demand bundle). For example, we know that under $\tilde{V}$, each buyer i has no envy towards any other buyer k: $\langle \tilde{v}_i, \tilde{x}_i \rangle \geq \langle \tilde{v}_i, \tilde{x}_k \rangle$. By Proposition 1 each buyer i has envy at most $\|\Delta v_i\|_1$ under V when using $(\tilde{x}, \tilde{p})$. All envies are thus bounded by $\|\Delta V\|_{1,\infty}$. Regret and proportionality can be bounded similarly which implies guarantees under $\tilde{V}$.

Market equilibria also guarantee Pareto optimality. Unfortunately, we cannot give any meaningful guarantee on how much social welfare improves under Pareto-improving allocations for $\tilde{V}$. The following real and abstracted matrices provide an example of it:

$$V = \begin{bmatrix} 1 & \epsilon & \epsilon \\ 0 & 1 & \epsilon \end{bmatrix}, \quad \tilde{V} = \begin{bmatrix} 1 & \epsilon & 0 \\ 0 & 1 & \epsilon \end{bmatrix}.$$

If we set $B_1 = B_2 = 1$, then for supply-aware market equilibrium, we end up with competition only on good 2, and we get prices $\tilde{p} = (0, 2, 0)$ and allocation $\tilde{x}_1 = (1, 0.5, 0)$, $\tilde{x}_2 = (0, 0.5, 1)$. Under V this is a terrible allocation, and we can Pareto improve by using $x_1 = (1, 0, 0.5)$, $x_2 = (0, 1, 0.5)$, which increases overall social welfare by $\frac{1}{2} - \epsilon$, in spite of $\|\Delta V\|_1 = \epsilon$.

On the other hand, we can show that under any Pareto-improving allocation, some buyer i improves by at most $\|\Delta V\|_{1,\infty}$. To see this, note that for any Pareto improving allocation x, under $\tilde{V}$ there existed at least one buyer i such that $\langle \tilde{v}_i, \tilde{x}_i - x_i \rangle \geq 0$, and so this buyer must improve by at most $\|\Delta v_i\|_1$ under V.

5.4 Historical Notes

The proportional response (PR) algorithm was introduced by Wu and Zhang (2007). It was later shown to be effective for BitTorrent sharing dynamics (Levin et al., 2008), and it was eventually shown to be an instantiation of the *mirror descent* algorithm by Birnbaum et al. (2011), who also show the last-iterate $\frac{1}{T}$ rate. Birnbaum et al. (2011) also show how mirror descent can be applied to the Shmyrev formulation. For a general introduction to mirror descent, see e.g., Bubeck (2015) or the lecture notes of Ben-Tal and Nemirovski (2001). Alternatively, the lectures notes from Kroer (2020) also cover this derivation, as well as the general duality derivations required to obtain Shmyrev from EG.

Abstraction in the context of Fisher market equilibrium was introduced by Kroer et al. (2019). However, abstraction has been applied in other related problems. Candogan et al. (2016) consider replacing sets of agents in trading networks with representative buyers, in order to do comparative statics. There has also been work on abstraction for non-market-based allocation problems (Walsh et al., 2010; Lu and Boutilier, 2015; Peng and Sandholm, 2016), where the results largely center around good abstraction of LP or MIP-based allocation problems. Abstraction has also been studied in the computation of game-theoretic equilibria, where it has been used extensively in practice (Gilpin et al., 2007; Brown et al., 2015; Brown and Sandholm, 2018), and studied from a theoretical perspective (Lanctot et al., 2012; Kroer and Sandholm, 2014, 2016, 2018).

A brief introduction to low-rank models can be found in Udell (2019). Udell et al. (2016) gives a more thorough exposition and describes more general model types. Beyond these papers, there is a large theory of low-rank models that show a number of interesting results. There is a class of nuclear-norm-regularized convex optimization problems that can recover the original matrix with only a small number of entry samples (Candès and Recht, 2009; Recht, 2011). One might think that this would then be the preferred method in practice, but surprisingly non-convex models are often preferred instead. These non-convex methods also have interesting guarantees on statistical recovery under certain assumptions. An overview of non-convex methods is given in Chi et al. (2019). Low-rank market equilibrium models were also studied in Kroer and Peysakhovich (2019), where it is shown that large low-rank markets enjoy a number of properties not satisfied by small-scale markets.

6 Conclusion

We have described how market models can be used to generate insights and compute solutions that are relevant to internet platforms. We have given examples in the context of ad auctions, recommendation systems and fair division problems. The general area is very rich and there is ample research published about these applications. For additional context on online advertising, particularly as it relates to display ads, we direct the reader to a survey by Choi et al. (2020). For some context on market equilibrium in electricity markets, see Azizan et al. (2020) and references therein. A somewhat dated overview of work in the theoretical computer science community on computing market equilibrium can be found in Nisan et al. (2007). The references in Sect. "Historical Notes" provide some examples of more recent work. For more context on fair division, we recommend several surveys covering various aspects of this problem. Aleksandrov and Walsh (2020) gives a very recent overview of *online* fair division, a problem highly related to the topics covered here. For more general fair division coverage, see, e.g., Procaccia (2013), Brandt et al. (2016), and Walsh (2020).

To use the ideas discussed in this chapter in practice, it is important to feed models with the right input. Some data may be available from historical information, logs, and other measurements, while other data may need to be estimated. To do the latter, one can rely on several statistical and machine learning techniques. In relation to advertising, clustering, recommender systems, and particularly to calibrate values of goods to buyers, we refer the reader to the chapter by Bastani et al. (2021) in this volume about the interplay between machine learning and operations.

To conclude we mention some open problems and directions, as of this article. For the general problem of computing market equilibrium, there are several interesting open questions. While the general Fisher market equilibrium can be computed efficiently, as discussed in earlier chapters, the refinement of market equilibria that is needed for SPPE is harder to handle. It was shown by Conitzer et al. (2018) that maximizing various objectives over the set of SPPE equilibria is NP-complete, but the complexity of finding any SPPE is currently unknown. In a similar vein, one might investigate the existence of approximate methods for finding SPPE, even if exact SPPE are hard to find. A related question is how to reconcile these potential hardness results with the fact that under independence assumptions, Balseiro and Gur (2019) show that it is possible to have buyers converge to an SPPE-like equilibrium in an online setting. Understanding the exact boundaries of what can be learned online is an important practical question.

Another interesting line of work would be to generalize pacing equilibria to more realistic allocation mechanisms. In practice, each item is not sold via independent first or second price auctions. Instead, the number of items sold per auction is the number of slots available in a given impression, and a single ad is typically only allowed to win one of those slots. This breaks the correspondence with market equilibrium, but an appropriate notion of pacing still exists. It would be interesting to understand which, if any, results from market equilibrium carry over to this setting. There is a rich space of possible questions to ask for this problem, based on the type of multi-item allocation mechanism being run, the presence of reserve prices, and so on.

Finally, we mentioned in the introduction that various mechanisms have been used in practice to allocate goods to buyers. It would be interesting to further explore the relations between the models presented in this chapter on market mechanisms and their underlying situations in the technology industry to bond auctions, supply function equilibria, and other market approaches.

References

Aleksandrov, M., & Walsh, T. (2020). Online fair division: A survey. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(9), 13557–13562.

Allouah, A., Avadhanula, V., Kroer, C., Pupyrev, S., Shah, P., & Zhang, X. (2022). *Robust and fair work allocation*. Mimeo.

Azizan, N., Su, Y., Dvijotham, K., & Wierman, A. (2020). Optimal pricing in markets with nonconvex costs. *Operations Research, 68*(2), 480–496.

Balseiro, S. R., & Gur, Y. (2019). Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science, 65*(9), 3952–3968.

Balseiro, S. R., Besbes, O., & Weintraub, G. Y. (2015). Repeated auctions with budgets in ad exchanges: Approximations and design. *Management Science, 61*(4), 864–884.

Balseiro, S. R., Kim, A., Mahdian, M., & Mirrokni, V. (2017). Budget management strategies in repeated auctions. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*, pp. 15–23.

Bastani, H., Zhang, D. J., & Zhang, H. (2021). Applied machine learning in operations management. In V. Birge Babich & J. G. Hilary (Eds.), *Innovative technology at the interface of finance and operations*. *Series in supply chain management*. Berlin: Springer.

Bei, X., Garg, J., & Hoefer, M. (2019). Ascending-price algorithms for unknown markets. *ACM Transactions on Algorithms, 15*(3), 1–33.

Ben-Tal, A., & Nemirovski, A. (2001). *Lectures on modern convex optimization: Analysis, algorithms, and engineering applications*. Philadelphia: SIAM. Retrieved November 2020 from https://www2.isye.gatech.edu/~nemirovs/LMCOLN2020WithSol.pdf

Bertsekas, D. P., Nedic, A., & Ozdaglar, A. (2003). *Convex analysis and optimization*. Nashua: Athena Scientific.

Birnbaum, B., Devanur, N. R., & Xiao, L. (2011). Distributed algorithms via gradient descent for Fisher markets. In *Proceedings of the ACM Conference on Electronic Commerce (EC'11)* (pp. 127–136). New York: ACM.

Borgs, C., Chayes, J., Immorlica, N., Jain, K., Etesami, O., & Mahdian, M. (2007). Dynamics of bid optimization in online advertisement auctions. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)* (pp. 531–540).

Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (2016) *Handbook of computational social choice*. Cambridge: Cambridge University Press.

Brown, N., & Sandholm, T. (2018). Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science, 359*(6374), 418–424.

Brown, N., Ganzfried, S., & Sandholm, T. (2015). Hierarchical abstraction, distributed equilibrium computation, and post-processing, with application to a champion no-limit Texas hold'em agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS'15)* (pp. 7–15). International Foundation for Autonomous Agents and Multiagent Systems.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning, 8*(3–4), 231–357.

Budish, E. (2011). The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy, 119*(6), 1061–1103.

Budish, E., Cachon, G. P., Kessler, J. B., & Othman, A. (2016). Course match: A large-scale implementation of approximate competitive equilibrium from equal incomes for combinatorial allocation. *Operations Research, 65*(2):314–336.

Budish, E., Che, Y.-K., Kojima, F., & Milgrom, P. (2013). Designing random allocation mechanisms: Theory and applications. *American Economic Review, 103*(2), 585–623.

Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics, 9*(6), 717.

Candogan, O., Epitropou, M., & Vohra, R. V. (2016). Competitive equilibrium and trading networks: A network flow approach. In *Proceedings of the ACM Conference on Economics and Computation (EC'16)* (pp. 701–702).
Caragiannis, I., Kurokawa, D., Moulin, H., Procaccia, A. D., Shah, N., & Wang, J. (2019). The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation, 7*(3), 1–32.
Chen, L., Ye, Y., & Zhang, J. (2007). A note on equilibrium pricing as convex optimization. In *International Workshop on Web and Internet Economics (WINE'07)* (pp. 7–16). Berlin: Springer.
Chi, Y., Lu, Y. M., & Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing, 67*(20), 5239–5269.
Choi, H., Mela, C. F., Balseiro, S. R., & Leary, A. (2020). Online display advertising markets: A literature review and future directions. *Information Systems Research, 31*(2):556–575.
Cole, R., Devanur, N. R., Gkatzelis, V., Jain, K., Mai, T., Vazirani, V. V., & Yazdanbod, S. (2017). Convex program duality, Fisher markets, and Nash social welfare. In *ACM Conference on Economics and Computation (EC'17)*. New York: Association for Computing Machinery.
Conitzer, V., Kroer, C., Sodomka, E., & Stier-Moses, N. E. (2018). Multiplicative pacing equilibria in auction markets. In *International Conference on Web and Internet Economics(WINE'18)*.
Conitzer, V., Kroer, C., Panigrahi, D., Schrijvers, O., Sodomka, E., Stier-Moses, N. E., & Wilkens, C. (2019). Pacing equilibrium in first-price auction markets. In *Proceedings of the ACM Conference on Economics and Computation (EC'19)*. New York: ACM.
Debreu, G. (1952). A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences, 38*(10), 886–893.
Devanur, N.R., Papadimitriou, C.H., Saberi, A., & Vazirani, V. V. (2008). Market equilibrium via a primal–dual algorithm for a convex program. *Journal of the ACM, 55*(5), 1–18.
Diamond, S., & Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research, 17*(83), 1–5.
Eisenberg, E. (1961). Aggregation of utility functions. *Management Science, 7*(4), 337–350.
Eisenberg, E., & Gale, D. (1959). Consensus of subjective probabilities: The pari-mutuel method. *The Annals of Mathematical Statistics, 30*(1), 165–168.
Fan, K. (1952). Fixed-point and minimax theorems in locally convex topological linear spaces. *Proceedings of the National Academy of Sciences of the United States of America, 38*(2), 121.
Flajolet, A., & Jaillet,, P. (2017). Real-time bidding with side information. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS'17)* (5168–5178). Red Hook: Curran Associates.
Freeman, R., & Shah, N. (2020). AAAI 2020 Tutorial on Recent Advances in Fair Resource Allocation. Retrieved November 2020 from https://users.cs.duke.edu/~rupert/fair-division-aaai20/Tutorial-Slides.pdf
Gao, Y., & Kroer, C. (2020). First-order methods for large-scale market equilibrium computation. In *Advances in Neural Information Processing Systems (NeurIPS'20)*.
Ghodsi, M., HajiAghayi, M.-T., Seddighin, M., Seddighin, S., & Yami, H. (2018). Fair allocation of indivisible goods: Improvements and generalizations. In *Proceedings of the ACM Conference on Economics and Computation (EC'18)* (pp. 539–556).
Gilpin, A., Sandholm, T., & Sørensen, T. B. (2007). Potential-aware automated abstraction of sequential games, and holistic equilibrium analysis of Texas hold'em poker. In *Proceedings of the 22nd National Conference on Artificial Intelligence* (vol. 1, pp. 50–57). Palo Alto: AAAI Press.
Glicksberg, I. L. (1952). A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proceedings of the American Mathematical Society, 3*(1), 170–174.
Goldman, J., & Procaccia, A. D. (2015). Spliddit: Unleashing fair division algorithms. *ACM SIGecom Exchanges, 13*(2), 41–46.
Kroer, C. (2020). IEOR8100: Economics, AI, and optimization lecture note 14: Large-scale fisher market equilibrium.

Kroer, C., & Peysakhovich, A. (2019). Scalable fair division for ‘at most one’ preferences. arXiv:1909.10925.

Kroer, C., & Sandholm, T. (2014). Extensive-form game abstraction with bounds. In *Proceedings of the ACM Conference on Economics and Computation (EC’14)* (pp. 621–638). New York: ACM.

Kroer, C., & Sandholm, T. (2016). Imperfect-recall abstractions with bounds in games. In *Proceedings of the ACM Conference on Economics and Computation (EC’16)* (pp. 459–476). New York: ACM.

Kroer, C., & Sandholm, T. (2018). A unified framework for extensive-form game abstraction with bounds. In *Advances in Neural Information Processing Systems (NeurIPS’18)* (pp. 612–623).

Kroer, C., Peysakhovich, A., Sodomka, E., & Stier-Moses, N. E. (2019). Computing large market equilibria using abstractions. In *Proceedings of the ACM Conference on Economics and Computation (EC’19)* (pp. 745–746).

Lanctot, M., Gibson, R., Burch, N., Zinkevich, M., & Bowling, M. (2012). No-regret learning in extensive-form games with imperfect recall. In *Proceedings of the 29th International Coference on Machine Learning (ICML’12)* (pp. 1035–1042). Madison: Omnipress.

Levin, D., LaCurts, K., Spring, N., & Bhattacharjee, B. (2008). Bittorrent is an auction: Analyzing and improving bittorrent’s incentives. In *Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication* (pp. 243–254).

Lu, T., & Boutilier, C. (2015). Value-directed compression of large-scale assignment problems. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Menlo Park: AAAI Press.

McElfresh, D. C., Kroer, C., Pupyrev, S., Sodomka, E., Sankararaman, K. A., Chauvin, Z., Dexter, N., & Dickerson, J. P. (2020). Matching algorithms for blood donation. In *Proceedings of the ACM Conference on Economics and Computation (EC’20)* (pp. 463–464).

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (2018). Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 2243–2251).

Mosek ApS. (2010). The MOSEK optimization software. http://www.mosek.com

Murray, R., Kroer, C., Peysakhovich, A., & Shah, P. (2020a). Robust market equilibria with uncertain preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Murray, R., Kroer, C., Peysakhovich, A., & Shah, P. (2020b). Robust market equilibria: How to model uncertain buyer preferences. Retrieved December 2020, from https://research.fb.com/blog/2020/09/robust-market-equilibria-how-to-model-uncertain-buyer-preferences

Nisan, N., Roughgarden, T., Tardos, E., & Vazirani, V. V. (2007). *Algorithmic game theory*. Cambridge: Cambridge University Press.

O’Donoghue, B., Chu, E., Parikh, N., & Boyd, S. (2016). Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications, 169*(3), 1042–1068.

Paes Leme, R., Sivan, B., & Teng, Y. (2020). Why do competitive markets converge to first-price auctions? In *Proceedings of The Web Conference 2020, WWW’20* (pp. 596–605). New York: Association for Computing Machinery.

Parkes, D. C., Procaccia, A. D., & Shah, N. (2015). Beyond dominant resource fairness: Extensions, limitations, and indivisibilities. *ACM Transactions on Economics and Computation, 3*(1), 1–22.

Peng, F., & Sandholm, T. (2016). Scalable segment abstraction method for advertising campaign admission and inventory allocation optimization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI’16)* (pp. 655–661). Menlo Park: AAAI Press.

Procaccia, A. D. (2013). Cake cutting: Not just child’s play. *Communications of the ACM, 56*(7), 78–87.

Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research, 12*, 3413–3430.

Shmyrev, V. I. (2009). An algorithm for finding equilibrium in the linear exchange model with fixed budgets. *Journal of Applied and Industrial Mathematics, 3*(4), 505.

Talluri, K., & Van Ryzin, G. (1998). An analysis of bid-price controls for network revenue management. *Management Science, 44*, 1577–1593.

Udell, M. (2019). Big data is low rank. SIAG/OPT Views and News.

Udell, M., Horn, C., Zadeh, R., & Boyd, S. (2016). Generalized low rank models. *Foundations and Trends® in Machine Learning, 9*(1), 1–118.

Varian, H. R. (1974). Equity, envy, and efficiency. *Journal of Economic Theory, 9*(1), 63–91.

Vazirani, V. V. (2010). Spending constraint utilities with applications to the adwords market. *Mathematics of Operations Research, 35*(2):458–478.

Walsh, T. (2020). Fair division: The computer scientist's perspective. arXiv:2005.04855.

Walsh, W. E., Boutilier, C., Sandholm, T., Shields, R., Nemhauser, G., & Parkes, D. C. (2010). Automated channel abstraction for advertising auctions. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (pp. 887–894). Menlo Park: AAAI Press.

Wu, F., & Zhang, Li. (2007). Proportional response dynamics leads to market equilibrium. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing (STOC'07)* (pp. 354–363).

Large-Scale Price Optimization for an Online Fashion Retailer

Hanwei Li, David Simchi-Levi, Rui Sun, Michelle Xiao Wu, Vladimir Fux, Torsten Gellert, Thorsten Greiner, and Andrea Taverna

1 Introduction

Markdown and promotional pricing have been popular in the operations management field, where researchers and companies try to understand consumers' purchasing behaviour by building demand forecasts and optimize the markdowns. Within the field, the online fashion retailing environment has some unique features and challenges. Firstly, in the online retailing environment, it is common for a company to manage a large number (typically hundreds of thousands) of products across different markets and product categories. The scale of the problem imposes requirements on the efficiency of the optimization framework as it must be executed on a weekly or even daily basis. To make things more complicated, there exist business targets that tie individual products together. For instance, from the business planning perspective, the company may set a certain revenue target for specific product categories or individual countries, which is a constraint that applies to all the products within the category and/or country. The challenge is that jointly optimizing all the products may be intractable and decomposition methods need to be applied. Another challenge is that, even for single-product discount optimization, we would have specific business constraints that require careful modelling. As an example, due to the fear of confusing customers or receiving negative customer response, discounts of a single product should not change drastically from week to week. If

H. Li (✉), D. Simchi-Levi, R. Sun, M. X. Wu
MIT Data Science Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: hanweili@mit.edu; dslevi@mit.edu; ruisun@mit.edu; wuxiao@mit.edu

V. Fux, T. Gellert, T. Greiner, A. Taverna
Zalando SE, Berlin, Germany
e-mail: vladimir.fux@zalando.de; torsten.gellert@zalando.de; thorsten.greiner@zalando.de; andrea.taverna@zalando.de

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_8

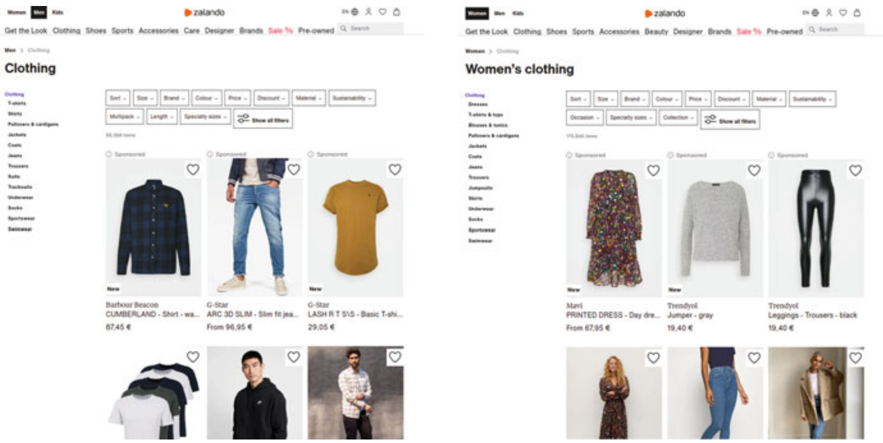

Fig. 1 Zalando's homepages for the "Men" (left) and "Women" (right) segments

these business constraints lead to a non-linear optimization model even for single products, the corresponding "global" problem with multiple products will have a higher level of complexity and would be impossible to solve. Finally, due to the nature of fashion industry and the large scale of products, there is a long tail of products with very few or no historical sales data. This makes both demand forecast and optimization challenging.

In this work, the MIT Data Science Lab (http://dsl.mit.edu/) has been collaborating with Zalando (www.zalando.de), a Europe's leading online fashion platform, which delivers products to more than 28 million active customers in 17 countries. On the fashion platform, customers can find a wide assortment of around 600,000 articles from more than 2500 brands for a total of €6.5 billion yearly sales as of 2019. Zalando offers customers a comprehensive selection of apparel, footwear, cosmetics, and accessories for women, men, and children with free shipping and return shipping. A screenshot of the web page is presented in Fig. 1. Zalando's logistics network with five central logistics centers enables efficient delivery to all customers throughout Europe, supported by local distribution focused on local customer needs in northern Italy, France and Sweden. The triad of fashion, technology, and logistics offers added value to the customers and brand partners.

The goal of the collaboration is to manage the prices of products through a sales season, so that profit is maximized and overstock at the end of the season is minimized. Additionally, the process aims to distribute discounts in such a way that business targets (e.g. growth in a certain market) are reached. The implementation of this process in the past has been a hybrid of an automated optimization process, which recommends profit optimal discounts for individual products, and a manual selection process from these discount recommendations. The manual selection process builds on the experience of pricing experts for the individual markets which use their experience to select the discounts so that the business targets for the markets are reached.

There are a few challenges in the manual process. First, as Zalando's business grow, scaling this process is not easy. Second, because Zalando has multiple targets, by product category and country level, intuition and experience are not enough. Finally, it is not clear that the current process generates effective discounts. Therefore, our objective is to develop a fully automated system that satisfies the various business requirements and allows Zalando to maximize impact on the bottom line.

1.1 Problem Statement

As most of the online retailers, Zalando has a regular price update cycle that aims to ensure that each product (or Stock Keeping Units, SKU) receives an optimal discount. Weekly prices updates are typically synchronized between all markets, while prices in different countries (or markets) are set independently. The shop operates with a global stock assumption that every product can be sold in any country without limitations by sharing a common inventory pool. This specifically means that we do not have control over the sales directly: it is not an option to turn down a customer to reserve the product for customers in other countries, and we can only impact sales through appropriate discounts.

Weekly selection of prices for such a large assortment is near impossible to do manually. That is the reason why the first generation of automatic price recommendation system was introduced in Zalando, aiming to maximize total profit. The insights were to discount heavily the SKUs that are expected to have large overstock (e.g. due to optimistic buying decisions, or drastic changes in season's weather) and to discount conservatively with SKUs which are selling well in the shop. Such automatic system proved to be beneficial and led to significant increases in financial indicators, and contributed to successful growth of the company.

However, such system lacks an important feature, which is crucial for a large scale multinational business. Having a drastic financial impact, pure discounting does not allow to steer towards certain financial targets and as such does not allow easily to include strategic company goals in the pricing process. This leads to a situation where discounts produced by the system still need manual intervention from commercial planners, in order to steer discounts in the direction of financial targets. We also refer to the latter as "global" constraints in this paper since our model will have to solve at a more "global level" instead of by single products.

The increase in manual efforts reveals a need for a new generation of pricing system, which we call "target steering" system. The goal of such system is not only to make price recommendations for overstock mitigation but also to provide discounts, which satisfy weekly financial targets (like certain revenue levels in a selection of countries). This paper describes the design and implementation details of such a large scale price recommendation system, which proved to be a challenging task at such a large scale.

Given the demand forecasts under a collection of discrete prices, the objective of Zalando is to find the discount levels that maximize the total profit over the entire selling horizon while taking into account both "local" and "global" business constraints. On the "local" single SKU level, due to operational sharing inventory pool assumption, Zalando needs to balance the sales across different countries. Specifically, we have to make sure that when customers arrive, no inventories are reserved strategically for demands in other countries or later time periods (this is referred to as "stock hedging"). To address this challenge, we formulate the price optimization problem as a mixed integer programming (MIP) problem, and incorporate the balancing constraints to solve the "stock hedging" issue. In addition, since our forecaster is at size-aggregated SKU level, due to each SKU's limited availability of different sizes under low stock, we propose a new stock-dependent method to adjust the demand forecasts and provide more accurate input into the pricing system.

On the "global" level, when we jointly optimize the discounts of different SKUs in selected categories, the total revenue and weighted average discounts of the SKUs should meet certain "global" steering targets. This is because pricing managers need to manage the discounts across articles in order to deliver the forecasted discount spend and achieve the revenue targets. The weighted average discounts (or say "discount spend") is an important factor for pricing managers' planning decisions. To reduce the computational complexity raised by these global constraints, we decompose the global problem into parallel sub-problems of each single SKU by using Lagrangian decomposition, and propose an efficient algorithm to find the optimal Lagrangian multipliers.

Finally, to reduce computational complexity, we developed an aggregation framework to cluster SKUs based on categories and similarities. This framework enables solving the problem via a 3-step process where in the first step we solve the aggregated problem in a way that decouples the problem into a problem for each product category. In the second step we apply the Lagrangian method to decouple the problem for each SKU. Finally, in the last step we solve a large number of SKU specific problems. Importantly, the aggregation method is designed to enable planners to test a variety of pricing strategies before finalizing the season's targets for each product category and country.

We also conduct field experiments to validate the optimization system in real business cases. In both the online and offline environment, we design and implement experiments to validate whether the proposed framework is capable of steering towards certain business targets. The results in the target group satisfied the targets, where global constraints are satisfied and the weighted average discount values are close or within the target bounds most of the time. Previously the different global constraints are managed manually by the commercial team using intuitions and heuristic processes. This new methodology will automate the process of pricing, reducing the manual work and making scaling up much more manageable.

In the remainder of this section, we provide a literature review on related research. We briefly introduce the demand forecast model Zalando develops in Sect. 2, which generates the demand as inputs to the optimization framework.

Section 3 includes details on the pricing model for single SKU optimization, while Sect. 4 describes the price optimization model with global steering targets and how to solve it. Details on the implementation of our pricing decision support tool as well as an analysis of field experiments are included in Sect. 5. Section 6 explores an aggregation model that solves the problem by clustering SKUs. Finally, Sect. 7 concludes the paper with a summary of our results and potential future works.

1.2 Literature Review

Promotional pricing is a sales strategy in which brands temporarily reduce the price of a product or service to attract prospects and customers. By lowering the price for a short time, a brand artificially increases the value of a product or service by creating a sense of scarcity. Consumer-goods companies and retailers realize that getting pricing, markdowns, and promotions right across all brands and channels is critical to survive and thrive. In-depth overviews of this literature can be found in Talluri and Van Ryzin (2006) and Özer et al. (2012). Previous research contributes to the development and implementation of pricing decision support tools for retailers. Smith and Achabal (1998) developed clearance prices and inventory management policies. Natter et al. (2007) implemented a decision-support system for dynamic retail pricing and promotion planning. In the meantime, more and more companies have adopted industry software to facilitate their pricing decisions. For instance, software companies like LOKAD and BlueYonder provide software systems as solutions to demand prediction and optimization within the supply chains. However, such products may not be able to tailor to company's needs fully, and in many cases company demands its own system of forecast and optimization. Making the right pricing decision is a critical step for modern companies to succeed and enable smart decision-making taking advantage of the visibility of data (Boute et al., 2020).

Our work is to apply revenue management techniques for optimizing prices in the online fashion retailing environment across different countries under inventory constraints. The problem we study in this paper includes the following distinguishing features that together define our unique position in the revenue management literature.

First, in the online retailing environment, the richness and availability of data enables us to build demand forecasts using historical observations. This is related to the online learning setting for price optimization. There is an increasing stream of research on online learning models which often assumes an unknown linear demand model. Den Boer (2015) provides a comprehensive survey on this topic. Papers in this research stream study the fundamental trade-off between experimenting to improve estimates of the unknown demand model (the exploration) and leveraging current estimates to maximize revenue (the exploitation) (Bu et al., 2020). There are many papers that developed online pricing models under the assumption of a linear demand model (e.g., Keskin & Zeevi 2014; Den Boer 2015; Qiang & Bayati 2016; Den Boer & Keskin 2019). Nambiar et al. (2019) and Ban and

Keskin (2020) assumed a generalized linear demand model. In this paper, Zalando builds a deterministic demand function based on neutral-networks, which makes it powerful to incorporate all the internal and external data to improve prediction; however, it is highly non-linear. This non-linearity makes the optimization model more challenging to deal with.

Second, the pricing process usually contains two key components: demand forecasting and price optimization. An important input for the success of the price optimization model is the predicted demand values. Recent advances in machine learning techniques and richness of data have motivated innovative data-driven approaches to forecast demand and optimize price. For example, Ferreira et al. (2016) study a pricing problem for an online flash fashion retailer, Rue La La. In that paper, they apply random forests to estimate customer demand under different price levels, and propose an efficient optimization algorithm based on mixed-integer programs to make discount decisions. Caro and Gallien (2012) study a clearance pricing problem for fast fashion retailer, Zara. In that work, they build a demand forecast model to address the lack of price-sensitivity data, and then feed the demand forecast to an optimization model to determine price markdowns. Cheung et al. (2017) study a promotion pricing problem for Groupon, a large e-commerce marketplace for daily deals. In that paper, they develop a pricing policy that dynamically learns customer demand using real-time sales data under limit price experimentation. Ma et al. (2018) apply random forest models to predict demands for a CPG company followed by a pricing optimization model.

Third, another interesting feature of the problem is the large scale due to the nature of fashion products, often involving large assortments. An individual product is sold across multiple countries with a shared inventory pool, and the inventory allocation across different countries and different time periods need to be *balanced*. Namely, the company cannot reserve inventory for a specific country or for a specific time period. Furthermore, the pricing decisions of different products are tied together under certain business steering targets across countries and product categories. Therefore, instead of solving the problem by articles, the model should be able to solve multiple products at a much large scale a cross countries within reasonable time. There are very few papers resolving this challenge in the literature.

Finally, this work is also related to the literature investigating the operational challenges in online fashion retailing industry. Apart from the several papers mentioned above, there has been more and more empirical work in the context of fashion retailing. Caro and Martínez-de Albéniz (2015) provide a comprehensive overview of the business models for fast-fashion industry. Boada-Collado and Martínez-de Albéniz (2020) examine the impact of inventory levels on demand in the fashion retailing setting. Fisher et al. (2018) validate the pricing competition model in online retailing setup through field experiments. Our paper combines data driven approaches and optimization modelling methods, which is then validated through offline tests with historical data and real world field experiments.

2 Demand Forecast Model

Before introducing the price optimization framework, we first give a brief overview of the demand forecast model, which generates deterministic demand predictions under all price levels and serve as the inputs to the optimization model. To make the price optimization model succeed, the accuracy of the demand forecaster is critical. There are many commercial promotion software available in the market. However, they either can not meet the high requirement of accuracy or are limited to a certain type of business environment. Zalando has its unique challenges (large scale assortment, long tail of low sales products, operating in multiple countries, long planning horizon, etc.), which makes it important to develop our own operational processes/models.

Zalando uses a collection of forecast models that provide (size aggregated) article-level forecasts on a weekly basis. The forecaster is based on the Transformer architecture Vaswani et al. (2017) with some adjustments to make it suitable for time series forecasting. Transformer is a recently developed machine learning tool based on neural networks. It has been widely used in natural language processing and achieves outstanding performance. In Zalando's context, Transformer takes as input the sales history of every product in all countries and some product-specific feature information (e.g. brand, color, style, product category) to predict the demand for future weeks. The model is retrained every week with new incoming data to predict, taking into the consideration of discount levels, past sales history, countries, weeks, and product specific features mentioned above. It uses weighted least square error as the measure of accuracy, and has been tested to outperform the previous forecast models Zalando was using, including gradient based auto-regressive models, random forests and other types of neural networks. We present the relative performance metrics in Table 1, where ASF4 is the name for the gradient based auto-regressive benchmark model, and LSTM is another type of neural network model. We mask the percentage error values so that ASF4 has the calibrated level of zero, then present the absolute percentage difference in terms of errors for the other two models. The table includes both first week forecast accuracy and aggregate level forecast accuracy for the whole selling season. It is observed that the Transformer model obtains not only the lowest error with huge improvement from the other two models, but also the smallest bias, which is essential for the optimization procedure.

In the business practice, we have the following two observations on the demand. First, since markets are based on different countries, it is unlikely that demand in one country will depend on the prices in other countries. In fact, it is observed that same

Table 1 Performance comparisons between demand forecast models

Models	1st week error	Seasonal error	Bias
ASF4	0 (calibrated)	0 (calibrated)	14%
LSTM	−30%	+9%	12%
Transformer	−30%	−9%	−3%

applies to different time periods and different products. The assumption is that the forecast in one country does not depend on discounts of other countries, nor should the forecast for one week depend on the forecast in other weeks. This assumption allows us to characterize demand independently across countries and time periods in the Transformer architecture. The second observation is that discount levels are not continuous, and usually take discount steps of 5% (e.g., 15%, 50%). This allows us to model the optimization problem into a MIP to select the optimal discount among the discrete discount levels. For this MIP we require the forecaster to tabulate the predicted demand under every possible discount level, for every country and every future week into large tables, which will feed as inputs to the optimization model in the following section.

3 Single SKU Discount Optimization

We start with the price optimization model for each single SKU. Recall that each SKU's discount need to be optimized for a total of 17 countries (denoted as C), and generally a season of 40 weeks (denoted as T). The forecaster in the previous section will generate the predicted demand for each discount level on the price ladder (denoted by L). Given L discrete discount levels, we need to decide the optimal discount for each product over a selling horizon of T weeks across C countries. Throughout the paper, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. Let P_c denote the original (undiscounted) price of the SKU in country $c \in [C]$, and let d_l be the discount value for discount level $l \in [L]$. Specifically, we have a total of $L = 15$ discount levels, ranging from 0 to 70% off, with a step size of 5%, in which $d_1 = 0$ denotes the undiscounted price and $d_L = 0.7$ denotes 70% off the original price.

Let $D_{c,t,l}$ and $R_{c,t,l}$ be the demand and return rate forecast, respectively, for country $c \in [C]$ in week $t \in [T]$ under discount level $l \in [L]$. We assume that demand $D_{c,t,l}$ and return rates $R_{c,t,l}$ are deterministic and provided as inputs to the optimization model. Given the sales in week t, the corresponding returns are distributed in the following weeks according to the return base vector RB. In other words, RB_1 fraction of the sales will be returned in the upcoming week. The RB vector is also assumed to be fixed and provided by Zalando as an input.

Our decisions are binary variables $z_{c,t,l} \in \{0, 1\}$, which indicate the choices of discount level l in country c on week t, and sales variables $x_{c,t,l}$, which characterizes the sales under discount level l in country c on week t. Let y_t be the stock level at the beginning of week t, and y_{end} be the stock leftover after selling horizon T. At the beginning of week t, we have stock replenishment A_t that is predetermined before the whole selling horizon. At the end of selling horizon, the remaining stock y_{end}

has a salvage value of SV per unit. Let $\pi_{c,t,l}$ be the profit of selling one item of the SKU in country c and week t under discount level l.[1]

To maximize the total profit of selling the product, we formulate the single SKU discount optimization problem into a MIP as follows.

$$(P)\ \max_{x,z} \sum_{c,t,l} \pi_{c,t,l}\, x_{c,t,l} + y_{end}\, SV \tag{2}$$

$$s.t.\quad y_t = y_1 - \sum_{c,s<t,l} x_{c,s,l} + \sum_t A_t + \sum_{c,s<t,l} (\sum_{i=1}^{t-s} RB_i) R_{c,s,l} x_{c,s,l} \qquad \forall t = 2, \ldots, T \tag{3}$$

$$\sum_{c,l} x_{c,t,l} \leq y_t \qquad \forall t \tag{4}$$

$$x_{c,t,l} \leq z_{c,t,l} D_{c,t,l} \qquad \forall c, t, l \tag{5}$$

$$\sum_l z_{c,t,l} = 1, \qquad \forall c, t \tag{6}$$

$$z_{c,t,l} \in \{0, 1\}, x_{c,t,l} \geq 0 \qquad \forall c, t, l \tag{7}$$

The objective is to maximize the total profit, both in and after the selling season. Constraint (3) specifies the stock dynamics for each time period, where stock level at the beginning of week t is equal to the initial stock y_1 minus sales, plus replenishment and returns from previous weeks. Constraint (4) requires that the total sales for a specific time period have to be less than or equal to the remaining stock. Constraint (5) limits the sales variable for each c, t, l (country, time and discount level) combination. So that when $z_{c,t,l} = 0$, the sales must also be zero, and when $z_{c,t,l} = 1$, the sales will be less or equal to the forecast demand. Finally, Constraint (6) describes that only one discount level is allowed to be selected in each country and in each week.

A natural question is why do we need to model the extra sales variable $x_{c,t,l}$, given that in reality we do not have control over it. We illustrate this point through a counterexample. Suppose that we do not model the sales variable, the simplified MIP will have the following form:

[1]Specifically, we have

$$\pi_{c,t,l} = \frac{1}{(1+\mathrm{CCR})^{t/52}} \left(\frac{\sum_l P_c(1-d_l)(1-\mathrm{CO})(1-R_{c,t,l})}{1+\mathrm{VAT}_c} \right) - \frac{1}{(1+\mathrm{CCR})^{t/52}} \left(R_{c,t,l}\mathrm{CR}_c - \mathrm{CF}_c \right) \tag{1}$$

where CCR and VAT are constants. CO, CR and CF are the coupon loss, return and fulfillment cost, respectively.

$$\begin{aligned}
\max_{z} \quad & \sum_{c,t,l} z_{c,t,l}\ p_l\ D_{c,t,l} \\
s.t. \quad & \sum_{c,t,l} z_{c,t,l}\ D_{c,t,l} \le Y \\
& \sum_{l} z_{c,t,l} = 1 && \forall c, t \\
& z_{c,t,l} \in \{0, 1\} && \forall c, t, l
\end{aligned}$$

Suppose we have a toy model with a single period, single country, and two price levels $p_1 = 20, p_2 = 10$, so that $C = T = 1, L = 2$. The inventory level is $Y = 120$, and the demand forecaster gives us $D_{l=1} = 50$ and $D_{l=2} = 140$. In other words, if we set the high price of 20, we will have a demand of 50, and the low price of 10 will yield a demand of 140. With the simplified MIP, the optimal solution is to set the price high $p = p_1 = 20$ and the total profit is 1000. However in reality we could set to the low price and only satisfy partial demand, with $p = p_2 = 10$ and a total profit of 1200. From this counterexample we can see the limitation of the simple MIP in terms of the flexibility to capture different sales levels. We can think of this limitation as a result of having discrete price levels, since if prices are continuous, we could always set to the correct price for the demand to just deplete all the stock. As a result, we need to model sales $x_{c,t,l}$ as a decision variable into the model.

3.1 Business Constraints

The price optimization model (P) is a basic model that captures the stock dynamics and establishes the relationship between discount decisions and sales. From the business perspective, it is necessary to set certain limitations on the discounts in this basic model. The motivation could be either to avoid bad customer perceptions and experience, to adjust to specific promotional sales events, or simply from the business requirements. We summarize several types of single SKU business constraints as follows.

Minimum/Maximum Discounts To allow flexibility of discount levels, the discount range for the basic formulation is set to be from 0 to 70%. However in practice, the allowable discount range could be much narrower. Some limitations could come from important brand agreements, associated with brand images, others can come from specific country (market) regulations. For example, we would not expect a newly on-shelf fashion product to have deep discounts. To capture this constraint, the model includes country-week specific minimum/maximum discount bounds to restrict the discount ranges.

Maximum Upward/Downward Steps Intuitively, customers will get upset if they find the price changes drastically within a short amount of time. In addition, significant discount increases may lead to an explosion of returns, which will incur non-trivial costs on the business. In other words, discount differences in consecutive weeks should not be very large. We could upper bound these discount differences by country-week specific maximum upward/downward steps. For example, if one SKU has a discount of 20% in week one, and maximum allowable upward and downward step are both 15%, then the feasible discount range for week two will be from 5 to 35%.

Discount Barriers When discount levels are too marginal (like 5%), customers' perception of the promotion may be compromised. To address this issue, we impose a country-week specific discount barrier, so that the discount values to be either 0% or above these certain discount barriers. For example, by setting a discount barrier of 20%, we disallow discount levels 5%, 10% and 15%.

3.2 *Stock Hedging*

In the previous section, we assume in model (P) that we can optimize over sales via decision variables $x_{c,t,l}$. Although this assumption allows to have a linear formulation of (P), it may cause "stock hedging" problems in the final solution, meaning the product's inventory is reserved for certain countries and certain weeks, which violates the operations in practice. The online shop operates with global stock, which is accessible by all countries, and as it is not typical to reject sales from a specific country, even if it may be profitable to do so. For example, a customer in Germany arrives in week one when there are stocks available, the model (P) may reject her demand by setting $x = 0$, because it is more profitable to sell this unit of inventory in Spain, or in week two. In reality, Zalando does not have this flexibility to turn down customers and "hedge" the stock, so the optimal solution from (P) is often not practical to implement, and we need to integrate further constraints to deal with the stock hedging problem.

Specifically, we go through the sales dynamics in each week as follows, and there will be two possible scenarios. When the inventory is sufficient in week t, i.e., $y_t \geq \sum_{c,l} z_{c,t,l} D_{c,t,l}$, it requires $x_{c,t,l} = z_{c,t,l} D_{c,t,l}$, i.e., sales equal to demand $x_{c,t,l} = D_{c,t,l}$ in each country for the selected price levels l with $z_{c,t,l} = 1$. The other scenario happens when the inventory is insufficient in week t, i.e., $y_t \leq \sum_{c,l} z_{c,t,l} D_{c,t,l}$, as a result not all the demand in week t will be satisfied. Here we assume that the customer arrival process in all countries are evenly distributed across the week, and we will have $x_{c,t,l} = y_t \cdot \left(z_{c,t,l} D_{c,t,l} / \sum_{c,l} z_{c,t,l} D_{c,t,l}\right)$, i.e., sales split proportionally to demand in each country for the selected price levels.

Combining the above two scenarios, the realized sales in practice can be described by the following equation:

$$x_{c,t,l} = z_{c,t,l} D_{c,t,l} \min\{1, \frac{y_t}{\sum_{c,l} z_{c,t,l} D_{c,t,l}}\}. \tag{8}$$

In our model formulation (P), we assume sales are also decision variables, and simply put constraints $x_{c,t,l} \leq z_{c,t,l} D_{c,t,l}$ on the sales decisions $x_{c,t,l}$. This formulation allows an extra degree of freedom to allocate the stock across different week and countries. However, in the presence of insufficient inventory, the optimal solution might be to reserve the inventory for certain countries and certain weeks, as the example above illustrates. If this is the case, we will see in the final solution that $x_{c,t,l} < D_{c,t,l}$ even when y_t can satisfy the demand of all the countries in the corresponding week. This violates the real-world sales pattern in (8), and we refer to this violation as the "stock hedging" problem.

We refer to the requirement in (8) as the sales-balancing conditions, and since restrictions are non-linear, we cannot directly integrate them into our basic MIP formulation (P). To address this problem, We break down the conditions into country-balancing conditions and week-balancing conditions, and then add constraints into our formulation to capture these two conditions separately.

Country Balancing Constraints For the stock hedging problem across countries, the issue occurs when there are unbalanced sales across countries. For example, suppose we only have two markets, and the demand is 100 in Germany and 50 in Spain, we would expect the realized sales to be also 2:1. In other words, if inventory is sufficient (larger than 150), the sales will be equal to the demand. If inventory is insufficient (say 100), the demand will be satisfied proportionally (66 in Germany and 33 in Spain).

We address the stock hedging issues across different countries by adding the following constraint into formulation (P):

$$\sum_{l \in [L]} \frac{x_{c,t,l}}{D_{c,t,l}} = \sum_{l \in [L]} \frac{x_{1,t,l}}{D_{1,t,l}} \text{ for } c \in [C] \setminus \{1\},\ t \in [T]. \tag{9}$$

Intuitively, this constraint requires that sales through all countries should follow the same depleting rate (proportional to their demands). It makes sure that given a set of demands and inventory levels, there's only one way to distribute the stock across countries, which is to distribute proportionally to their demands.

Week Balancing Constraints For the stock hedging problem across weeks, the issue occurs when the model reserves stock for future weeks, when there are still unsatisfied demands in the current week.

We address the stock hedging issues across different weeks by adding a set of big-M constraints. We introduce binary variable ϕ_t for each time period. Specifically, $\phi_t = 1$ denotes sufficient inventory in week t and $\phi_t = 0$ denotes insufficient inventory in week t. We then add the following constraints:

$$\sum_{c,l} (D_{c,t,l} z_{c,t,l} - x_{c,t,l}) \leq M_{demand} \cdot (1 - \phi_t) \qquad \forall t = 1, \ldots, T \tag{10}$$

$$\sum_{c,l} x_{c,t,l} \geq y_t - M_{stock} \cdot \phi_t \qquad \forall t = 1, \ldots, T, \tag{11}$$

where M_{demand} and M_{stock} are large constants that upper-bound the total demand and remaining inventory in each week, respectively. When $\phi_t = 1$, constraint (10) in combination with (5) forces sales to be equal to demand and (11) is relaxed. When $\phi_t = 0$, (10) is relaxed, and (11) in combination with constraint (4) forces sales to be equal to the remaining inventory.

Corollary 1 *Set of constraints* (9)*,* (10) *and* (11) *is identical to desired sales pattern described by* (8) *given the original problem formulation.*

The proof can be found in the Appendix. By adding constraints (9), (10) and (11), we fix the stock hedging problem and capture the observed sales pattern in (8), while still maintaining the linear structure for the optimization model.

3.3 Limited Size Availability

In the context of fashion industry, there is a clear separation between articles on unit and aggregated level: throughout the paper we have been using the notion of SKU for configuration level stock keeping units (e.g. sneakers of a specific brand, including all sizes). In practice, a SKU can be managed in a lower size level (e.g. white T-shirt of a specific brand sized "M") while in principle it is possible to set prices on size levels. The latter is not typically done due to the drastic increase of the problem scale.

We assume in model (P) that each SKU represents the same product with different sizes, and both the demand forecast and the inventory of each SKU are given on the aggregated level of all sizes. In reality, however, certain sizes of the SKU might be unavailable when the stock level is low, and this might raise discrepancies between the observed sales and the demand forecast values. For instance, consider a SKU that contains a white T-shirt of the same style with sizes S, M and L. When stock level is much higher than the demand, it is likely that all sizes are available in the requested quantities. When the stock level is low and close to the demand, certain sizes may not be on stock sufficiently, and thus the corresponding demand cannot be fulfilled entirely. In this case, the given demand forecast values overestimate the true demand, and we need to scale down these values based on the product's stock level.

One way to adjust the demand forecast is to use a "stock response" function, which is defined as a function that maps the product's stock level to a smaller value to approximate inventories in reality. Specifically, given stock level Y_t, the stock response function outputs a multiplier $\mathrm{sr}_t(Y_t)$ that is multiplied to demand forecasts $D_{c,t,l}$. In practice, Zalando has been implementing an exponential stock response function as follow:

$$\mathrm{sr}_t(Y_t) = 1 - \exp\left(-\alpha \cdot (Y_t/\mathcal{N})^{\beta}\right), \tag{12}$$

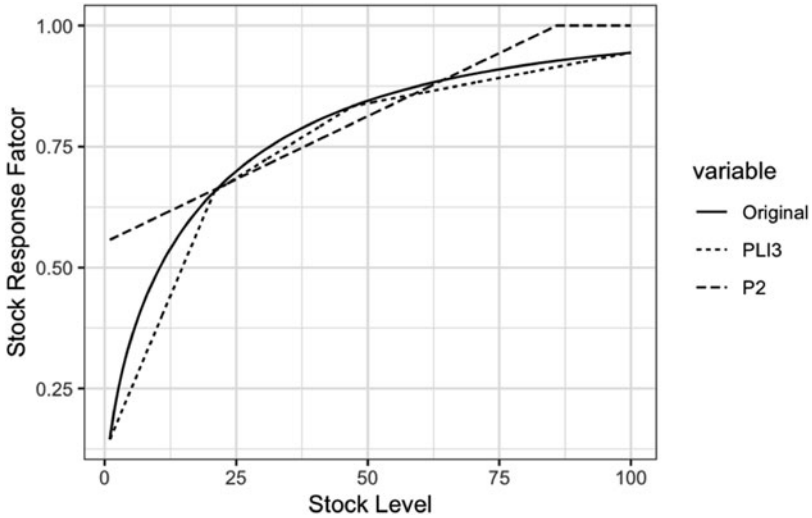

Fig. 2 Stock response function and the piecewise linear approximation

where $\mathcal{N}$ is the cardinality (number of different sizes) of the SKU, and α and β are parameters fine-tuned by fitting the historical data. A graphical illustration is plotted as the solid curve in Fig. 2. The intuition is that when stock level is very high, all sizes are expected to be available, therefore the stock response factor is close to one. On the other hand when stock level is low, it is more likely for certain sizes to become unavailable and the corresponding demand cannot be satisfied, therefore realized demand will reduce by multiplying a factor of the stock response value.

3.3.1 Piecewise Linear Approximation

One challenge of the approach above is that the stock response function is non-linear and will be multiplied by the demand and decision variables in the formulation, which will also become non-linear. To deal with this challenge, we adopt the approximation algorithm in Kontogiorgis (2000) to approximate the stock response function with a piecewise linear function that has K segments. The detailed algorithm is provided in the appendix. Figure 2 shows the approximation Piecewise Linear Interpolate (PLI) result with three segments, compared to the benchmark $P2$ two-piece approximation.

There is a natural trade-off on how many pieces we should select for the approximation. We compare the maximum error, the running time of the model and the objective difference for different approximation schemes. The result is presented in Table 2.

Table 2 Comparison between stock response approximation schemes

Methods	Max error	Running time (seconds)	Objective
P2	0.412	502	9.03%
PLI3	0.117	671	5.35%
PLI4	0.063	710	1.76%
PLI5	0.037	874	1.07%

3.3.2 Formulate Stock Response Factors

We explicitly formulate the stock response factors using the following constraints that describe the factor as a linear combination of the breakpoints of the piecewise linear curve. Let $0 = y^{(1)} \leq \ldots \leq y^{(K)} \leq y^{(K+1)} = \overline{y}_k$ be the breakpoints of the interval $[0, \overline{y}_k]$, and $f^{(1)}, \ldots, f^{(K)}, f^{(K+1)}$ the corresponding function value, i.e., $f^{(i)} = \mathrm{sr}\left(y^{(i)}\right)$ for $i \in [K]$. Given inventory y_t at the beginning of week t, we have

$$y_t = \sum_{i=1}^{K+1} \mu_i y^{(i)} \tag{13}$$

$$\mathrm{sr}_t = \sum_{i=1}^{K+1} \mu_i f^{(i)} \tag{14}$$

$$\sum_{i=1}^{K+1} \mu_i = 1 \tag{15}$$

$$0 \leq \mu_i \leq 1 \mathrm{for} i = 1, \ldots, K+1 \tag{16}$$

In addition, (13) and (14) should be a linear combination of **two consecutive breakpoints**. Let E_i with $i = 1, \ldots, K$ be the binary variables that indicate whether or not $y^{(i)}$ is selected as the left breakpoint. We then have constraints

$$\sum_{i=1}^{K} E_i = 1 \tag{17}$$

$$E_i \in \{0, 1\} \mathrm{for} i = 1, \ldots, K \tag{18}$$

$$\mu_1 \leq E_1 \tag{19}$$

$$\mu_i \leq E_{i-1} + E_i \mathrm{for} i = 2, \ldots, K \tag{20}$$

$$\mu_{K+1} \leq E_k. \tag{21}$$

3.4 Integrating Stock Hedging and Stock Response

In the previous two subsections, we integrate extra constraints to deal with stock hedging and limited size availability issues, respectively. Unfortunately, the stock hedging behaviour is impacted by the stock response constraint. To see this, it is assumed in stock hedging that either the stock is depleted or the demand is fulfilled. However, with the stock response factor, the condition no longer holds. In fact it should be updated that either the stock is depleted or the demand modified by the stock response factor is fulfilled.

Due to the interference between stock hedging and limited size availability, we need to introduce an extra set of variables and constraints. Let $D'_{c,t,l}$ denote the demand forecast after the modification of the corresponding stock response factors, and we then have the following constraints that describe the modified demand forecast:

$$D'_{c,t,l} \leq \mathrm{sr}_t \cdot D_{c,t,l} \qquad \forall c, t, l \tag{22}$$

$$D'_{c,t,l} \leq z_{c,t,l} \cdot D_{c,t,l} \qquad \forall c, t, l \tag{23}$$

$$D'_{c,t,l} \geq \mathrm{sr}_t \cdot D_{c,t,l} - (1 - z_{c,t,l}) \cdot M_{\mathrm{sr}} \cdot D_{c,t,l} \qquad \forall c, t, l \tag{24}$$

Given $D'_{c,t,l}$ the modified demand forecast, we need to update constraint (10) to

$$\sum_{c,l} D'_{c,t,l} - x_{c,t,l} \leq M_{demand} \cdot (1 - \phi_t), \quad \forall\, t \in [T] \tag{25}$$

and additionally, constraint (5) to

$$x_{c,t,l} \leq D'_{c,t,l}, \quad \forall\, c, t, l. \tag{26}$$

With the constraints above included, we now have the full-scale single SKU optimization problem that captures all relevant business constraints. In summary, the constraints for the ingle SKU MIP are (3)–(7), (9), (13)–(21), and (22)–(26). The benefit of this linear formulation is the computitional time. In practice, solving each single SKU MIP with 17 countries, 40 weeks, and 15 price levels using commercial solver will take several seconds. We can speed up the process by utilizing parallel computing power, since by far each SKU is optimized individually. There are other approaches to speed up the single SKU optimization, and we will introduce an effective one in the next subsection.

3.5 Telescopic Method

For single SKU optimization, we can aggregate on the time scale by using a telescopic method to reduce the computational time. The telescopic method is motivated by several observations in single SKU level optimization. First, the optimization model is solved only to obtain optimal discounts for the first week, and will be re-solved with new data inputs every week. As a result, the most relevant decisions will be the first week discounts. Second, it is observed that demand and return forecasts for later weeks are less accurate, and intuitively we should "care less" about later weeks. Last, computational burdens are huge concerns in practice. In general, the optimization model needs to include the full season that SKUs are offered to customers, and could last for an entire year. In other words, the number of time periods can be as large as $T = 52$ weeks. Although one can resort to linear programming (LP) relaxations for faster computations, with extra complexity of the formulation introduced in Sects. 3.2 and 3.3, LP relaxations will be less reliable, and the computation time for solving MIP will explode as the number of weeks grow. In contrast, telescopic methods "combines" later weeks into fewer optimization periods, which will significantly speed up the optimization routine.

Take an example where the optimization planning horizon is $T = 40$. If we introduce extra constraints $z_{c,t,l} = z_{c,20,l}$ for all $t \geq 20$, that is, forcing all the discounts after week 20 to be the same, then the new optimal solution will be suboptimal for the original problem and will yield a lower objective. Nevertheless, the optimal solution for the first week will not be far from the true optimal solution. Hence we can implement such an approximation by combining all weeks after week 20 into one period.

Denote $\tau \in [t_1, \ldots, t_{\hat{T}}]$ to be the starting weeks for each period, and we have $\hat{T}$ total telescopic periods, each with n_τ weeks. For example, $[1, 2, 7]$ means that we combine weeks 2–6 into one period and all weeks after (weeks 7–40) into another period. We denote the hatted version to be the telescopic updated version, and update the model inputs to their aggregated version:

$$\hat{D}_{c,\tau,l} = \sum_{i=t_\tau}^{t_{\tau+1}-1} D_{c,i,l} \qquad \hat{A}_\tau = \sum_{i=t_\tau}^{t_{\tau+1}-1} A_i \tag{27}$$

$$\hat{R}_{c,\tau,l} = \frac{1}{n_\tau} \sum_{i=t_\tau}^{t_{\tau+1}-1} R_{c,i,l} \qquad \hat{\pi}_{c,\tau,l} = \frac{1}{n_\tau} \sum_{i=t_\tau}^{t_{\tau+1}-1} \pi_{c,i,l} \tag{28}$$

Intuitively, we approximate demand, replenishment, and unit profit of a telescopic period by summing up the values of each individual week. For returns, we compute the average return rates of each individual week since these are fractional values. In addition, we approximate the business constraint inputs, including max/min discount levels, max upward/downward steps and discount barriers, by taking the average across the corresponding weeks. The telescopic methods has proven to be an effective approximation heuristic and has been adopted by Zalando to speed up the weekly optimization routine.

4 Global Optimization

In previous sections, we formulate the discount optimization problem for single SKUs and address the potential issues caused by the linear MIP formulation. If in practice each SKU behave independently, we could separately optimize each SKU to obtain the optimal discounts for all the SKUs. However, there are "global" business constraints that tie different SKUs together. We introduce two kinds of global constraints in this paper. The first kind is from a business planning perspective, where each country has its own growth plans. For instance, the company may want the total sales for all the SKUs within a product category (e.g., women footwear) to reach a specific target. The second kind of business constraint comes from the cost perspective, where controlling the average discounts over a group of SKUs is critical to evaluate the cost spent on the campaigns. For example, the company may want to control the weighted average discount for a group of SKUs within a range (e.g.,15–17%) This value is denoted as sales Discount Rate (sDR) target. The way Zalando adopts to measure weighted average discount is to use discount rates weighted by their potential contributions to the total revenue (undiscounted price times sales), as defined below:

$$sDR_i = \frac{\sum_{(k,c,t)\in\mathcal{T}_i} \sum_l d_l \; x_{k,c,t,l} P_{k,c}}{\sum_{(k,c,t)\in\mathcal{T}_i} \sum_l x_{k,c,t,l} P_{k,c}}$$

sDR is critical from a business planning perspective, as it measures the relative cost of adopting a discount strategy. The higher the sDR value, the more Zalando needs to invest (or bear as opportunity costs) to the discount plan. In reality, sDR targets are closely monitored by Zalando high level business teams, and they usually impose certain values as targets to reach. For example, it might be required that for women footwear category, the weighted average discount (sDR) for Germany in the next four weeks is around 15%.

We discuss in this section the global problem in which we jointly optimize a group of SKUs such that the above two types of constraints are satisfied in certain countries and weeks. We follow the notation in Sect. 3 and add subscript k to denote the associated variables of SKU k. Define target set $\mathcal{T}$ to be the set of SKUs, with specific sets of countries and time periods that we want to reach a certain global target. Given target set $\mathcal{T}_i := \{(k, c, t) \mid k \in \mathcal{K}_i, c \in \mathcal{C}_i, t \in \mathcal{W}_i\}$, we have two types of steering targets: revenue target and sDR target.

- Revenue targets.

$$\sum_{(k,c,t)\in\mathcal{T}_i}\sum_{l} P_{k,c,t,l}x_{k,c,t,l} \geq GMV_i^- \tag{29}$$

- sDR targets

$$sDR_i^- \leq \frac{\sum_{(k,c,t)\in\mathcal{T}_i}\sum_l d_l\ x_{k,c,t,l}P_{k,c}}{\sum_{(k,c,t)\in\mathcal{T}_i}\sum_l x_{k,c,t,l}P_{k,c}} \leq sDR_i^+ \tag{30}$$

4.1 Lagrangian Relaxation

The solutions of different SKUs are coupled via the global steering targets, and directly solving the global optimization problem for hundreds of thousands of SKUs could be an impossible task. Moreover, parallel computing power cannot be utilized if we jointly optimize all the SKUs. We therefore decompose the global target constraints by using Lagrangian relaxation, namely, by introducing dual multipliers, each associated with a global target.

Let $\mathcal{P}_k$ be the feasible region of sub-problem k, which specifies the values of x_k and z_k subject to the constraints for a single SKU optimization problem ((3)–(7), (9), (13)–(21), and (22)–(26)) for each single SKU k. We can then write the global problem as the following.

$$\max \sum_{k,c,t,l} \pi_{k,c,t,l}x_{k,c,t,l} + y_{k,end}\mathrm{SV}_k \tag{31}$$

$$s.t. \quad \sum_{(k,c,t)\in\mathcal{T}_i}\sum_{l} P_{k,c,t,l}x_{k,c,t,l} \geq \mathrm{GMV}_i^- \quad \forall i \in I \tag{32}$$

$$\sum_{(k,c,t)\in\mathcal{T}_i}\sum_{l}(d_l - \mathrm{sDR}_i^-)\,x_{k,c,t,l} \geq 0 \quad \forall i \in I \tag{33}$$

$$\sum_{(k,c,t)\in\mathcal{T}_i}\sum_{l}(d_l - \mathrm{sDR}_i^+)\,x_{k,c,t,l} \leq 0 \quad \forall i \in I \tag{34}$$

$$x_k, z_k \in \mathcal{P}_k \tag{35}$$

Let λ^-, μ'^-, μ'^+ be the dual non-negative vectors associated with constraint (32), (33) and (34), respectively. Let $\theta = (\lambda, \mu^-, \mu_i^+)$. We then obtain the Lagrangian dual problem:

$$\min_{\theta\geq\mathbf{0}} g(\theta) := \max \sum_k\sum_{c,t,l}\Bigg(\pi_{k,c,t,l} + \sum_{\mathcal{T}_i}\lambda_i^- P_{k,c,t,l} + \sum_{\mathcal{T}_i}\mu_i^-(d_l - \mathrm{sDR}_i^-) \tag{36}$$

$$- \sum_{\mathcal{T}_i} \mu_i^+ (d_l - \mathrm{sDR}_i^+) \Big) x_{k,c,t,l} - \sum_{\mathcal{T}_i} \lambda_i \mathrm{GMV}_i^- + y_{k,end} \mathrm{SV}_k \tag{37}$$

$$s.t.\;\; x_k, z_k \in \mathcal{P}_k \tag{38}$$

where the Lagrangian dual function g is piece-wise linear, convex, continuous and non-smooth.

In general, Lagrangian multipliers θ are required to be non-negative to penalize target violations in the objective. In practice, the corresponding constraints could be infeasible, where the multipliers will go to infinity, or heavily violated in a given solution. In this case, the multipliers could have very large values, possibly orders of magnitude larger than the rest of the objective function, causing slow convergence or numerical instability. To guarantee that the problem of optimizing $g(\theta)$ is both *bounded* and *numerically stable*, we assume each Lagrangian multiplier θ_i to be bounded from above by a reference value $\bar{\theta}_i$ which can be easily computed from the data as follows. Consider a generic formulation for the ith target

$$\sum_{j \in J} \tau_{ji} x_j \le T_i$$

with primal variables x_j, $j \in J$, and multiplier θ_i. Let $f(\mathbf{x}) = \sum_{j \in J} c_j x_j$ be the objective function to be maximized. Dualizing the ith target constraint yields the modified objective function

$$\min_{\theta_i} \left(\max_{\mathbf{x}} \sum_{j \in J} c_j - \theta_i \tau_{ji}) x_j \right) - \theta_i T_i$$

Then, the reference value for the multiplier has to satisfy the condition

$$\bar{\theta}_i > \bar{\theta}_i^{\min} = \max_{j \in J} \frac{c_j}{\tau_{ji}}$$

The condition guarantees that if target i is violated then the multiplier λ_i can become large enough to dominate the coefficients of the primal variables $\mathbf{x}$, i.e. $\max_{\theta_i \in [0, \bar{\theta}_i]} \theta_i \tau_{ji} > c_j \; \forall j \in J$ and thus steer the optimization towards target-reaching solutions. From a business perspective, the multiplier θ_i represents the per-sold-item cost of violating target j by an additional unit, hence the condition states that the maximum allowed cost per sale for an additional unit of violation $\bar{\theta}_i$ for any article $j \in J$ needs to be greater than the per-unit profit c_j for the same sale to guarantee convergence to a target-reaching solution. In our experiments we set $\bar{\theta}_i = 3\bar{\theta}_i^{\min} \; \forall i \in I$.

4.2 Cutting Plane Algorithm

We use the Cutting Plane (CP) method (Kelley, 1960) to solve (36). The method iteratively constructs a piece-wise linear approximation of the Lagrangian dual function g and minimizes its value, yielding a new dual vector of multipliers at each iteration. Specifically, given dual variables θ_0^n and $y_0 = g(\theta_0^n)$ at iteration n, we add the constraint or "(optimality) cut" $y \geq g'(\theta_0)(\theta - \theta_0) + y_0$ to the CP model:

$$\left(\mathcal{M}^n\right)\ z^n := \min\ \ y \tag{39}$$

$$s.t.\ \ y \geq g'(\theta_0^i)(\theta - \theta_0^i) + y_0^i \text{ for } i = 1, \ldots, n \tag{40}$$

$$\mathbf{0} \leq \theta \leq \bar{\boldsymbol{\theta}} \tag{41}$$

We chose the CP algorithm because it can provide good solution quality and rapid convergence while being relatively easy to implement in the company's environment and requiring little overhead compared to the solution of the Lagrangian subproblems. See Frangioni et al. (2015) for a more in-detail evaluation of optimization methods for Lagrangian dual functions.

In the following we report relevant properties about the CP algorithm. First, as g is convex, we have that the series $\{z_n\}_{n \in N}$ is non-decreasing, i.e. $z_n \geq z_{n-1}\ \forall n \in N$. Let S^* be the optimal value for the original problem and θ^* the optimal solution for the Lagrangian dual problem. From strong-duality in convex problems, we will have $S^* = g(\theta^*)$, and $z^n \leq g(\theta^*) \leq g(\theta^n)$ for all iterations n, with the values converging to the optimum $g(\theta^*)$ in a finite number of steps. We can then define the optimality gap at iteration n as $\text{gap}_n = \dfrac{g(\theta^n) - z^n}{|z^n|}$. However, the MIP integer constraints introduce non-convexity to our problem, and we have instead the weak duality $S^* \leq g(\theta^*)$, and z^n converges to an upper bound of the optimal value of the original problem, i.e. $\exists\ \bar{n} \in N : z^n \geq S^n\ \forall n \geq \bar{n}$. Indeed, z^n converges to the value of the continuous relaxation of the equivalent Dantzig-Wolfe reformulation of the Lagrangian relaxation (Desrosiers & Lübbecke, 2005).

We report the algorithm scheme in Fig. 1. We initialize the algorithm with the "dummy solution" $\theta = \mathbf{0}$, which yields the "unconstrained optimum", or the profit-maximal solution where targets are ignored. Other initialization schemes are possible and could yield better results or faster convergence depending on the problem and the underlying data. We refer to the literature Frangioni et al. (2015) for further information. In our experiments, the zero-multipliers solutions are easier to use, as they do not require any prior knowledge or computation, and provide a benchmark of the optimization model without any global constraints.

As a stopping criterion we could use a MIP optimality gap by computing at each iteration n a lower bound for the original problem using some heuristic, possibly based on the current Lagrangian solution $(\theta_n, \mathbf{x}_n)$, and comparing it with the current best dual bound $\max_{n' \leq n} g(\theta'_n)$. Given the scale of our problem, computing accurate heuristic solutions during the iterations would be significantly expensive. We then consider the following stopping conditions: (1) the number of iterations has reached an upper limit N and (2) the change between multipliers in subsequent iterations is below a minimum threshold ϵ.

4.3 Primal Heuristics

While optimizing the dual function leads to find primal solutions with small violations and good objectives, the process does not guarantee to find primal solutions that are either optimal or feasible for the original problem. To tackle this challenge, we develop primal heuristics to construct a good global solution using all the results in the previous iterations.

Let $s_n = (p_n, v_{n1}, v_{n2}, \ldots, v_{n|J|})$ denote the result in iteration n where p_n is the profit objective and v_{nj} is the violation in global target j. Let $\bar{p} = \max_{n \in N}\{p_n\}$ be the highest profit in all iterations. Let σ_j be the target value of global target j. In addition, we record the profit and violations of each SKU k. Let p_{kn} be the profit from SKU k in iteration n and l_{knj} violation from SKU k for target j in iteration n. We then solve the following problem to obtain an optimal combination of the solutions.

$$\min \sum_{j \in |J|} \delta_j + \delta_p \tag{42}$$

$$s.t.\ \ \delta_p = 1 - \frac{\sum_{k \in K} \sum_{n \in N} r_{kn} p_{kn}}{\bar{p}} \tag{43}$$

$$\delta_j = \frac{\sum_{k \in \mathcal{T}_j} \sum_{n \in N} r_{kn} l_{knj}}{\sigma_j} \quad \forall j \in J \tag{44}$$

$$\sum_{n \in N} r_{kn} = 1 \tag{45}$$

Algorithm 1: Cutting plane

1: **Initialization.** Optimization model $\mathcal{M}$; set $n = 1$, $z^0 = -\infty$; set dual values $\theta^0 = 0$
2: **for** $n = 1, \ldots, N$ **do**
3: solve the Lagrangian relaxation problem to obtain $g(\theta^n)$;
4: calculate subgradient $g'(\theta^n)$;
5: **if** $n = N$ or $\|\theta^n - \theta^{n-1}\| > \epsilon$ **then**
6: add optimality cut to model $\mathcal{M}^{n-1}$ yielding model $\mathcal{M}^n$
7: solve model $\mathcal{M}^n$ to obtain dual value z^{n+1} and dual solution θ^{n+1}
8: **else**
9: break;
10: **end if**
11: **end for**

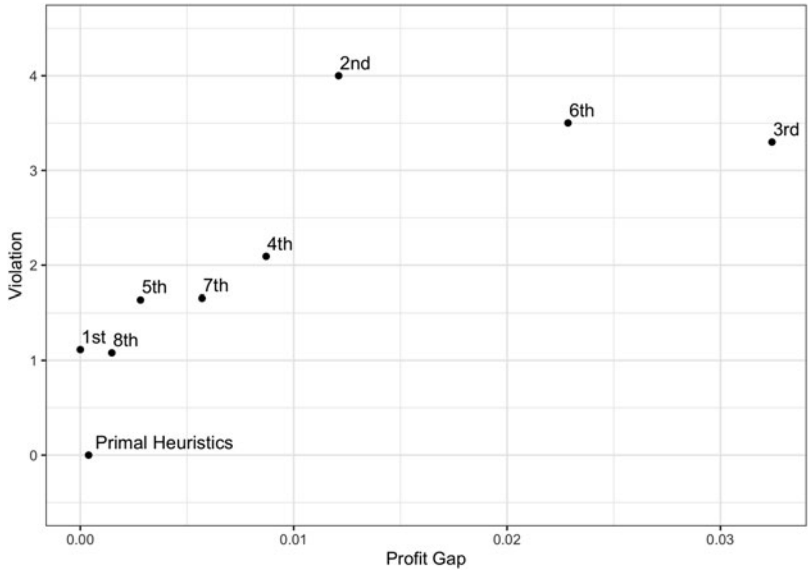

Fig. 3 Profit gaps and violations across iterations and primal heuristic

$$r_{kn} \in \{0, 1\}; \;\; \delta_j, \delta_p \geq 0 \tag{46}$$

Figure 3 showcases one run of a global optimization problem with 1000 SKUs and four sDR targets. It took eight iterations to converge within the optimality gap. The x-axis plots the percentage profit gap compared to the highest profit the model has ever seen. The y-axis measures the total amount of violations of the four sDR constraints. It is clear that when the algorithm stops at the eighth iteration, the primal solution is not particularly preferable. However, after computing a new solution with the Primal Heuristic, the violations are reduced to zero without compromising the profit objective very much.

5 Field Experiments

We have proposed a variety of techniques in previous sections to model, solve and improve the discount optimization process. Zalando is very collaborative and eager to conduct both offline and online experiments to validate the proposed framework. In this section we report experiments performed at Zalando. These are the first stage of the experiments, with the main purpose to check the functionality of the system, as well as the ability to reach the global targets. In later field experiments the main goal will shift towards measuring the improvement in total profits compared to the old system.

5.1 Offline Large-Scale Experiments

Before carrying out real field experiments, we first test the optimization framework offline by using historical data. From an experimental design perspective, we selected a sample of the assortment of the Kids category because it includes articles of all types, from shoes to accessories, and therefore its variety is comparable to the one of the whole shop. Other categories are split both by gender or article type (Men Textile, Women Shoes, Accessories). In this sense, a sample from Kids category is likely to be more representative for the whole shop than a sample of the same size from another category. The optimization spans 14 countries and 26 weeks. The goal is to generate profit-optimal target-reaching discounts to be uploaded for the first week of the optimization horizon. In the experiments, the problem has one global sDR target for the first week. To speed up the process and also facilitate the weekly optimization routine, the implementation exploits large-scale cloud-based parallelization to solve Lagrangian subproblems in parallel. We use Amazon EMR as the execution platform, which allows to provision Hadoop clusters with a specified amount of total cores and memory. Our implementation is written in Python, using the Apache Spark framework. In our experiments, we used the C5 instance type provided by AWS.

We present here the results of two large-scale tests we run on our algorithm. The goal of these tests is to mimic the settings of the following field experiments, and the experiments started on different weeks. Run time measured as total time for the overall system to confirm the completion of each run after launch. It can be seen that our algorithm manages to reach the targets in both cases (Table 3).

5.2 Online Field Experiment Results

The online experiments were launched in consecutive weeks, where the model was solved using 2000 parallel processing cores and 4.0 TiB for each experiment. Bold values indicate when the model and the actual values in the target group satisfied the targets. It can be observed that the target (sDR) value for both the model and the test articles are close or within the target bounds most of the time. For the test group, the actual sDR is much closer to the target than the control one, especially so for experiment #3, proving the practical effect of the decisions taken by the model. More precisely, for test #3 the users later confirmed the targets were discovered to

Table 3 Offline experimental results

Experiments	SKUs	Iterations	Run time (s)	Number of cores	Total memory	Target deviation
1	51,745	5	105	2000	4.0	0.00
2	12,798	5	90	1500	3.0	0.00

Table 4 Online field experimental results

Experiments	SKUs	Iterations	Run time	Target deviation		
					Actuals	
				Model	Test	Control
1	12,632	4	1 h	−0.09%	**0.00%**	−3.27%
2	12,757	1	40 m	**0.00%**	**0.00%**	0.00%
3	8961	4	1 h	−0.14%	−2.85%	−10.43%

be much harder than expected with regard to the assortment, meaning our model managed to significantly move the sDR for the test group despite the targets being difficult to reach. To compare profit we require the Treatment and Control groups to be subjected to the same (sDR) targets, as they have a direct impact on profit (Table 4).

In practice, during the field test the Control group is managed directly by the commercial team using different heuristic processes that can yield different sDRs than the one for the Treatment group. Indeed, we observed large deviations in the Control group from the sDR target that can largely be attributed to this heuristic process and other process-related issues. For this reason, we cannot directly compare profit among the two groups.

6 Aggregation Model

In reality, the company is dealing with a global problem with potentially $K > 600,000$ SKUs, $C = 17$ countries, $T = 40$ weeks, and $L = 15$ discount levels. Moreover, the global optimization in previous sections may take several iterations to terminate. Practically it imposes heavy burdens on computational resources, and demands simplification or certain levels of aggregation. For the global problem, we propose an aggregation framework to cluster similar SKUs within each category into dummy SKUs. This framework brings value from several perspectives. Firstly it could significantly reduce the number of SKUs in the global problem, saving much computational time and resources. Secondly, there is a large fraction of “long tail” SKUs which are less popular and have few or no historical sales. Since these SKUs are expected to have lower demand prediction accuracy and lower sales, therefore are less important to the business operations. By clustering, we combine these SKUs and make centralized price decisions. Finally, aggregation results are also helpful to business planning. Aggregated outputs provide quick suggestions to the business users on their high level promotional planning for certain countries or product categories.

6.1 Clustering

On the global optimization level, the company has country specific and category specific targets. For instance, Fig. 4 illustrates an example of a global optimization with 14 country targets (as indicated for each column) and four category targets (as indicated for each row). The values in individual cells are category-country specific targets and are unknown before solving the global optimization problem. The idea of aggregation model is to approximate the global problem by clustering SKUs within each category. The resulting model is much smaller due to clustering, whose outputs approximate the cell category-country targets. Finally we could decompose the global problem into category specific problems and solve for each SKU's optimal discounts by using cell specific targets.

The first part of the aggregation model is to find a clustering algorithm. Ideally the SKUs within each cluster should have similar optimal discounts since we are combining their information. After exploring the features that impact optimal discount decisions, we include demand, price, inventory, and unit profit as features to perform clustering. After removing the outliers (SKUs with extremely low inventories or demands), we normalize the values on each dimension and apply K-Means clustering algorithm for each category.

6.2 Aggregation Approximation

After we form clusters in each product category, we need to apply an aggregation method to represent SKUs within each cluster. We consider two approaches. The first approach is to aggregate the SKUs in each cluster by taking the N closest SKUs to the cluster's center and scaling up demand, inventory, and other parameters accordingly. This approach is simple to implement; however, it only utilizes the information of a small number of SKUs, and will not yield a good approximation

	Country 1	Country 2	...	Country 14	
Category 1	20%	30%		40%	→ Category Target 1
Category 2	15%	25%		35%	→ Category Target 2
Category 3	20%	30%		40%	→ Category Target 3
Category 4	25%	35%		45%	→ Category Target 4
	↓ Country Target 1 20%	↓ Country Target 2 30%		↓ Country Target 14 40%	

Fig. 4 Example of country and category global targets

Table 5 Comparison of sDR targets

	Country 1			Country 2		
	Benchmark	Aggregation	Difference	Benchmark	Aggregation	Difference
Women textile	0.138	0.129	−0.009	0.161	0.171	0.01
Women footwear	0.148	0.177	0.029	0.183	0.181	−0.002

if the cluster is spread out. The second approach is to aggregate the SKUs in each cluster into a huge dummy SKU. The challenge of this method is how to combine the inputs from different SKUs. In practice, we adopt the second approach and take the demand-weighted average of price, inventory, and other parameters.

6.3 Experimental Results

The data set contains 1000 SKUs from 11 categories across 14 countries and 40 weeks of planning horizon. The business users impose four global sDR targets, two on individual product categories (Women Textile and Women Footwear) and two on individual countries.

We test the aggregation framework by comparing to the benchmark instance, where we apply the dis-aggregated method and jointly optimize SKUs from all the categories. In both instances we implement methods for global optimization in Sect. 4, namely Lagrangian relaxations and primal heuristics on the MIP. We compare the running time and sDR targets for the two instances (Table 5).

After solving the first stage of the aggregated model, value has already been created in terms of category level business insights. If needed or further required by the business users, we could solve for each category on the dis-aggregated model, using sDR inputs from the aggregated model. The optimal objective difference for both product categories is less than 0.5%. We present the discount distribution histogram in Figs. 5 and 6.

7 Conclusion and Future Directions

Zalando has successfully implemented all the algorithms described in this work into their weekly routine of price discount recommendation system. The optimization model outputs optimal discount decisions for hundreds of thousands of products across more than 14 markets on a weekly basis. As the preliminary field experiments suggest, the model is capable of steering the discounts to satisfy the global targets. In conclusion, our work addresses several key challenges for the online fashion retailer, Zalando, with a very large number of products and different levels of business

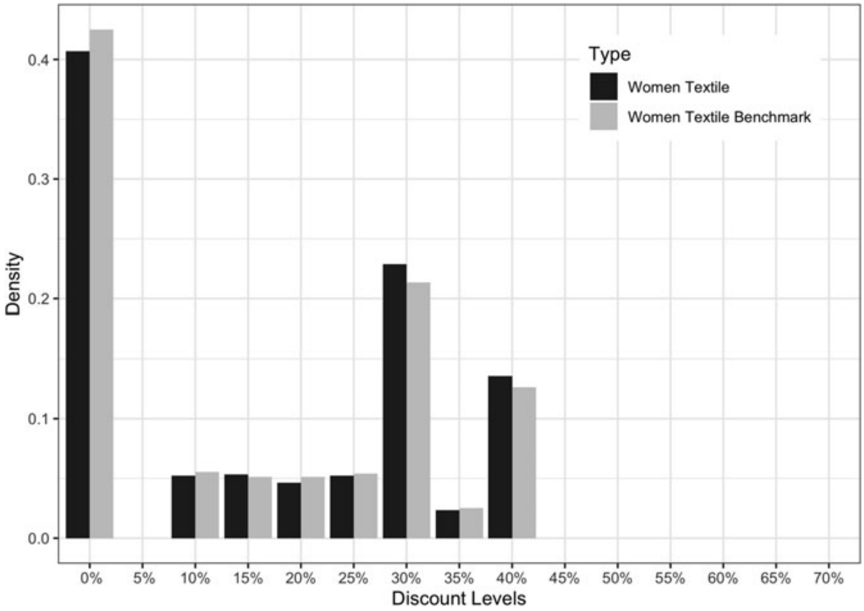

Fig. 5 Discount distribution comparison for women textile category

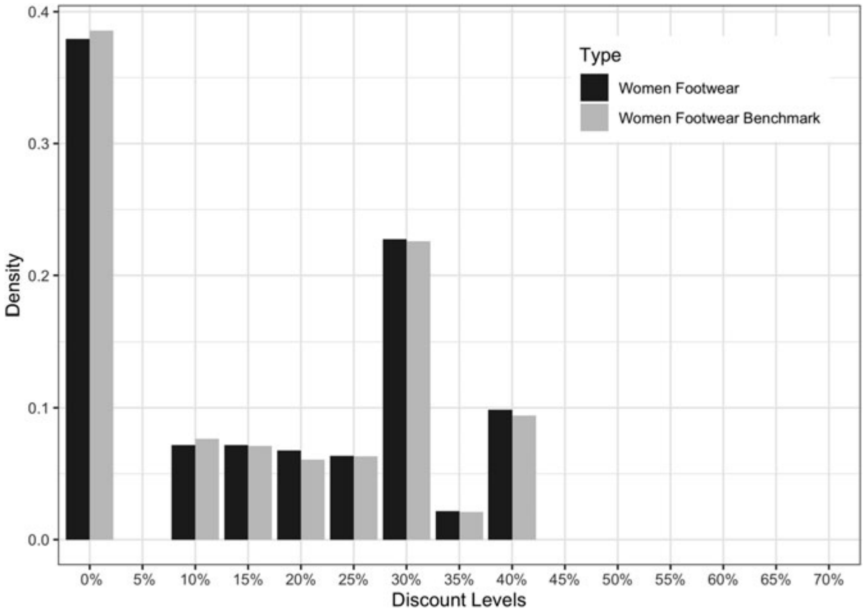

Fig. 6 Discount distribution comparison for women footwear category

constraints. We first manage to model the single product optimization problem as an MIP instance, which correctly captures the challenge of stock hedging and stock response. On the "global" level with constraints across multiple products, we apply the Lagrangian decomposition together with the cutting plane method to efficiently find the solution within optimality gap for this large scale optimization problem. We also adopt a heuristic to combine solutions across iterations to yield a better solution. Finally we propose an aggregation framework that not only will drastically improve the computational time, but also provide high level business insights. The pilot field experiment empirically validates that the optimization framework successfully steers the discounts towards the business targets, and the model will be integrated into the company's weekly operation pipeline. We provide a new system of algorithms that automate the decision making process for a globally operating fashion platform. This system can be applied to many other similar business environments in the future. To show more convincing results, we plan to test the algorithms on the company platform and investigate the profit improvement under this new decision making system. Another important future direction is how to reduce the scale of the problem or increase the efficiency of our system so we can achieve better speed without losing much optimality.

Appendix

Return Forecaster

An essential part of the Zalando business is the flexible return policy: for most of the countries the customers enjoy a 100-day free return after the purchase. Returned articles (e.g. if they do not fit), if they pass the quality control, are coming back in the stock and can be sold later. Such a policy obviously has an impact on the pricing of the articles and plays an important part in Zalando's pricing system. In order to keep track to stock, we also need to forecast returns. The return model we use consists of two components. The so-called return rate model predicts the probability that a given cSKU is returned, and a return times model that predicts when a given return is expected to arrive. Assuming all returns happen within a six-week window, we model return arrival by a six-dimensional vector where the i_{th} entry corresponds to the probability that a given article arrives in the $i - 1_{th}$ week after sale. The return rate models the probability of a return at a cSKU-country level. It is a decision-based model that uses two sources of information gathered over the last 52 weeks:

1. cSKU-specific information: once we observed a sufficient number of past sales, we use (observed) past returns to estimate return rates.
2. Fallback: if we do not have sufficient past sales, we use the return rate of all articles within the same article type and brand.

Proof of Corollary 1

Let us first consider a specific time slot t' and prove the corollary for the case, when inventory is sufficient to satisfy all demand, i.e. $y_{t'} \geq \sum_{c,l} z_{c,t',l} D_{c,t',l}$.

In this case in (8) sales become equal to demand (if corresponding discount decision is activated):

$$x_{c,t',l} = z_{c,t',l} D_{c,t',l} \quad \forall c \in [C]. \tag{47}$$

Which implies (assuming non-negative demand) that $\frac{x_{c,t',l}}{D_{c,t',l}} = z_{c,t',l}$, i.e. ratio between sales and demand is equal to 1 if the corresponding discount variable is chosen and zero otherwise. Let us assume without loss of generality, that such discount level is l_c for each country, i.e.

$$z_{c,t',l_c} = 1 = \frac{x_{c,t',l_c}}{D_{c,t',l_c}}. \tag{48}$$

At the same time, constraints (10)–(11) are forcing $\phi_{t'} = 1$:

$$\sum_c D_{c,t',l_c} = \sum_c x_{c,t',l_c} \tag{49}$$

$$\sum_c x_{c,t',l_c} \geq y_{t'} - M_{stock}, \tag{50}$$

and given that sales can never exceed demand ($D_{c,t,l} \geq x_{c,t,l}$), lead to the fact that $D_{c,t',l_c} = x_{c,t',l_c} \forall c \in [C]$.

Let us a take a look on what happens with (9). For every constraint, only one summand on each side of it is positive, since only one and only one $x_{c,t',l_c} > 0$ (one discount can be chosen per country). Thus the (9) becomes:

$$\frac{x_{c,t',l_c}}{D_{c,t',l_c}} = \frac{x_{1,t,l_c}}{D_{1,t,l_c}} \text{ for } c \in [C] \setminus \{1\},\ t \in [T], \tag{51}$$

which is true when $D_{c,t',l_c} = x_{c,t',l_c}$.

Let us now consider the case stock scarcity, i.e. when there is not enough stock to satisfy all demand ($y_{t'} \leq \sum_{c,l} z_{c,t',l} D_{c,t',l}$). In this case (8) becomes:

$$x_{c,t',l} = z_{c,t',l} D_{c,t',l} \frac{y_{t'}}{\sum_{c,l} z_{c,t',l} D_{c,t',l}}. \tag{52}$$

Let us again assume that for simplicity that l_c discount level is chosen for each country c:

$$x_{c,t',l_c} = D_{c,t',l_c} \frac{y_{t'}}{\sum_{c'} D_{c',t',l_{c'}}} \tag{53}$$

$$x_{c,t',l} = 0 \quad \forall l \in [L] \setminus \{l_c\}, \tag{54}$$

which is identical (given assuming demand values) to:

$$\frac{x_{c,t',l_c}}{D_{c,t',l_c}} = \frac{y_{t'}}{\sum_{c'} D_{c',t',l_{c'}}}, \tag{55}$$

from which we can also deduce that $\sum_c x_{c,t',l_c} = y_{t'}$.

For constraints (9)–(11) we have:

$$\frac{x_{c,t',l_c}}{D_{c,t',l_c}} = \frac{x_{1,t',l_c}}{D_{1,t',l_c}} \qquad \text{for } c \in [C] \setminus \{1\}, \tag{56}$$

$$\sum_c D_{c,t',l_c} - x_{c,t',l_c} \leq M_{demand} \tag{57}$$

$$\sum_c x_{c,t,l_c} = y_{t'}, \tag{58}$$

where the last one comes from the fact that sales cannot exceed stock (4). The latter set of equations is identical to:

$$\frac{x_{c,t',l_c}}{D_{c,t',l_c}} = \frac{x_{1,t',l_c}}{D_{1,t',l_c}} \qquad \text{for } c \in [C] \setminus \{1\}, \tag{59}$$

$$\sum_c (D_{c,t',l_c} - x_{c,t',l_c}) \leq M_{demand} \tag{60}$$

$$\sum_c x_{c,t',l_c} = y_{t'}, \tag{61}$$

from what we get

$$\frac{x_{c,t',l_c}}{D_{c,t',l_c}} = \frac{x_{1,t',l_1}}{D_{1,t',l_1}} \qquad \text{for } c \in [C] \setminus \{1\},$$

$$\sum_c x_{1,t',l_c} \frac{D_{c,t',l_c}}{D_{1,t',l_1}} = y_{t'},$$

and finally

$$x_{c,t',l_c} = D_{c,t',l_c} \frac{y_{t'}}{\sum_{c'} D_{c',t',l_{c'}}} \qquad \text{for } c \in [C] \setminus \{1\}$$

Piecewise Linear Approximation

The objective is to use a piecewise linear function to approximate function (12). Note that parametric form of the function is known, and it is continuous, twice differentiable, strictly increasing and concave. We adopt the approximation method in Kontogiorgis (2000) by selecting breakpoints on the curve, and connect them into a piecewise linear function. We adopt the infinite norm as the measure of distance for intervals of each piece,

$$\|g\|_{[a,b]} := \max_{x\in[a,b]} |g(x)| \tag{62}$$

which is upper bounded by $\frac{1}{8}(\Delta\tau_k)^2(\|f''\|_{[\tau_k,\tau_{k+1}]})$. Intuitively, if the function has higher curvature on a certain interval, the distance (or the approximation error) of that interval will also be larger. Therefore, we would like to put more breakpoints where the function is "more non-linear". Formally, the paper suggests minimizing

$$\{\max_k (\Delta\tau_k)^2(\|f''\|_{[\tau_k,\tau_{k+1}]})\}$$

with breakpoints τ_k such that

$$(\Delta\tau_k)(\|f''\|_{[\tau_k,\tau_{k+1}]})^{1/2} = \text{constant}, \quad \forall k = 1,\ldots,K. \tag{63}$$

This can be approximated by

$$\int_l^{\tau_k} |f''|^{1/2}dx = \frac{k-1}{K-1}\int_l^u |f''|^{1/2}dx, \quad \forall k = 2,\ldots,K. \tag{64}$$

Algorithm 2:

initialization: Uniform breakpoints
while $\Delta\epsilon \geq \bar{\epsilon}$ **do**
  for k = *2,...,K-1* **do**
    compute change of slope $\alpha_k := \frac{f(\tau_{k+1})-f(\tau_k)}{\tau_{k+1}-\tau_k} - \frac{f(\tau_k)-f(\tau_{k-1})}{\tau_k-\tau_{k-1}}$ compute $\beta_k := |\alpha_k|/(\tau_{k+1}-\tau_{k-1})$
  end
  set $\beta_1 = \beta_2$, $\beta_K = \beta_{K-1}$
  for k = *1,...,K-1* **do**
    compute $h(x) = \beta_k + \beta_{k+1}$
  end
  compute $G(x) := \int_l^x (h(s))^{1/2}ds$
  solve $G(\tau_k) = \frac{k-1}{K-1}G(u)$ for $k = 2, \ldots, K$ to get τ_k
  calculate approximation error ϵ and the change of error $\Delta\epsilon$
end

We note that in this approach, the piecewise linear approximation function uses breakpoints **on** the original function curve, which might not be the "optimal" way of approximating function $\mathrm{sr}(\cdot)$. Also, in practice, we do not have to require the breakpoints to be on the curve; however, there could be potential improvements.

References

Ban, G. Y., & Keskin, N. B. (2020). Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Available at SSRN 2972985*.

Boada-Collado, P., & Martínez-de Albéniz, V. (2020). Estimating and optimizing the impact of inventory on consumer choices in a fashion retail setting. *Manufacturing & Service Operations Management, 22*(3), 582–597.

Boute, R. N., Gijsbrechts, J., & Van Mieghem, J. A. (2020). Digital lean operations: Smart automation and artificial intelligence in financial services. In V. Babich, J. Birge, & G. Hilary (Eds.), *Innovative Technology at the interface of Finance and Operations*. Springer Series in Supply Chain Management. Springer Nature.

Bu, J., Simchi-Levi, D., Xu, Y. (2020). Online pricing with offline data: Phase transition and inverse square law. In *International Conference on Machine Learning* (PMLR) (pp. 1202–1210).

Caro, F., & Gallien, J. (2012). Clearance pricing optimization for a fast-fashion retailer. *Operations Research, 60*(6), 1404–1422.

Caro, F., & Martínez-de Albéniz, V. (2015). Fast fashion: Business model overview and research opportunities. In *Retail supply chain management* (pp. 237–264). Springer.

Cheung, W. C., Simchi-Levi, D., & Wang, H. (2017). Dynamic pricing and demand learning with limited price experimentation. *Operations Research, 65*(6), 1722–1731.

Den Boer, A., & Keskin, N. B. (2019). Dynamic pricing with demand learning and reference effects. *Available at SSRN 3092745*.

Den Boer, A. V. (2015). Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science, 20*(1), 1–18.

Desrosiers, J., & Lübbecke, M. E. (2005). A primer in column generation. In *Column generation* (pp. 1–32). Springer.

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management, 18*(1), 69–88.

Fisher, M., Gallino, S., & Li, J. (2018). Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management Science, 64*(6), 2496–2514.

Frangioni, A., Gendron, B., & Gorgone, E. (2015). *On the computational efficiency of subgradient methods: a case study in combinatorial optimization* (CIRRELT, Centre interuniversitaire de recherche sur les réseaux d'entreprise).

Kelley Jr, J. E. (1960). The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics, 8*(4), 703–712.

Keskin, N. B., & Zeevi, A. (2014). Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research, 62*(5), 1142–1167.

Kontogiorgis, S. (2000). Practical piecewise-linear approximation for monotropic optimization. *INFORMS Journal on Computing, 12*(4), 324–340.

Ma, W., Simchi-Levi, D., & Zhao, J. (2018). Dynamic pricing under a static calendar. *Preprint arXiv:1811.01077*.

Nambiar, M., Simchi-Levi, D., & Wang, H. (2019). Dynamic learning and pricing with model misspecification. *Management Science, 65*(11), 4980–5000.

Natter, M., Reutterer, T., Mild, A., & Taudes, A. (2007). Practice prize report an assortment-wide decision-support system for dynamic pricing and promotion planning in DIY retailing. *Marketing Science, 26*(4), 576–583.

Özer, Ö., Ozer, O., Phillips, R. (2012) *The Oxford handbook of pricing management*. Oxford University Press.

Qiang, S., & Bayati, M. (2016). Dynamic pricing with demand covariates. *Available at SSRN 2765257*.

Smith, S. A., & Achabal, D. D. (1998). Clearance pricing and inventory policies for retail chains. *Management Science, 44*(3), 285–300.

Talluri, K. T., & Van Ryzin, G. J. (2006). *The theory and practice of revenue management* (vol. 68). Springer Science & Business Media.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural Information Processing Systems* (pp. 5998–6008).

Microbanks in Online Peer-to-Peer Lending: A Tale of Dual Roles

Jussi Keppo, Tuan Q. Phan, and Tianhui Tan

1 Introduction

A dynamic banking sector needs new entrants, like any other competitive industry (McWilliams, 2018). Yet, following the 2008 global financial crisis, the growth of new commercial banks has screeched to a halt, and only a handful of new banks have been chartered over the past decade in the U.S.[1] The Federal Deposit Insurance Corporation (FDIC) has thus sought to address the de novo drought by promoting new bank formation.[2] However, in most cases, starting a new bank is a challenging endeavor with a long regulatory review process, mainly attributed to legal barriers (e.g., entry restriction, capital requirement, and more), and further complicated by the uncertain economic environment. While much attention has been placed on policies and regulations, empirical research has shed little light on the nature of

[1]Refer to FDIC website: https://www.fdic.gov/bank/statistical/stats/2020mar/fdic.pdf

[2]Remarks by FDIC Chairman Martin J. Gruenberg at the FDIC Community Banking Conference, "Strategies for Long-Term Success," Arlington, VA. https://www.fdic.gov/news/speeches/spapr0616.html

J. Keppo (✉)
NUS Business School and Institute of Operations Research and Analytics, National University of Singapore, Singapore, Singapore
e-mail: keppo@nus.edu.sg

T. Q. Phan
Faculty of Business and Economics, The University of Hong Kong, Pok Fu Lam, Hong Kong
e-mail: tphan@hku.hk

T. Tan
NUS School of Computing, National University of Singapore, Singapore, Singapore
e-mail: dcstant@nus.edu.sg

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_9

bank formation as a banking behavior in an unregulated, laissez-faire setting, due to the lack of observational data. In this study, we aim to explore bank formation in an empirical context, which allows for the free entry of new banks.

Peer-to-peer lending platforms connect borrowers to lenders. An interesting observation is that some users are conducting micro banking activities, freely performing dual roles as both borrowers and lenders, attempting to make profits by charging greater interest rates on money they lend than the interest rates paid on the money they borrow. They are referred to as *microbanks*. Notably, they face few regulatory restrictions or supervisory powers. Peer-to-peer lending platforms are also reluctant to restrict such activities. Admittedly, microbanks are not identical to traditional banks. The former is more akin to individuals conducting micro banking transactions by switching between the dual roles of borrower and lender in a well-defined financial system, namely the peer-to-peer lending platform, while the latter are chartered financial institutions whose primary tasks include safeguarding monetary deposits and lending money out with government permission. Notwithstanding the differences, microbanks are financial intermediaries that borrow money and issue loans to earn profits, but do not comply with formal regulation or supervision by authorities. In this study, we empirically examine the dynamics of free entry behaviors, by leveraging the unregulated microbanks. We collaborated with one of the largest online peer-to-peer lending platforms in China and obtained a large anonymized dataset of loans, borrowers, and lenders for the year of 2016, and identified a sample of microbanks.

Specifically, we predict the formation of microbanks at monthly intervals. A necessary condition for microbank formation is high creditworthiness. This is natural as it enables a lower interest rate of borrowing than lending, a necessary condition for a profitable banking business. Similarly to traditional banks, microbanks borrow short and lend long. We also demonstrate that the entry of a microbank could be persistent, as the current entry correlates positively with future entry. Namely, if a platform user is a microbank in the current month, she is likely to continue as a microbank in the next month. Also, a higher profit in the current month positively predicts microbank formation in the next month. To make a stronger causal claim, we create a quasi-experimental setting by leveraging the fact that the exact date to receive a repayment is exogenous to the microbanks. We examine the impact of positive and negative shocks to microbank liquidity. We find that a positive shock is positively associated with microbank formation.

Our study makes several contributions. First, analytical models on bank formation usually start with the assumption of free entry or deregulation. However, such theoretical development lacks empirical evidence because the banking industry is highly regulated. Our study contributes to banking literature by providing empirical evidence of entry behavior in an unregulated environment. Second, the concept of microbank, as an informal financial intermediary, is not new. Microbanks mostly exist offline as part of the informal banking landscape. They are not under public scrutiny and, hence, lack observable records. Their behaviors have largely fallen out of the academic spotlight. However, by using data generated by peer-to-peer lending platforms, today, microbanks can be analyzed. Our study unveils the behaviors of microbanks, adding to the broader conversation surrounding informal banking. Third, from the standpoint of peer-to-peer lending platform owners, risk control is critical to promote platform-wide safety and soundness. Given the dual roles of microbanks, their impact on financial contagion and systemic risk cannot be overlooked. Our study serves as the first step to understanding the behaviors of microbanks, providing practical insights to platform owners.

This paper is organized as follows: we provide a comprehensive background to the concept of banking, new bank formation, fintech, and peer-to-peer lending in section "Background". We then introduce our data context in section "Empirical Context and Data". Next, we explore the formation of microbanks and present our findings in section "Formation of a Microbank". We document the impact of shocks on microbank formation in section "Impact of Positive vs. Negative Shocks" and conclude our study in section "Concluding Remarks".

2 Background

2.1 Concept of Banking

The concept of banking has a long history. In ancient Greece, Athenian banks were already found to conduct credit and deposit activities (Cohen, 1997). At the operational level, a bank is defined as *an institution whose current operations consist of granting loans and receiving deposits from the public* (Freixas & Rochet, 2008). Given the nature of financial intermediation, banks have two main functions, namely deposit creation and asset management (Fama, 1980). Similarly, Santomero (1984) explained the existence of banking firms by distinguishing between the firms as asset transformers and as brokers, and highlights the two-sided nature of these firms in modeling their behavior. While contemporary banks perform many complex functions including liquidity and payment services, asset transformation, risk management, monitoring, and information processing (Freixas & Rochet, 2008), our study specifically focuses on a microbank's dual roles as a borrower and as a lender, which correspond to a bank's core activities of accepting deposits and issuing loans, respectively.

The importance of the dual roles has been widely recognized in developing various paradigms and models for bank behavior. For example, Klein (1971) viewed a commercial bank as a subset of financial intermediaries that secures funds from surplus spending units in the form of time deposits, demand deposits, and ownership claims, and transmits them to deficit spending units by investing in a wide variety of earning assets, including loans and securities. The latter constitutes the main source of bank income. Thus, the Monti-Klein model presents a utility model of a monopolistic bank confronted with a demand for loans and a supply of deposits (Klein, 1971; Monti, 1972). In this paradigm, a banking firm is assumed to be an expected value-maximizer considering revenues on loans minus the expenses incurred from deposits (Santomero, 1984). Another paradigm adapts portfolio theory and regards the bank as a portfolio manager. The bank is expected to select a mean-variance efficient portfolio of risky financial products such as loans and deposits by explicitly taking risk into account (Hart & Jaffee, 1974; Pyle, 1971). In their survey, Baltensperger (1980) and Santomero (1984) summarized several extensions of the Monti-Klein model and the Pyle-Hart-Jaffee model. Further, Santomero (1984) formulated a more general objective function by maximizing a multi-period function of asset inflows minus liability outflows minus the costs incurred.

Notably, managing risks is important for banks. Baltensperger (1980) argued that a bank, as a debtor or borrower, faces a withdrawal or liquidity risk that is associated with its liabilities, while a bank, as a creditor or lender, is also exposed to the credit or default risk that is associated with the assets it holds. More systematically, the risks can be classified into four major types: liquidity risks, default risks, interest rate risks, and market risks (Freixas & Rochet, 2008). The deposit activity of banks is affected by liquidity risk, as banks have to make unexpected cash payments to meet any unexpected massive withdrawal requested by their depositors. The credit activity of banks is affected by default risk, as a borrower might not be able to repay her principal and/or interest. In addition, the maturity transformation activity of banks is affected by interest rate risks, as short-term (deposits) interest rates rise above long-term (loans) interest rates. Lastly, market risks and systemic risks affect the marketable assets and liabilities of banks, as the entire market would be influenced at the same time. On the other hand, banking authorities derive a rich set of regulatory instruments to deal with such risks and the inherent fragility of banks to develop a safe and sound banking system. As commented by Taggart (1984), the regulatory system is perhaps what truly distinguishes the banking sector from other wider financial services industry. In fact, it is practically impossible to examine the behaviour of banks without considering the role of banking regulation (Freixas & Rochet, 2008).

2.2 New Bank Formation

The specifics to start a new bank vary from country to country, but in general, a chartered bank application entails a long organization process and permission from regulatory authorities. The entry restriction and capital requirement are typical regulatory impediments to the entry of new banks. In the banking sector, the entry restriction is an early intervention for authorities to prevent instability and limit competition (Barth et al., 2005). At the same time, a capital requirement requires bank entrants to raise and retain a non-negligible minimum amount of capital to enter the market and support its risk profile, operations, and future growth. For example, in the U.S., the initial capital would be in the millions of dollars (Adams & Gramlich, 2016). Taken together, the regulatory burden means that new bank formation is a non-trivial task.

The existing literature examines the impact of entry barriers mainly through the lens of competition (e.g., Cetorelli & Gambera, 2001; Guiso et al., 2004; Jayaratne & Strahan, 1998). The entry, growth, and exit of de novo banks are susceptible to the economic environment (e.g., Adams & Gramlich, 2016; DeYoung, 1999; Lee & Yom, 2016). As a result, the factors affecting entry into local markets are derived from market conditions such as demographics, market concentration, and merger activity. For example, Amel & Liang (1997) found that market size, market growth, and high profits are strong entry determinants. Berger et al. (2004) found that merger activity has an economically significant positive impact on the probability of entry. However, other than analytical models in which the authors made assumptions to remove restrictions to free entry (Besanko & Thakor, 1992), scant literature provides empirical evidence on bank formation by isolating the effect of regulations. One exception is the work by Adams & Gramlich (2016), which inferred that non-regulatory factors such as low interest rates and a weak economy could depress the entry of new banks, though the authors also acknowledged that the standalone regulatory effect is hardly quantifiable. Hence, bank formation in the absence of regulations remains under-explored.

2.3 Fintech and Peer-to-Peer Lending

Fintech is a portmanteau of financial technology, which describes an emerging phenomenon in the twenty-first century. Broadly, it refers to any technological innovation in the financial sector, including financial services operations (e.g. digital banking and credit scoring), payment services (e.g. mobile payment and digital currency), deposit and lending services (e.g. open banking and peer-to-peer lending), and financial market and investment-related services (e.g. high-frequency trading and social trading). Today, technological forces are undoubtedly transforming almost all areas of financial services, and are mediating between markets, regulators, firms, and investors. As evidenced, Boute et al. (2021) showcased the digitalization journey in financial service operations. Byrum (2021) discussed the use of artificial intelligence in financial portfolio management. Bao et al. (2021) identified the opportunities and challenges in accounting fraud detection using machine learning techniques. Kou (2021) highlighted the privacy and transparency related topics in fintech econometrics. The multi-disciplinary nature, multi-faceted problematization, and multi-level analysis have prompted the burgeoning literature on fintech.

Among the wide range of fintech innovations that design and deliver financial products and services through disintermediation, extension of access, and more (Demirguc-Kunt et al., 2018), peer-to-peer lending platforms have received increasing attention. As a specific type of crowdfunding platform that rallies the public for collective funding through the Internet, online peer-to-peer lending enables individual lenders to make unsecured and collateral-free loans directly to unrelated individual borrowers without the intervention of traditional financial intermediaries such as banks. When a borrower posts a loan request, the platform also allows multiple lenders to fund the loan collectively. While such platforms are efficient and cost-effective, the dis-intermediary and anonymous nature of these platforms exaggerate the extent of information asymmetry compared to traditional financial intermediaries, which operate in a tougher regulatory environment. As a result, a growing strand of research focuses on the mitigation of information asymmetry in online peer-to-peer lending. For example, Lin et al. (2013) posited that online friendship can be a signal of credit quality to alleviate adverse selection and information asymmetry, and Iyer et al. (2016) investigated the potential use of nonstandard soft information for lenders to infer borrower creditworthiness. Another stream of literature explores individual behaviors on the platform, such as rational herding (Zhang & Liu, 2012) and home bias (Lin & Viswanathan, 2016) among lenders.

Many studies that examine peer-to-peer lending are conducted from the perspective of either the lender or borrower. In contrast to the implicit assumption that a user can only perform one role, as a lender or a borrower, our focus is instead on the potential existence of dual roles, as both a lender and a borrower simultaneously. We contend that the concept of banking can be extended from institutions to individuals. We refer to this specific type of platform users as microbanks, by adapting the operational definition of banks at the individual level. However, the lack of research on microbanks is surprising, given the openness and disintermediation of peer-to-peer lending platforms.

Our study unravels the existence of microbanks empirically. It is helpful to start examining the nature of banking with unregulated assumption (Fama, 1980). Thus, we situate our study in a loosely regulated context that imposes few restrictions on entry. In our study, we focus on microbank formation on a peer-to-peer lending platform.

3 Empirical Context and Data

For this study, we collaborated with one of the largest online peer-to-peer lending platforms in China (hereinafter referred to as "the platform"). Since its official launch in 2007, the platform has connected a large number of lenders, stretching across cities and counties in China, to (collectively) offer micro loans to meet borrowers' immediate credit needs. Borrowers repay the principal loan and interest in monthly installments. As of 2017, the platform has attracted approximately 10.5 million borrowers and 560,000 individual lenders.

The loan application and transaction processes on the platform are simple, convenient, and fast. A borrower can post a loan request with her expected loan amount, interest rate, number of monthly repayments, together with a title and loan description, anytime and anywhere on a website or mobile app, and receive the outcome of initial screening shortly thereafter. The platform does not require any form of collateral or proof of income from the borrower to secure a loan. Instead, borrowers are required to provide their national identity card information and mobile phone number for verification. In addition, borrowers are encouraged to provide optional information about themselves, including education, location, occupation, marital status, e-mail address, and the mobile phone number of one or two alternative contacts. Hereby, the platform builds an extensive database by combining information volunteered by borrowers and data collected from various third-party sources. The platform deploys a proprietary algorithm to automate the initial screening and credit decision.[3]

A list of active loan requests, called listings, is shown to the lenders, with loan request details and bidding progress status, including borrowing amount, interest rate, number of installments, credit grade, percent completed, and time left. The minimum amount to lend is just RMB 50,[4] and creditors usually provide small amounts to several listings to diversify their risk. As a result, in most cases, a listing will be funded by a number of lenders. When the listing is successfully funded, the borrower receives the amount transferred from lenders. Subsequently, each creditor expects to receive a proportional amount of the borrower's repayment of principal and interest in monthly installments.

For this study, we obtained the platform's anonymized backend data, including listings' details, records of loan issuances and repayments in 2016, as well as user demographics such as age and gender. The details of listings include loan amount, interest rate, number of monthly installments, and audit time. The records of loans issued capture the exact amount that a creditor provides to a listing. The repayment information includes the due amount, due time, paid amount and payment time for each installment of a listing. We constructed a sample by focusing only on a subset of platform users (i.e. microbanks).

We determine whether a platform user i becomes a microbank in month t, denoted by $isBank_{i,t}$, a binary dependent variable. We define the microbanks at the monthly level, as those that conduct both borrowing and lending activities in the same month, with the monthly average interest rate of lending greater than the monthly average interest rate of borrowing. The former part of the definition is a realization of dual roles, and the latter part relates to the interest rate risk. We want to minimize the exposure to interest rate risk mainly due to two reasons. First, the occurrence of interest rate risk is not a frequent event, and in general, traditional banks make their profits from maturity transformation (Freixas & Rochet, 2008). Second, in our context, where a microbank can decide to perform or not perform both the borrowing and lending activities on a monthly basis, we need a consistent definition that accounts for the rationality behind this behavior. In this regard, the interest rate difference can serve as the motivation to become a microbank to make profits. In our study, rather than conditional prediction, we predict for the unconditional probability of microbank formation in the next month ($t + 1$), $isBank_{i,t+1}$, and includes the current microbank indicator, $isBank_{i,t}$, as an independent variable.

[3]There was a lack of well-established credit bureau scores in China during the sample period of this study.

[4]It is equivalent to US$ 7.2, based on the currency exchange rate on December 31, 2016.

The aforementioned operationalization results in a total of 6172 focal platform users, each of whom act as a microbank for at least one month in 2016. It is evident that being a microbank is not a common practice in the platform – only around 13% of lenders in 2016 took out loans in the same year.[5] Summary statistics are detailed in Table 1. Proportion refers to the share of microbanks present in month *t*. The entry rate is defined as the ratio of the number of non-microbanks at *t-1* that become microbanks at *t* to the total number of non-microbanks at *t-1*. Similarly, the exit rate refers to the share of microbanks in the previous month *t-1* that do not continue acting as microbanks in the current month *t*. These non-negligible rates, especially the exit rate, imply that acting as microbanks is a dynamic behavior. The share of microbanks is consistently no more than one-third.[6] Microbanks can enter and exit at little cost, as compared to institutional banks, which have even lower rates (Lee & Yom, 2016).

4 Formation of a Microbank

4.1 Variable Description

We attempt to explore microbank formation as a kind of individual, temporal and voluntary behavior. We posit that the likelihood of being a microbank would be affected by the activities and behaviors on the platform. We classify the activities of a microbank into four categories, based on the dual roles played in prior and current months. First, as a lender, the microbank can perform lending activities in the current month *t*, captured by $lendAmt_{i,t}$, $lendInstal_{i,t}$, and $lendCnt_{i,t}$. Second, as a borrower, the microbank can perform borrowing activities in the current month *t*, captured by $borrowAmt_{i,t}$, $borrowInstal_{i,t}$, and $borrowCnt_{i,t}$. Third, as a lender in prior months, the microbank will receive repayments from debtors. We construct two sets of variables, $collectAmtOT_{i,t}$ and $collectCntOT_{i,t}$, as well as $collectAmtD5_{i,t}$ and $collectCntD5_{i,t}$, by considering the on-time repayments and delayed repayments respectively. Fourth, as a borrower in prior months, the microbank needs to make repayments for her loans. Similarly, we derive two sets of variables, $repayAmtOT_{i,t}$ and $repayCntOT_{i,t}$, as well as $repayAmtD5_{i,t}$ and $repayCntD5_{i,t}$. In addition, we define a simple profitability measure, $profit_{i,t}$, by focusing on the earned interests net of paid interests in month *t*. Furthermore, we estimate the microbank's capital and

[5]The sample size can be increased by relaxing the time period requirement. For example, 20,755 platform users can be identified as microbanks on a yearly basis, while our sample consists of 6172 microbanks defined at the monthly level. Although the latter is smaller in size, it allows us to explore the time-varying behaviors throughout 2016.

[6]Due to data constraints, we do not explicitly predict for the individual's entry or exit behavior, which is in the form of conditional probability. Both entry and exit rates are conditional on the behavior of microbanks in the previous month, leading to a reduced sample size for conditional prediction.

Table 1 Entry and exit dynamics

	(Feb)	(Mar)	(Apr)	(May)	(Jun)	(Jul)	(Aug)	(Sep)	(Oct)	(Nov)	(Dec)
Proportion	21.5%	26.9%	27.04%	24.79%	24.72%	31.93%	32.15%	29%	17.61%	12.99%	11.92%
Entry rate	10.39%	15.19%	13.74%	13.15%	12.52%	18.92%	14.83%	13.23%	10.2%	8.52%	8.18%
Exit rate	36.58%	30.37%	36.81%	43.8%	38.24%	28.44%	30.95%	37.7%	64.25%	66.05%	62.97%

normalize the monetary variables accordingly. For each month, we calculate *total inflow*, defined as the sum of the amounts borrowed and the repayments received at t, and *total outflow*, defined as the sum of the amounts lent and the repayments made at t. We use the larger of the two as the proxy of capital for microbank i at t. Finally, we construct a directed network based on the flow of funds among microbanks, and the directed edge indicates the direction of fund which goes from one microbank to another. We tend to capture peer influence based on degree centrality, i.e. $indegreeBank_{i,t}$ and $outdegreeBank_{i,t}$. We use a 10-month period from February to November 2016,[7] and summarize the list of variables and descriptive statistics among the 6172 microbanks in Table 2.

4.2 Preliminary Analyses

We estimate a logistic model, specified in Eq. 1, where the dependent variable, $isBank_{i,t+1}$, is central to our interest. To start, we break down the analysis by fixing the time window to be exactly one month to explore if there exists any pattern or consistency over time. Hence, we run a cross-sectional model on 10 subsets of our sample data, each of which represents a monthly snapshot of the sample.

$$logit\left(isBank_{i,t+1}\right) = \alpha + \beta_1\, isBank_{i,t} + \beta_2\, profit_{i,t} + \beta_3\, activities_{i,t} + \beta_4\, centralities_{i,t} + \beta_5\, controls_i \tag{1}$$

We compare the results from February to November in 2016 (Table 3). First, the coefficient of $isBank_{i,t}$, which indicates the status of microbank in the current month t, is consistently significantly positive. This shows that being a microbank in the current month increases the likelihood to continue forming a microbank in the following month, implying stickiness to remain the role of microbank. Second, our results suggest the more profits earned by a microbank in the current month, the more likely it is to act as a microbank in the next month. This also demonstrates our hypothesis that making profits can serve as the motivation to become a microbank on the platform. Next, the lending amount is positively related to microbank formation, whereas the borrowing amount is negatively related.

[7]The earlier and later data (in January and December in 2016) are discarded due to the lack of loan repayment behaviors and the truncation on the lead of dependent variable respectively.

Table 2 Descriptive statistics (N = 61,720)

Name	Description	Mean	SD
lead.isBank	Whether the focal user is a microbank in month $t+1$	0.239	0.427
isBank	Whether the focal user is a microbank in month t	0.249	0.432
profit	The difference between the amount of earned interests and amount of paid interests during month t, normalized by the microbank's proxy of capital	0.015	0.053
lendAmt	The total amount the microbank issues during month t, normalized by the microbank's proxy of capital	0.420	0.432
lendInstal	The average number of installments associated with loans issued by the microbank during month t	6.066	5.451
lendCnt	The number of loans issued by the microbank during month t	294.947	951.421
borrowAmt	The total amount that the microbank borrows during month t, normalized by the microbank's proxy of capital	0.219	0.356
borrowInstal	The average number of installments associated with the loans received by the microbank during month t	2.590	4.953
borrowCnt	The number of loans received by the microbank during month t	1.649	4.459
collectAmtOT	The total amount of installment repayments that are paid on time during month t, normalized by the microbank's proxy of capital	0.383	1.068
collectCntOT	The number of installment repayments that are paid on time during month t	1025.813	3014.882
collectAmtD5	The total amount of installment repayments that are delayed for more than 5 days during month t, normalized by the microbank's proxy of capital	0.026	0.081
collectCntDt	The number of installment repayments that are delayed for more than 5 days on time during month t	71.659	223.699
repayAmtOT	The total amount that the microbank repays on time during month t, normalized by the microbank's proxy of capital	0.836	21.069
repayCntOT	The number of loans that the microbank repays on time during month t	3.837	9.391
repayAmtD5	The total amount that the microbank delays to repay for more than 5 days during month t, normalized by the microbank's proxy of capital	0.223	21.893
repayCntD5	The number of loans that the microbank delays to repay for more than 5 days during month t	0.021	0.258
outdegreeBank	The number of microbanks that the focal microbank lends money to during month t	0.061	1.394
indegreeBank	The number of microbanks that the focal microbank borrows money from during month t	0.082	0.822
age	Age of the microbank	33.668	7.955
isFemale	Whether the microbank is a female	0.207	0.405

Table 3 Monthly logistic regression results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Month	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*
	Dependent variable: lead.isBank									
isBank	1.065^{***}	0.868^{***}	0.755^{***}	0.931^{***}	0.666^{***}	0.584^{***}	1.146^{***}	0.810^{***}	0.361^{***}	0.368^{**}
	(0.170)	(0.148)	(0.133)	(0.137)	(0.147)	(0.125)	(0.126)	(0.116)	(0.129)	(0.148)
profit	1.993	2.062	1.446	0.233	2.910^{***}	3.034^{***}	2.442^{**}	2.736^{*}	3.766^{***}	5.608^{***}
	(1.305)	(1.316)	(1.207)	(0.942)	(1.111)	(1.121)	(1.116)	(1.407)	(1.368)	(1.277)
lendAmt	1.475^{***}	0.954^{***}	0.797^{***}	0.689^{***}	1.071^{***}	0.660^{***}	0.789^{***}	1.011^{***}	0.813^{***}	0.326^{*}
	(0.154)	(0.157)	(0.147)	(0.138)	(0.129)	(0.134)	(0.136)	(0.137)	(0.156)	(0.172)
borrowAmt	−0.060	−0.223	-0.401^{**}	-0.766^{***}	-1.174^{***}	-0.965^{***}	-0.966^{***}	-0.465^{**}	-1.287^{***}	-1.330^{***}
	(0.220)	(0.199)	(0.196)	(0.199)	(0.224)	(0.193)	(0.189)	(0.205)	(0.223)	(0.210)
Activities	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Centralities	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6172	6172	6172	6172	6172	6172	6172	6172	6172	6172
Log Likelihood	−2376.46	−2378.87	−2488.11	−2276.75	−2455.26	−2361.49	−2407.03	−2363.13	−1861.35	−1780.22
Akaike Inf. Crit.	4794.923	4799.735	5018.215	4595.492	4952.524	4764.969	4856.052	4768.269	3764.697	3602.431

Note: isBank is an indicator that equals 1 if the focal user is a microbank in the current month. *profit* is the difference between the amount of earned interest and amount of paid interest in the current month, normalized by the microbank's proxy of capital. Activity variables comprise total lending amount, average lending installments, lending times, total borrowing amount, average borrowing installments, borrowing times, amount of on-time collections, number of on-time collections, amount of delayed collections, number of delayed collections, amount of on-time repayments, number of on-time repayments, amount of delayed repayments, and number of delayed repayments. Centrality variables include indegree centrality and outdegree centrality. Control variables comprise age and gender. The estimates of activity, centrality and control variables are omitted due to inconsistency and insignificance, except for total lending amount and total borrowing amount. The estimates of centrality variables are not consistently significant, probably because microbanks are anonymous to each other on the platform

$^{*}p < 0.1; ^{**}p < 0.05; ^{***}p < 0.01$

4.3 Fixed-Effects Logit Model

To analyze panel data, we use a fixed-effects logistic regression model specified in Eq. 2. The explanatory variables include *isBank*$_{i,t}$, *profit*$_{i,t}$, *activities*$_{i,t}$, and *centralities*$_{i,t}$. We estimate their parameters by taking the individual fixed effects α_i (and time fixed effects γ_t) into consideration.[8]

$$\begin{aligned} logit\left(isBank_{i,t+1}\right) = \beta_1\, isBank_{i,t} + \beta_2\, profit_{i,t} + \beta_3\, activities_{i,t} \\ + \beta_4\, centralities_{i,t} + \gamma_t + \alpha_i \end{aligned} \tag{2}$$

The results are shown in Table 4, based on an unconditional maximum likelihood estimation (columns 1 and 2) (Stammann et al., 2016) and a conditional likelihood estimation (columns 3 and 4) (Gail et al., 1981).[9] Unlike the results in the previous section, the microbank indicator in the current month is weakly positively associated with microbank formation in the next month. The effect is diminishing after controlling for the time effect in columns 2 and 4. A plausible explanation is that the time effect, in addition to individual effect, is more influential in explaining the variations in microbank formation. Next, profitability has a positive impact on the microbank formation. It may imply that platform users who are capable of making money are more likely to recognize the potential of microbanks. Alternatively, it may also suggest that users are aware of the importance of liquidity to microbanks even without the imposition of regulations. Interestingly, those who tend to borrow smaller amounts for a shorter period but lend larger amounts for a longer term are more likely to become microbanks. On the other hand, receiving repayments or paying installments exerts a negative influence on predicting the formation of microbanks, regardless of being on time or delayed. The seemingly counter-intuitive negative results might correspond to the perceptible exit rates in Table 1. It could also be attributed to the fact that the repayments to receive or installments to pay are accumulated from lending and borrowing activities in any of previous months. Yet, the timings of repayment receipt and installment payment affect a microbank's cash flow liquidity and profitability. In the following section, we further investigate the relationship between cash flow liquidity and microbank formation.

[8] To identify the structural parameters, a total of 6170 observations of 617 microbanks with non-varying response are dropped.

[9] Statistics were done using R 3.4.0 (R Core Team, 2017), with the bife (v0.7; Stammann et al., 2020) and survival (v 2.41–3; Therneau et al., 2017) packages.

Table 4 Fixed-effects logistic regression results

	(1)	(2)	(3)	(4)
	Dependent variable: lead.isBank			
isBank	0.097^{*}	0.018	0.087.	0.018
	(0.048)	(0.049)	(0.045)	(0.046)
profit	1.565^{**}	1.499^{**}	1.396^{**}	1.348^{**}
	(0.505)	(0.527)	(0.477)	(0.496)
lendAmt	1.591^{***}	1.59^{***}	1.385^{***}	1.379^{***}
	(0.061)	(0.062)	(0.056)	(0.057)
lendInstal	0.009^{*}	0.015^{**}	0.009^{*}	0.014^{**}
	(0.004)	(0.005)	(0.004)	(0.005)
lendCnt	0.00008^{***}	0.00009^{***}	0.00007^{**}	0.00007^{**}
	(0.00002)	(0.00002)	(0.00002)	(0.00002)
borrowAmt	-0.509^{***}	-0.506^{***}	-0.432^{***}	-0.431^{***}
	(0.073)	(0.075)	(0.069)	(0.07)
borrowInstal	-0.013^{*}	-0.007	-0.01^{*}	-0.006
	(0.004)	(0.005)	(0.004)	(0.004)
borrowCnt	0.291^{***}	0.286^{***}	0.245^{***}	0.24^{***}
	(0.008)	(0.009)	(0.008)	(0.008)
collectAmtOT	-0.234^{***}	-0.231^{**}	-0.207^{***}	-0.204^{**}
	(0.067)	(0.072)	(0.063)	(0.068)
collectCntOT	-0.000002	0.000004	-0.000004	0.0000008
	(0.00002)	(0.00002)	(0.00002)	(0.00002)
collectAmtD5	-7.782^{***}	-7.356^{***}	-7.024^{***}	-6.613^{***}
	(0.695)	(0.743)	(0.654)	(0.698)
collectCntD5	-0.002^{***}	-0.002^{***}	-0.002^{***}	-0.001^{***}
	(0.0002)	(0.0003)	(0.0002)	(0.0002)
repayAmtOT	-0.011^{*}	-0.012^{*}	-0.01^{*}	-0.011^{*}
	(0.005)	(0.005)	(0.005)	(0.005)
repayCntOT	-0.038^{***}	-0.036^{***}	-0.032^{***}	-0.03^{***}
	(0.004)	(0.004)	(0.004)	(0.004)
repayAmtD5	-2.308.	-2.315.	-2.225.	-2.214.
	(1.304)	(1.319)	(1.282)	(1.291)
repayCntD5	-0.858	-0.915	-0.761	-0.808
	(0.692)	(0.709)	(0.663)	(0.675)
outdegreeBank	-0.088^{**}	-0.086^{**}	-0.079^{**}	-0.077^{**}
	(0.031)	(0.031)	(0.029)	(0.029)
indegreeBank	-0.023	-0.033.	-0.02	-0.03.
	(0.017)	(0.018)	(0.017)	(0.017)
Time fixed effects	No	Yes	No	Yes
Individual fixed effects	Yes	Yes	Yes	Yes
Observations	55,550	55,550	55,550	55,550
Log Likelihood	$-20{,}751.6$	$-20{,}334.1$	$-15{,}094.7$	$-14{,}725.9$

Note: isBank is an indicator that equals 1 if the focal user is a microbank in the current month. *profit* is the difference between the amount of earned interest and amount of paid interest in the current month, normalized by the microbank's proxy of capital. Activity variables comprise total lending amount, average lending installments, lending times, total borrowing amount, average borrowing installments, borrowing times, amount of on-time collections, number of on-time collections, amount of delayed collections, number of delayed collections, amount of on-time repayments, number of on-time repayments, amount of delayed repayments, and number of delayed repayments. Centrality variables include indegree centrality and outdegree centrality. Time and individual fixed effects are included in the models as indicated

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

5 Impact of Positive vs. Negative Shocks

5.1 Quasi-Experimental Setting

Like traditional banks, microbanks are exposed to default risk, which is non-trivial as borrowers on the platform face little recourse and lenders have no collateral to collect after a default event.[10] As mentioned, a microbank can perform four categories of activities. As a lender or borrower, the microbank can decide to lend or borrow anytime. However, as a prior borrower, the microbank has little freedom to alter the schedule to repay her monthly installments. Meanwhile, as a prior lender, the microbank has no control over the exact time when her borrowers make repayments. That is, the time when the microbank receives the proportional repayments is almost exogenous to her. Therefore, we only focus on the latter two activity categories. To be specific, the inflows obtained by the microbank refer to the repayments she receives, while the outflows refer to the repayments she makes. By comparing the actual time of inflows against the scheduled time of outflows of the microbank, two types of unexpected shocks could occur. Namely, for repayments that are expected after the microbank's earliest own payment due date, if the microbank receives any of them earlier than her earliest own payment due date, there is an unexpected shock that exerts a positive influence on the microbank's liquidity, known as positive shock. Similarly, for repayments that are expected earlier than the microbank's latest own payment due date, when the microbank does not fully receive those amounts before her latest own payment due date, there is an unexpected shock that negatively affects her liquidity, known as negative shock.

Figure 1 depicts an example where the microbank has four repayments to receive and two repayments to make. As illustrated, the repayments to receive are ordered by their due dates chronologically, denoted as inflow (I), inflow (II), inflow (III), and inflow (IV), and the installments to repay are also denoted sequentially as outflow (A) and outflow (B). We argue that the microbank is exposed to a positive shock when the microbank receives any of inflow (II), inflow (III) and inflow (IV) earlier than the due date of outflow (A). In contrast, the microbank experiences a negative shock when she fails to collect the repayments of both inflow (I) and inflow (II) before the due date of outflow (B). In this way, we manage to create a quasi-experiment setting by leveraging the positive and negative shocks. In the next section, we investigate the impact of such liquidity shocks on the formation of microbanks.

[10]The recovery rate (i.e. the proportion recovered by the lender when the borrower defaults) is also important. However, in this study, we examine a simplified scenario by assuming a zero recovery rate.

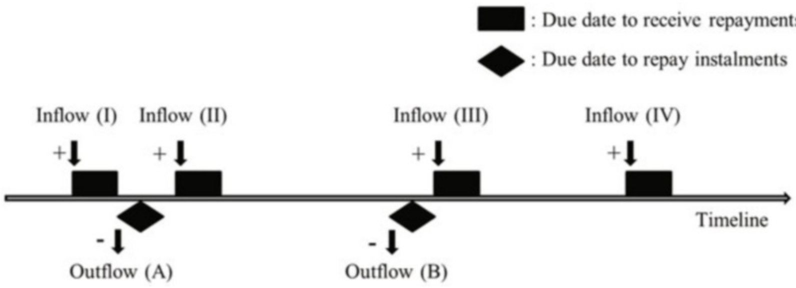

Fig. 1 Sequence of inflows and outflows

5.2 *Difference-in-Differences Model with Matched Sample*

We define a microbank that has ever been exposed to such shocks as a treated microbank, and create the variable, $treated_{i,t}$ to indicate that the microbank i receives the shock in the month t.[11] We construct a matched data sample based on the general idea of propensity score matching (PSM). To be specific, we first divide the treatment group into 10 subgroups according to the month when microbanks receive such a treatment. Second, for each treatment subgroup, we compute a score for the propensity of a microbank to be treated using predictors on demographics, microbank status, behaviors, and network centrality.[12] Third, we adopt a one-on-one nearest neighbor matching with replacement to find each of the treated microbanks ($treated_{i,t} = 1$) with an untreated microbank ($treated_{i,t} = 0$) that has the closet propensity score.

We run a difference-in-differences model on the matched sample, specified in Eq. 3. For a treated microbank, we set the time variable, $after_{i,t}$ as zero for all the months preceding the treatment, and one otherwise. For each (matched) pair, the microbank in the control group is assigned with the same value of $after_{i,t}$ as her counterparty in the treatment group. The other variables are similar to those in section "Formation of a microbank".

$$logit\left(isBank_{i,t+1}\right) = \alpha + \beta_1\, isBank_{i,t} + \beta_2\, treated_{i,t} * after_{i,t} + \beta_3\, profit_{i,t} + \beta_4\, activities_{i,t} + \beta_5\, centralities_{i,t} + \beta_6\, controls_i + \gamma_t \tag{3}$$

[11] For simplicity, we focus on the first occurrence of the shock in this study.

[12] The original data is in long form. Given a month τ, we only consider the predictors prior to $t = \tau$. We transform the data into wide form and run a simple logit regression to estimate the score. To impose a requirement that $isBank_{i,\tau}$ is the same for a matched pair, we run the matching process separately on the two subsets of data where $isBank_{i,\tau} = 1$ and $isBank_{i,\tau} = 0$ separately.

We derive several variants of the treatment variable, $treated_{i,t.}$ First, we define that $treated_{i,t}$ equals 1 if a positive shock occurs and at least one repayment is received earlier than the microbank's own earliest due date for more than seven days (i.e. treatedPosBin). Second, we define that $treated_{i,t}$ equals 1 if a negative shock occurs and at least one repayment is received later than the microbank's own latest due date for more than seven days (i.e. treatedNegBin). In addition to the two binary variable definitions, we also define the third and fourth forms of treatment using amount ratio, namely the total amount of early (or late) repayments the microbank receives earlier (or later) than the due date of her own first (or last) repayment, divided by the amount of the microbank's repayments to make during the same month. To define the third form of treatment, we set $treated_{i,t}$ as 1 if a positive shock occurs and the early amount ratio is larger than 25% (i.e. treatedPosAmt).[13] Similarly, we define the fourth form of treatment by setting $treated_{i,t}$ to be 1 if a negative shock occurs and the late amount ratio is larger than 5% (i.e. treatedNegAmt).[14]

We present the results in Table 5. Our results suggest that the impact of a positive shock on microbank formation is significantly positive on being a microbank in the next period, whereas that of a negative shock is non-significant. In general, early repayment is a more frequent event than late repayment on the platform. The asymmetric effects could be due to the asymmetric distribution of positive and negative shocks. Microbanks are more likely to receive a positive shock than a negative shock, as evidenced by the different cutoffs chosen for treatedPosAmt and treatedNegAmt. On one hand, positive shock increases liquidity and improves one's confidence to become a microbank in the near future. On the other hand, those microbanks-to-be are not bad at risk management, and they are capable of absorbing the negative effect created by a negative shock. Hence, a negative shock does not significantly affect one's decision towards microbank formation, which implies that microbanks can voluntarily manage the potential risks, even without the imposition of regulations.

6 Concluding Remarks

The banking sector is intensively regulated today. This is hardly surprising, given the critical impact banks have on economic development and human welfare. Regulatory and supervisory policies cover almost every aspect, from the entry of new banks, to how they exit, in hopes of enhancing bank operations and lowering

[13] We vary the cutoffs such as 33%, 50% and 66%. The size of treatment group is decreasing, but the results are almost consistent.

[14] We try larger cutoffs such as 10%. The size of the treatment group is smaller, but the results are consistent. However, when the cutoff is even larger, the size will be reduced dramatically. For example, when the cutoff is 25%, only 17% of original sample receives a negative shock.

Table 5 DID results

	(1)	(2)	(3)	(4)
Treatment	*treatedPosBin*	*treatedNegBin*	*treatedPosAmt*	*treatedNegAmt*
	Dependent variable: lead.isBank			
isBank	1.063^{***}	1.481^{***}	1.482^{***}	1.568^{***}
	(0.032)	(0.037)	(0.036)	(0.041)
treated*after	0.357^{***}	−0.038	0.270^{***}	0.095
	(0.088)	(0.086)	(0.088)	(0.078)
treated	0.274^{***}	0.514^{***}	0.561^{***}	0.383^{***}
	(0.050)	(0.050)	(0.051)	(0.046)
after	-0.759^{***}	-0.317^{***}	-0.718^{***}	-0.429^{***}
	(0.086)	(0.084)	(0.085)	(0.076)
profit	0.300^{***}	0.248^{***}	0.213^{***}	0.195^{***}
	(0.020)	(0.022)	(0.021)	(0.023)
Activities	Yes	Yes	Yes	Yes
Centralities	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	95,680	76,800	74,580	64,420
Log Likelihood	−37,973.650	−31,228.010	−30,835.480	−26,595.900
Akaike Inf. Crit.	76,021.300	62,530.020	61,744.970	53,265.810

Note: isBank is an indicator that equals 1 if the focal user is a microbank in the current month. *profit* is the difference between the amount of earned interest and amount of paid interest in the current month, normalized by the microbank's proxy of capital. Activity variables comprise total lending amount, average lending installments, lending times, total borrowing amount, average borrowing installments, borrowing times, amount of on-time collections, number of on-time collections, amount of delayed collections, number of delayed collections, amount of on-time repayments, number of on-time repayments, amount of delayed repayments, and number of delayed repayments. Centrality variables include indegree centrality and outdegree centrality. Control variables comprise age and gender. The estimates of activity, centrality and control variables are omitted

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

systemic fragility (Barth et al., 2005). From the perspective of banks, however, regulatory compliance also incurs a variety of non-trivial costs (Elliehausen, 1998). The debate on the impact of bank regulatory practices, in conjunction with the complex motivations underlying these regulations (Barth et al., 2005), implies that in the modern banking sector, banking behaviour is more or less shaped by the existence of regulations, and further complicated by the broader political economy context. Empirical data on banking can hardly be collected in isolation from any regulatory force, leaving researchers to wonder how banks will behave in the absence of regulations and restrictions. On the other hand, technological advancements in the financial sector are helping to improve the traceability of financial activities at the individual level. The plethora of immediately accessible data provides us the opportunity to conduct an empirical investigation on microbanks as one of the informal financial intermediaries in the underexplored field of micro and informal finance, and opens up a promising avenue for future empirical research. Our research can also help build insights into the informal banking sector, especially prevalent in third-world and developing economies.

Acknowledgments We thank Dr. Xuesong Lu for providing valuable suggestions for data cleaning. We gratefully acknowledge the funding support from the Singapore Ministry of Education (Grant R-252-000-A08-112).

References

Adams, R. M., & Gramlich, J. (2016). Where are all the new banks? The role of regulatory burden in new Bank formation. *Review of Industrial Organization, 48*(2), 181–208.

Amel, D. F., & Liang, J. N. (1997). Determinants of entry and profits in local banking markets. *Review of Industrial Organization, 12*(1), 59–78.

Baltensperger, E. (1980). Alternative approaches to the theory of the banking firm. *Journal of Monetary Economics, 6*(1), 1–37.

Bao, Y., Hilary, G., & Ke, B. (2021). Artificial intelligence and fraud detection. In V. Babich, J. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations* (Springer series in supply chain management, forthcoming). Springer Nature.

Barth, J. R., Caprio, G., & Levine, R. (2005). *Rethinking bank regulation: Till angels govern*. Cambridge University Press.

Berger, A. N., Bonime, S. D., Goldberg, L. G., & White, L. J. (2004). The dynamics of market entry: The effects of mergers and acquisitions on entry in the banking industry. *Journal of Business, 77*(4), 797–834.

Besanko, D., & Thakor, A. V. (1992). Banking deregulation: Allocational consequences of relaxing entry barriers. *Journal of Banking and Finance, 16*(5), 909–932.

Boute, R. N., Gijsbrechts, J., & Van Mieghem, J. A. (2021). Digital lean operations: Smart automation and artificial intelligence in financial services. In V. Babich, J. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations* (Springer series in supply chain management, forthcoming). Springer Nature.

Byrum, J. (2021). AI in financial portfolio management: Practical considerations and use cases. In V. Babich, J. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations* (Springer series in supply chain management, forthcoming). Springer Nature.

Cetorelli, N., & Gambera, M. (2001). Banking market structure, financial dependence and growth: International evidence from industry data. *Journal of Finance, 56*(2), 617–648.

Cohen, E. (1997). *Athenian economy and society: A banking perspective*. Princeton University Press.

Demirguc-Kunt, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2018). *The global Findex database 2017: Measuring financial inclusion and the Fintech revolution*. The World Bank.

DeYoung, R. (1999). Birth, growth, and life or death of newly chartered banks. *Economic Perspectives, 23*(3), 18–35.

Elliehausen, G. (1998). The cost of Bank regulation: A review of the evidence. *Federal Reserve Bulletin, 84*(4), 252.

Fama, E. F. (1980). Banking in the theory of finance. *Journal of Monetary Economics, 6*(1), 39–57.

Freixas, X., & Rochet, J. C. (2008). *Microeconomics of banking* (2nd ed.). The MIT Press.

Gail, M. H., Lubin, J. H., & Rubinstein, L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika, 68*(3), 703–707.

Guiso, L., Sapienza, P., & Zingales, L. (2004). Does local financial development matter? *The Quarterly Journal of Economics, 119*(3), 929–969.

Hart, O. D., & Jaffee, D. M. (1974). On the application of portfolio theory to depository financial intermediaries. *The Review of Economic Studies, 41*(1), 129–147.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: inferring the quality of small borrowers. *Management Science 62*(6), 1554–1577.

Jayaratne, J., & Strahan, P. E. (1998). Entry restrictions, industry evolution, and dynamic efficiency: Evidence from commercial banking. *Journal of Law and Economics, 41*(1), 239–273.

Klein, M. A. (1971). A theory of the banking firm. *Journal of Money, Credit and Banking, 3*(2), 205–218.

Kou, S. (2021). FinTech econometrics: Privacy preservation and the wisdom of the crowd. In V. Babich, J. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations* (Springer series in supply chain management, forthcoming). Springer Nature.

Lee, Y. Y., & Yom, C. (2016). *The entry, performance, and risk profile of de novo banks*. FDIC Center for Financial Research Paper, WP 2016-03.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science, 59*(1), 17–35.

Lin, M., & Viswanathan, S. (2016). Home Bias in online investments: An empirical study of an online Crowdfunding market. *Management Science, 62*(5), 1393–1414.

McWilliams, J. (2018, December 6). We can do better on De Novos (opinion). *American Banker*. https://www.americanbanker.com/opinion/fdic-chairman-jelena-mcwilliams-we-can-do-better-on-de-novos

Monti, M. (1972). Deposit, credit and interest rate determination under alternative bank objective functions. In *Mathematical methods in investment and finance* (pp. 431–454). North-Holland.

Pyle, D. H. (1971). On the theory of financial intermediation. *The Journal of Finance, 26*(3), 737–747.

Santomero, A. M. (1984). Modeling the banking firm: A survey. *Journal of Money, Credit and Banking, 16*(4), 576–602.

Stammann, A., Heiss, F., & Mcfadden, D. (2016). Estimating fixed effects logit models with large panel data. Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel.

Stammann, A., Czarnowske, D., Heiss, F., & McFadden, D. (2020). *bife: Binary Choice Models with Fixed Effects*. R package version 0.7. https://cran.r-project.org/web/packages/bife/index.html

Taggart, R. A. (1984). Modeling the banking firm: Comment. *Journal of Money, Credit and Banking, 16*(4), 612–616.

Therneau, T. M., Lumley, T., Elizabeth, A., & Cynthia, C. (2017). *survival: Survival Analysis*. R package version 2.41-3. https://cran.r-project.org/web/packages/survival/index.html

Zhang, J., & Liu, P. (2012). Rational herding in microloan markets. *Management Science, 58*(5), 892–912.

FinTech Econometrics: Privacy Preservation and the Wisdom of the Crowd

Steven Kou

1 Introduction

FinTech is a new buzzword that can mean different things to different people. It is so new that even the spelling of the word is not unanimously adopted; for example, it can also be written as fintech or financial technology. To a person doing information technology, FinTech might be mainly about mobile payment systems, to a person processing loan applications classification and regression trees, and perhaps to many other people simply Bitcoins.

It is perhaps fruitful to borrow a quote from Jack Ma, the founder of Alibaba Group: "FinTech takes the original financial system and improves its technology. TechFin is to rebuild the system with technology" (South China Morning Post, Dec 2, 2016). In other words, in his view FinTech is about a new type of finance and TechFin is about a new type of technology. By the way, in the same interview, Jack Ma said that what he has done at Alibaba is TechFin.

In this paper I shall take a similar view as Jack Ma did. More precisely, the focus of the paper is on the new financial theory and analytical tools arisen due to the rapid advance in information technology; neither detailed technology issues (such as mobile payment systems and distributed databases) nor philosophical issues (e.g. whether FinTech is beneficial to the society) will be discussed here.

Obviously the topics to be chosen are biased by the author's own expertise and objectives. In this regard there are many FinTech topics that will not be

A Chapter in "Innovative Technology at the Interface of Finance and Operations," Forthcoming.

S. Kou (✉)
Department of Finance, Questrom School of Business, Boston University, Boston, MA, USA
e-mail: kou@bu.edu

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_10

discussed here: (1) Special data mining and machine learning tools useful for financial applications. Since many data scientists know more about these than I do, I only refer the reader to Gu et al. (2018), Horel and Giesecke (2020), Fan et al. (2020) and references therein. (2) Portfolio optimization for robo-advising, which requires eliciting the risk preferences from investors and provides asset allocation solutions that are consistent with the investors' common wisdom. The literature is too large to be completely mentioned here, and I refer the reader to Capponi et al. (2020), Dai et al. (2021b,c), Strub et al. (2020) and references therein. (3) Economics of cryptocurrencies and stablecoins. This is an important research topic, as cryptocurrencies play an essential role in preventing double spending in public blockchains using the proof of work protocols. See Hou et al. (2020), Trimborn and Härdle (2018), a survey paper by Harvey (2016), government backed stablecoins in Garratt (2016), and various designs of stablecoins discussed in Lipton et al. (2018), Cao et al. (2018). (4) Corporate finance and blockchains; see, e.g., Ritchken and Wu (2021) a survey paper by Yermack (2017). (5) Interaction of FinTech and operations management; see Babich and Hilary (2020) and Babich and Birge (2021). (6) Peer-to-peer lending; see Keppo et al. (2021) and references therein.

In addition, there is also a large literature on peer-to-peer equity finance (e.g. initial coin offerings), crowdfunding, crowdsourcing and the economics of mining protocols for public blockchains; instead of citing some references here, one can simply search online and find many papers related to these topics.

Instead, in this survey paper I shall focus on the following two topics in econometrics related to privacy and transparency issues, arisen due to information flows in FinTech. More precisely, we shall discuss: (1) Econometrics for sensitive financial data with privacy preservation in the era of big data. Here we want to restrict certain information flows. (2) The wisdom of the crowd and prediction markets, in the presence of new available information from anonymous individual-level trading data. For example, all transactions of a given account on a public blockchain are available anonymously, and in sports the betting positions for different groups (such as people betting on home teams, away team, and the bookmakers) are also announced regularly.

2 Econometrics with Privacy Preservation

In the era of big data how to collect and analyze data while still preserving people's privacy needs is an important issue in science and social science; see, e.g., Posner (1981), Acquisti et al. (2015), Acquisti et al. (2016), and the references therein. For example, people may not want to share their sensitive data, such as salary information, trading positions, and voting preferences. The recent governmental investigation of Facebook and the EU General Data Protection Regulation (GDPR) are further examples of the concerns from the public about data privacy.

What is unique in finance is that very often it is mandatory for firms to share some aspects of their sensitive data. More precisely, there is a central administration, to which individual firms have to report their sensitive data (e.g. trading profits and losses), and then the central administration, after doing some statistical analysis, will broadcast the summary statistics (e.g. Value-at-Risk) to the general public. For example, according to the Basel regulations, everyday banks have to report certain trading outcomes to regulators, who will broadcast aggregated risk analysis publicly. It should be noted that in most cases, some summary statistics of the data, perhaps not the original sensitive data, are needed for the central administration to do a study. This should be in contrast with sensitive individual medical data, which may not be required to be reported to the central administration to share with the public, unless there is a public health concern such as a pandemic; in other words, generally speaking, individuals do not have to "share medical secrets." That is why distributed statistical inference (which allows data to stay inside individual computers, but not sharing) is popular in medical statistics, but not necessarily the multi-party secure computation (which requires sharing of encrypted data).

In this section I shall survey some of the recent results related to econometrics for the encrypted data, in which individual parties only report encrypted data publicly so that the central administration can do exact statistical inference without knowing the original sensitive data.

2.1 Background

There are ways to collect the data to get certain summary statistics without revealing the original data. For example, it is well known, perhaps even hundreds of years ago, that a simple algorithm is to add an independent random variable from an identical distribution with mean zero to each sample before collecting the data. If the sample size is large enough (e.g. millions), then by adding all collected (encrypted) numbers together the random noises tend to cancel out and the sample mean can be recovered approximately. Although this algorithm is simple and takes very little computational time, it is not easy to quantify the exact (deterministic) error of the algorithm, due to the difficulty in knowing how large the sample size is needed for the approximation to work well. Furthermore, the above algorithm cannot recover other quantities easily, such as the quantiles (especially at extreme levels) and the histogram of the whole distribution.

A major theoretical breakthrough is the theory of secure multi-party computation for secret sharing in Shamir (1979) and Yao (1982). Basically if an algorithm can be transformed into computing some functions based on individual private inputs involving only two operations, addition and multiplication, then in theory a protocol can be developed to implement the algorithm securely without any errors and without disclosing the private inputs to one another. Therefore, in principle, if a statistical inference procedure can be composited in terms of the addition and multiplication from individual inputs, then it can be achieved while preserving privacy.

However, the resulting algorithms from the multi-party secure computation may be computationally demanding, as it is generally difficult to reduce all operations into just two operations, addition and multiplication. As a real example, a secure multi-party computation protocol was implemented in 2008 in Denmark, where 1200 farmers participated in an auction to determine the market price of sugar-beet contracts without disclosing their own bid and ask prices. Nonetheless, the computation took about half an hour on a super computer.

Abbe et al. (2012) made a significant contribution to real-world applications of secure multi-party computation by desigining simple and efficient protocols, including a secure-sum protocol and three secure-inner-product protocols, “for computing aggregate risk measures based on standard sample moments such as means, variances, and covariances—the typical building blocks of financial risk measures”. Their protocols guarantee the privacy when the parties are semi-honest (also known as honest-but-curious), as each party can recover nothing about other parties’ input beyond what can be recovered from the output itself. Another appealing feature of the protocols is the robustness to collusion. The authors also argue that privacy is ensured in that the best information one can recover is based on his/her private input, and the output is a set of some uniform distributions.

By carefully designing a finite number of problem-driven functions so that the secure-sum protocol of Abbe et al. (2012) can be repeatedly applied, Cai and Kou (2019) propose an algorithm for privacy-preserving statistical inferences, including recovering the whole empirical distribution, linear regression, logistic regression, maximum likelihood estimation, quantile estimation, recovering the convolution, and distributed statistical inference. For simplicity, we refer to this algorithm as the ER (encryption and recovery) algorithm, because using this algorithm one can conduct statistical inference based on the *encrypted* data by *recovering* just enough information about the original sensitive data for accomplishing the task of the required statistical inference.

The basic ideas of Abbe et al. (2012) and Cai and Kou (2019) will be reviewed in the next section. Before we do that it is helpful to make a comparison with the theory of differential privacy, which was proposed by Dwork et al. (2006) to address the privacy preservation issue by adding random noises to ensure that whether certain parties are contained in a data set or not cannot be deduced reliably; see also, e.g., Duchi et al. (2014). Since differential privacy involves additional random noises, it incurs additional errors for data analysis; however, it is robust to malicious parties. By contrast, the ER algorithm achieves exact data analysis without additional errors, yielding the same numerical results as in the traditional (unprotected) statistical inference, but may break down in the face of malicious parties. Thus, the ER algorithm is more suitable in the setting where different parties have to report, by law, their computed true numbers to the central party.

2.2 Basic Ideas in Abbe et al. (2012) and Cai and Kou (2019)

Consider m parties and assume that for each $i = 1, \cdots, m$, party i has d sensitive numbers $x_{i1}, \cdots, x_{id}$ that constitute a sensitive vector x_i, i.e., $x_i = (x_{i1}, \cdots, x_{id})^T$. Here $m \in \mathbb{N}$, $m \geq 3$, $d \in \mathbb{N}$ and T denotes the transpose. These parties could be financial firms with sensitive numbers being, e.g., certain trading positions or some numbers in their balance sheets; they could also be voters with sensitive numbers representing their voting preferences. Using a linear transform if needed, without loss of generality we assume that all entries of the sensitive vectors $x_1, \cdots, x_m$ belong to the interval $[0, 1)$, i.e., $x_{ij} \in [0, 1)$ for any $i = 1, \cdots, m$ and $j = 1, \cdots, d$.

If all the vectors $x_1, \cdots, x_m$ were public, one would apply the traditional statistical theory to analyze these public data directly and make statistical inferences. However, these vectors $x_1, \cdots, x_m$ are sensitive, and each party does not want any other party or the central administration to know its own sensitive vector. Thus, a natural and fundamentally important issue is how one can gather information from the data while still preserving the privacy of these m parties.

2.2.1 Basic Idea in Abbe et al. (2012)

For simplicity, we consider the case of one dimension, i.e. $d = 1$. The algorithm in Abbe et al. (2012) has 3 steps.

Step 1. Each party sends one uniform random number to each of the rest of $m - 1$ parties, and, in return, receives $m - 1$ random numbers. More precisely, for all $i, \hat{i} = 1, \cdots, m$ with $i \neq \hat{i}$, party i provides party $\hat{i}$ with a uniformly distributed random number $R_{i\hat{i}}$ over $[0, m)$.

Step 2. Each party adds to the original sensitive number all the random numbers received, and then minus all the random numbers sent out. Afterwards, the party only reports the final outcome to the central administration. More precisely, for $i = 1, \cdots, m$, party i only discloses S_i to the central administration, where

$$S_i := x_i + \sum_{\hat{i} \neq i} (R_{\hat{i}i} - R_{i\hat{i}}).$$

Step 3. The central administration computes $Q := \sum_{i=1}^{m} S_i$.

Now clearly, using only the encrypted data, the central administration can recover the sample mean of the original sensitive data exactly, because

$$Q = \sum_{i=1}^{m} x_i,$$

as all the R's canceled out by design.

However, there is a significant drawback in the above simple algorithm, because Step 2 could review some statistical distribution about x_i, as the reported number S_i is positively correlated with x_i. To make S_i independent of x_i and having a uniform distribution, Abbe et al. (2012) modify Steps 2 and Steps 3 as

$$S_i := \left\{ x_i + \sum_{\hat{i} \neq i} (R_{\hat{i}i} - R_{i\hat{i}}) \right\} \bmod m, \qquad Q := \left\{ \sum_{i=1}^{m} S_i \right\} \bmod m,$$

respectively. Now it takes some algebra to show that we still have $Q = \sum_{i=1}^{m} x_i$.

It should be noted that there are two differences between the algorithm in Abbe et al. (2012) and a well known secure summation protocol in Banaloh (1986), where (1) there is no central administration and the random numbers are transmitted in a circular fashion between the parties, and (2) the range of the uniform numbers, instead of being $[0, m)$, is $[0, Q)$ with Q being an arbitrary large number.

2.2.2 Basic Idea in Cai and Kou (2019)

To extend the above idea from computing the sample mean to more general statistical inference procedures, it will be helpful to consider an example of the simple linear regression model $y_i = \beta_0 + \beta_1 z_i + \epsilon_i$ for$i = 1, \cdots, m$. Without loss of generality, we assume that $z_i \in [0, 1)$ and $y_i \in [0, 1)$ for $i = 1, \cdots, m$. If this assumption is not satisfied, z_i and y_i can be modified via linear transforms such that this assumption is satisfied, as long as z_i and y_i are bounded by known lower and upper bounds.

If all the sample data $x_i = (x_{i1}, x_{i2})^T := (z_i, y_i)^T$ for $i = 1, \cdots, m$ are public (here $d = 2$), then the least square estimators $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are given, respectively, by

$$\widehat{\beta}_1 = \frac{\frac{\sum_{i=1}^m z_i y_i}{m} - \frac{\sum_{i=1}^m z_i}{m} \frac{\sum_{i=1}^m y_i}{m}}{\frac{\sum_{i=1}^m z_i^2}{m} - \left(\frac{\sum_{i=1}^m z_i}{m} \right)^2} \quad \text{and} \quad \widehat{\beta}_0 = \frac{\sum_{i=1}^m y_i}{m} - \widehat{\beta}_1 \frac{\sum_{i=1}^m z_i}{m}.$$

If all the sample data are sensitive, we can recover $\widehat{\beta}_0$ and $\widehat{\beta}_1$ exactly while still preserving individual privacy by recovering the following functions:

$$\frac{1}{m} \sum_{i=1}^{m} z_i, \quad \frac{1}{m} \sum_{i=1}^{m} y_i, \quad \frac{1}{m^2} \sum_{i=1}^{m} z_i^2, \quad \frac{1}{m^2} \sum_{i=1}^{m} z_i y_i.$$

Because each term inside the above four summations only involves the sensitive numbers from the party i sensitive and does not involve anything about the sensitive numbers from other parties (i.e. there is no cross term such as $y_i y_j$), we can simply repeat the algorithm in Abbe et al. (2012) four times. Note that we have used $\frac{1}{m}$ and $\frac{1}{m^2}$ to normalize the above 4 terms so that each of them is between 0 and 1.

This is a basic insight in Cai and Kou (2019). In fact, their contribution is twofold.

(i) By choosing a set of functions (called ER functions) that only involves the sum of individual sensitive data, a general framework for statistical inference with privacy preservation can be achieved. The examples include recovering the whole empirical distribution, linear regression, logistic regression, maximum likelihood estimation, quantile estimation, recovering the convolution, and distributed statistical inference. Of course, these ER functions are different for different applications. To work out the details, in each case the involved statistical inference procedure has to be decomposed into smaller pieces, so that each piece can be written as ER functions. Sometimes, this is trivial, like the case of the simple linear regression; other times, this is more complicated, involving Fourier inversion, etc.

(ii) A detailed and rigorous proof for the privacy preservation results of the protocol is given, using the invariant equidistribution from number theory. Previously, the definition of privacy preservation is not very clear in most of the existing papers. Abbe et al. (2012) try to clarify this by suggesting to use uniform distribution to define privacy preservation. However, the observed encrypted data may have a singular distribution without a density function at all, thus motivating us to consider the invariant equidistribution.

To glimpse the difficulty of measuring the effectiveness of privacy preservation, consider the following simple example. There are 3 parties, each with sensitive number x_i, respectively, $1 \le i \le 3$. Each party uses a vector of random numbers y_i to generate an encrypted number x_i', to be reported to the central administration. Suppose the central administration computes the sample mean $\bar{x}$ from the encrypted data x_i' (without seeing the sensitive numbers x_i and encryption device numbers y_i), and broadcasts $\bar{x}$ to all parties. Then party 1 immediately knows the sum $x_2 + x_3$, party 2 knows $x_1 + x_3$, and party 3 knows $x_1 + x_2$, based on the public information $\bar{x}$ and their individual sensitive numbers.

Thus, in this case the privacy is preserved if (i) for party i, conditioning on the public information (including $\bar{x}$), the private information (including x_i and the random numbers that i send out and receive), and the information from other parties obtained by colluding with a hacker (including x_j', $j \neq i$), that joint distribution of the sensitive numbers of other people, $\{X_j, \ j \neq i\}$, that party i can infer is still uniformly distributed on the plane $\sum_{j \neq i} X_j = c_i$, where c_i is a given number $\sum_{j \neq i} x_j$; (ii) for the central administration, conditioning on the public information ($\bar{x}$) and private information collected by the central administration (x_i'), the joint distribution of the sensitive numbers of all people, $\{X_j, \ 1 \le i \le 3\}$, that the central administration is uniformly distributed on the plane $\sum_{j=1}^{3} X_j = 3\bar{x}$.

In general, for more general statistical problems the plan may, however, become several hyperplanes, or even separated manifolds (if the ER functions are non-linear), yielding some singular distributions which make the definition of uniform distribution more complicated.

It is worth pointing out two things: (1) Karr et al. (2005) seems to be the first paper on secure linear regression, by using the algorithm in Banaloh (1986). (2) The algorithms of Abbe et al. (2012) and Cai and Kou (2019) are all theoretically rooted in the general theory of secure multi-party computation and secret sharing. The focus of Cai and Kou (2019) is not to advance the theory of secure multi-party computation and secret sharing. Instead, the main objective and hence the contribution of the paper is to study how to design specific algorithms suitable for a general statistical inference framework for sensitive data analysis, which can be regarded as a counterpart of the traditional statistical inferences for public data analysis.

2.2.3 The ER Algorithm

The *ER algorithm* with the sensitive data $x_1, \cdots, x_m$ and the ER functions $g_i(x)$, with

$$g_i(x) \in [0, 1), \quad x \in [0, 1), \quad 1 \le i \le L,$$

is given as follows.

An ER Algorithm

- **E-Step (Encryption)**
  - **Substep 1**: Each party i for $i = 1, \cdots, m$ computes $g_1(x_i)$, $g_2(x_i)$, $\cdots$, $g_L(x_i)$.
  - **Substep 2**: For all $i, \hat{i} = 1, \cdots, m$ with $i \neq \hat{i}$, party i provides party $\hat{i}$ with L uniformly distributed random numbers over $[0, m)$: $R_{i\hat{i}}(1)$, $R_{i\hat{i}}(2)$, $\cdots$, $R_{i\hat{i}}(L)$. Here $R_{i\hat{i}}(k)$ for $i, \hat{i} = 1, \cdots, m$ with $i \neq \hat{i}$ and $k = 1, 2, \cdots, L$ are independent and identically uniformly distributed over $[0, m)$.

- **Substep 3**: For $i = 1, \cdots, m$ and $k = 1, \cdots, L$, party i only discloses $S_i(k)$ to the central administration, where

$$S_i(k) := \left\{ g_k(x_i) + \sum_{\hat{i} \neq i} (R_{\hat{i}i}(k) - R_{i\hat{i}}(k)) \right\} \bmod m. \tag{1}$$

  $S_i(k)$ for $i = 1, \cdots, m$ and $k = 1, \cdots, L$ are called ***encrypted data***, which are essentially all the information the central administration can obtain from the ER algorithm.

- **R-Step (Recovery)**
  – The central administration computes

$$Q(k) := \left\{ \sum_{i=1}^{m} S_i(k) \right\} \bmod m \quad \text{for } k = 1, \cdots, L. \tag{2}$$

  $Q(1), \cdots, Q(L)$ are called recovered information about the original sensitive data.

$Q(1), \cdots, Q(L)$ are called recovered information because it is shown in Cai and Kou (2019) that from the encrypted data the central administration can recover the following information about the original sensitive data:

$$Q(1) = \sum_{i=1}^{m} g_1(x_i), \quad Q(2) = \sum_{i=1}^{m} g_2(x_i), \quad \cdots, \quad Q(L) = \sum_{i=1}^{m} g_L(x_i). \tag{3}$$

Furthermore, a rigorous proof for the privacy preservation results of the ER algorithm using the invariant equidistribution from number theory is also given there.

Based on the recovered information about the original sensitive data, one can use certain statistical methods to do statistical inference. Since the recovered information depends on the ER functions one selects, to do statistical inference (e.g. linear regression, maximum likelihood estimation, logistic regression, recovery of quantiles, and recovery of histograms and empirical distributions) for the encrypted sample data, one needs to find appropriate ER functions $g_1(\cdot), \cdots, g_L(\cdot)$ to exactly recover the required information for inference. For example, if one adds an additional regression variable, then one needs to add more ER functions, and redo the ER algorithm. Thus, one may apply other approximate encryption methods first to do variable selections for the regression, and then use our ER algorithm to get the exact (not approximate) statistical inference results.

2.2.4 Privacy Preservation for All Parties from the Viewpoint of Each Individual Party in the Secure-Sum Protocol

Denote by View_i the view of party i or the information that party i obtains from the secure-sum protocol. We consider a worst-case scenario that a hacker reveals all $\{S_i\}$ to all parties; ideally without the hacker all $\{S_i\}$ are only know to the central administration. We shall see that the privacy is still preserved even in this scenario.

Without loss of generality, we take party 1 as an example. Then View_1 includes all S_i for $i = 2, \cdots, m$. Because party 1 knows its own sensitive number x_1 as well as the uniform random numbers received from and sent to all other parties, i.e., R_{i1} and R_{1i} for $i = 2, \cdots, m$, It can be shown that:

(i) View_1 can be equivalently expressed as

$$\text{View}_1 = \{I_i : \; i = 2, \cdots, m\}, \tag{4}$$

where I_i is defined as

$$I_i := (S_i - R_{1i} + R_{i1}) \bmod m.$$

(ii) Given R_{i1} and R_{1i} for $i = 2, \cdots, m$, S_i for $i = 2, \cdots, m$ can be recovered from I_i for $i = 2, \cdots, m$, and conversely given R_{i1} and R_{1i} for $i = 2, \cdots, m$, I_i for $i = 2, \cdots, m$ can be recovered from S_i for $i = 2, \cdots, m$. More precisely,

$$S_i = (I_i + R_{1i} - R_{i1}) \bmod m, \quad \text{for } i = 2, \cdots, m. \tag{5}$$

Cai and Kou (2019) prove that privacy is preserved for all parties from the viewpoint of each individual party. More precisely, $\text{View}_1 \equiv I$ and all the information that party 1 obtains from the secure-sum protocol, has an invariant equidistribution on its range; its range is symmetric with respect to the parameters $x_2, \cdots, x_m$ and given as follows:

$$\text{Range}(\text{View}_1) = \text{Range(I)},$$

where $\text{Range}(I)$ denotes the range of I and is given by

$$\text{Range(I)} = \left\{ (u_2, u_3, \cdots, u_m)^T : \left(\sum_{j=2}^{m} u_j \right) \bmod m = \sum_{j=2}^{m} x_j, \;\; u_j \in [0, m) \right\}.$$

Next, we consider the privacy preservation from the viewpoint of the central administration. Assume that there exists a central administration that has no input but can obtain $S_1, S_2, \cdots, S_m$, i.e., the information disclosed by all parties in the secure-sum protocol. Denote by View_0 the view of the central administration or the information the central administration obtains from the secure-sum protocol. Then View_0 can be expressed as:

$$\text{View}_0 = \{S_i : i = 1, \cdots, m\}.$$

Cai and Kou (2019) prove that privacy is preserved for all parties from the viewpoint of the central administration. More precisely, $\text{View}_0 \equiv (S_1, S_2, \cdots, S_m)^T$, i.e., all the information that the central administration obtains from the secure-sum protocol, has an invariant equidistribution on its range. Moreover, its range is symmetric with respect to the parameters $x_1, x_2, \cdots, x_m$ and given as follows:

$$\text{Range}(\text{View}_0) = \left\{ (u_1, \cdots, u_m)^T : \left(\sum_{j=1}^{m} u_j \right) \bmod m = \sum_{j=1}^{m} x_j,\ u_j \in [0, m) \right\}.$$

2.2.5 Recovering Histograms and Empirical Distributions

As an example, the ER algorithm can also be used to compute the empirical probabilities or the relative frequencies, i.e., the ratio of the number of the sensitive vectors in which a specified event happens to the total number m of sensitive vectors. For simplicity, we assume that $d = 1$; a general $d \in \mathbb{N}$ can be treated analogously.

Suppose that we expect to know the proportion of the sensitive numbers out of $x_1, \cdots, x_m$ that fall into a specified region $A \subset [0, 1)$. If the sample data are public, then the corresponding empirical probability is given by

$$\text{ep} = \frac{\sum_{i=1}^{m} \mathrm{I}_A(x_i)}{m},$$

where $\mathrm{I}_A(x)$ is the indicator function equal to 1 if $x \in A$ and equal to 0 otherwise. If the sample data are sensitive, we can also apply the ER algorithm to recover the empirical probability ep.

Apply the ER algorithm with $L = 1$ and $g_1(x) = \frac{\mathrm{I}_A(x)}{2}$ and consider the following estimator ep_e constructed from the encrypted data:

$$\text{ep}_\text{e} := \frac{2Q(1)}{m} = \frac{2\left\{\left\{\sum_{i=1}^{m} S_i(1)\right\} \bmod m\right\}}{m},$$

where $S_i(1)$ is defined in (1) and $Q(1)$ is defined in (2). Then we have $\text{ep}_\text{e} \equiv \text{ep}$ and moreover, the privacy is preserved in the sense of equi-invariant distribution. Here we select $g_1(x) = \frac{\mathrm{I}_A(x)}{2}$ rather than $g_1(x) = \mathrm{I}_A(x)$ to guarantee the ER function satisfies the assumption that $g_1(x) \in [0, 1)$ for any $x \in [0, 1)$.

In general, for a given set of bins we can apply the ER algorithm to construct the histogram from the encrypted data by repeatedly using the above method to compute all the empirical probabilities corresponding to these bins. Indeed, if the possible values that the sensitive numbers can take are known and the number of these possible values is finite, then we can set the bins to be sufficiently small such that each bin contains only one possible value that the sensitive numbers can take, and then the method above essentially yields the whole empirical distribution function of the sensitive numbers.

We assume that the sensitive numbers possessed by the m parties can take only N values $x_1^*, \cdots, x_N^* \in [0, 1)$ with $N \in \mathbb{N}$. This assumption is satisfied in many practical cases. For instance, if the single sensitive number represents certain price which is less than 1, then these N values could be $0, 0.01, 0.02, \cdots, 0.99$ with $N = 100$.

Recovering the whole empirical distribution function of the sensitive numbers is equivalent to recovering the empirical probabilities that the sensitive numbers are equal to each possible value. If the sample data are public, then the corresponding empirical probabilities are given by

$$\mathrm{ep}(k) = \frac{\sum_{i=1}^{m} \mathrm{I}_{\{x_i = x_k^*\}}}{m} \quad \text{for } k = 1, \cdots, N,$$

where $\mathrm{I}_{\{x=x_k^*\}}$ is the indicator function equal to 1 if $x = x_k^*$ and equal to 0 otherwise. We can apply the ER algorithm with $L = N$ and

$$g_k(x) = \frac{\mathrm{I}_{\{x=x_k^*\}}}{2} \quad \text{for } k = 1, \cdots, N.$$

Then we can use the following $\mathrm{ep_e}(k)$ to compute the empirical probability $\mathrm{ep}(k)$.

$$\mathrm{ep_e}(k) := \frac{2Q(k)}{m} = \frac{2\left\{\left\{\sum_{i=1}^{m} S_i(k)\right\} \bmod m\right\}}{m} \quad \text{for } k = 1, \cdots, N, \tag{6}$$

where $S_i(k)$ is defined in (1) and $Q(k)$ is defined in (2). By (3) we obtain

$$\mathrm{ep_e}(k) = \frac{2Q(k)}{m} = \frac{2}{m}\sum_{i=1}^{m} g_k(x_i) = \frac{2\sum_{i=1}^{m} \frac{\mathrm{I}_{\{x_i=x_k^*\}}}{2}}{m} = \mathrm{ep}(k).$$

As a by-product, we indeed obtain the set of all sensitive numbers without knowing which party possesses which sensitive number. For general $d \in \mathbb{N}$, this idea still applies except that N could be much larger.

After recovering the whole empirical distribution function of the sensitive numbers and obtaining the set of all sensitive numbers, one can in turn compute sample mean, sample variance, and other descriptive statistics such as the mode, various quantiles, the interquartile range, as well as various histograms with different bins.

2.3 Numerical Results

Due to the similarity, this section is only focused on the recovery of the whole empirical distribution function of the sensitive numbers discussed before.

2.3.1 Exact Recovery of the Whole Empirical Distribution and the Histograms via the ER Algorithm

We shall report some numerical results relating to the exact recovery of the whole empirical distribution of the sensitive numbers. With the exactly recovered empirical distribution, we are capable of recovering various descriptive statistics of the sensitive numbers such as the mode, various quantiles, the interquartile range, as well as various histograms with any fixed bins.

Figure 1 compares the recovered empirical distributions and the recovered histograms with those true ones for three groups of randomly generated sensitive numbers. It can be seen that we achieve the exact recovery of them in all the three cases. In addition, our ER algorithm is fast in that the CPU time used to recover one empirical distribution exactly is only approximately 2 s on a standard desktop computer.

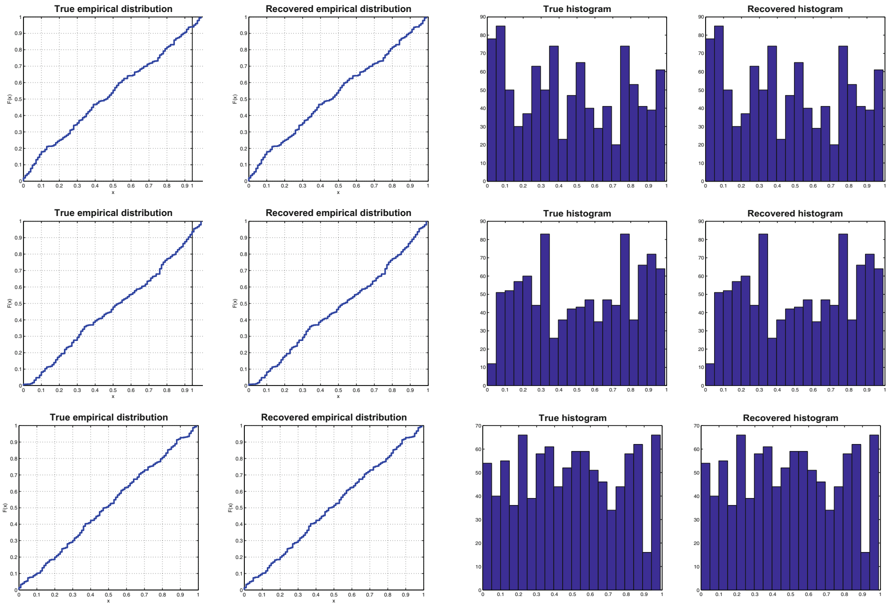

Fig. 1 Comparing the recovered empirical distributions and the recovered histograms with the true ones for three groups of $m = 1000$ sensitive numbers generated randomly. The number of bins used in the histograms is 20. Sensitive numbers can only take values of $0, 0.01, 0.02, \cdots, 0.99$. It can be seen that in all three cases, the central administration can exactly recover the whole empirical distributions and the histograms of the true sensitive numbers through the ER algorithm. The CPU time used to recover one empirical distribution exactly is approximately 2 s

2.3.2 Demonstration of Privacy Preservation

Next we shall attempt to demonstrate the privacy preservation from the perspectives of both the central administration and an individual party when only this individual party colludes with the hackers. Due to the high dimensions, we only demonstrate this partially by numerically illustrating some images of the related invariant equidistributions.

Consider an example where there are three parties ($m = 3$) with sensitive numbers (x_1 x_2 x_3)=(0.25 0.5 0.75), and (x_1^* x_2^* x_3^* x_4^*)=(0 0.25 0.5 0.75). Here x_1^*, x_2^*, x_3^* and x_4^* are four ($N = 4$) possible values the sensitive number can take. We try to estimate the empirical distribution by using $L = 4$ and $g_k(x_i) = \frac{1}{2}\mathrm{I}_{\{x=x_k^*\}}$ for $k = 1, \cdots, 4$. Thus, we have $Q(1) = 0$, $Q(2) = Q(3) = Q(4) = 1/2$.

From Fig. 2, we can see that all those four random vectors have invariant equidistributions or uniform distributions on their respective ranges. Moreover, their ranges are all symmetric with respect to the associated sensitive numbers. For example, the range of $(S_1(3), S_2(3), S_3(3))$ is $\sum_1^3 g_3 \equiv \{(u_1, u_2, u_3)^T : (u_1 + u_2 + u_3) \bmod 3 = Q(3) \equiv 0.5,\ u_i \in [0, 3)\}$, which is symmetric with respect to x_1, x_2 and x_3. Although the central administration knows $\sum_{i=1}^3 g_3(x_i) \equiv 0.5$ and from this it can infer that one sensitive number is 0.5 and two others are not 0.5, it cannot identify which party has the sensitive number 0.5. Likewise, the range of $(I_2(3), I_3(3))$ is $\sum_2^3 g_3 \equiv \{(u_2, u_3)^T : (u_2 + u_3) \bmod 3 = \sum_{i=2}^3 g_3(x_i) \equiv 0.5,\ u_i \in [0, 3)\}$, which is symmetric with respect to x_2 and x_3. Although party 1 that colludes with the hacker knows $\sum_{i=2}^3 g_3(x_i) \equiv 0.5$ and from this it can infer that exactly one of party 2 and party 3 has the sensitive number 0.5, it cannot distinguish between these two parties.

3 The Wisdom of the Crowd and Prediction Markets

The concept of the wisdom of the crowd indicates that one can gather information and make a prediction from a group's aggregated inputs, even if the individual inputs are noisy. The concept was suggested by Galton (1907), in which he mentioned that at a 1906 country fair in Plymouth the median guess of (about 800) people who participated in a contest to estimate the weight of an ox is 1207 pounds, accurate within 1% of the true weight of 1198 pounds. In the era of digital innovation, this concept has been widely used by social information sites, e.g. Wikipedia and Stack Exchange.

Although in many cases the idiosyncratic noise associated with each individual judgment gets canceled over the average, leading to a good answer from either polls or prediction markets, in general there is no guarantee that the crowd wisdom yields a better judgment than an individual does. This is especially true if the majority of the people do not know the right answer (e.g. most of people might not realize

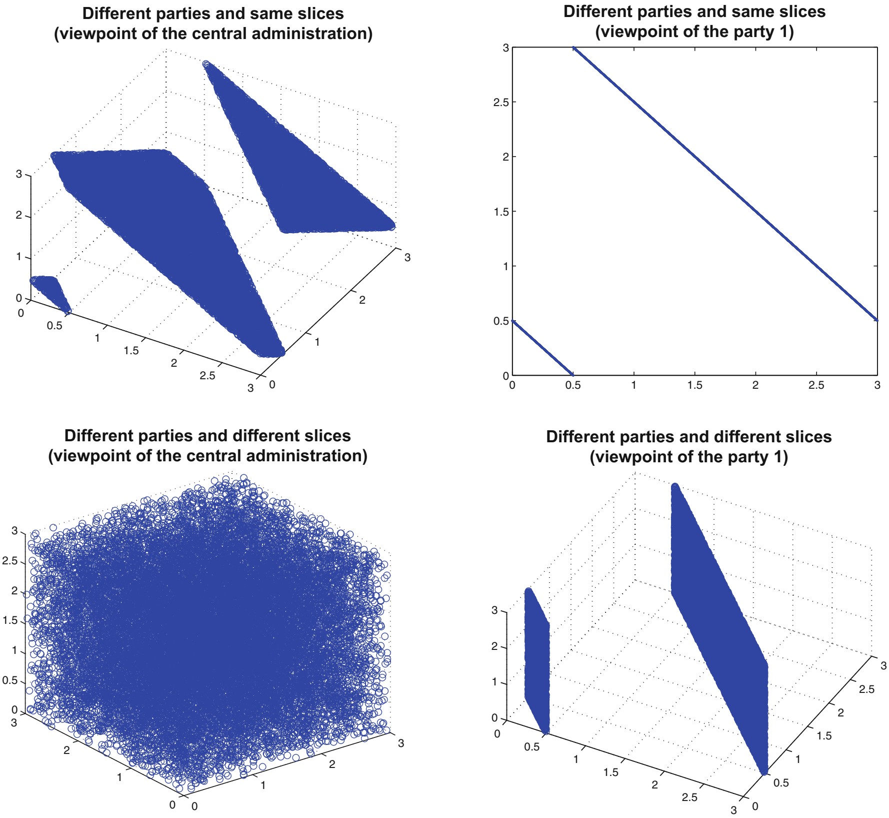

Fig. 2 Illustration of the viewpoints of the central administration (the two panels on the left) and party 1 when only party 1 colludes with the hackers (the two panels on the right). Different slices correspond to different ER functions. More specifically, the figure demonstrates the distributions of the following four random vectors by randomly generating 20,000 samples for each of them. (i) The top-left panel: $(S_1(3), S_2(3), S_3(3))$, whose range is a collection of disjoint hyperplanes; (ii) The top-right panel: $(I_2(3), I_3(3))$, whose range is a collection of disjoint hyperplanes; (iii) The bottom-left panel: $(S_2(3), S_3(3), S_2(4))$, whose range is a cube; (iv) The bottom-right panel: $(I_2(3), I_3(3), I_3(4))$, whose range is a Cartesian product of a collection of disjoint hyperplanes and an interval. We can see that all four random vectors have either invariant equidistributions or uniform distributions on their respective ranges.

that Philadelphia is not the capital of Pennsylvania) or the answer to a question is currently unknown to human beings (e.g. whether extraterrestrial life exists).

Surprisingly, a very creative solution based on the Bayesian Truth Serum (BTS) is given in Prelec et al. (2017). They point out that in a survey if, in addition to answering one survey question, a participant is also asked to make a prediction of the distribution of answers to the survey question of the whole participants, then a consistent estimator related to a quantity in the objective world can be obtained, under a suitable set of regularity conditions on the subjective and objective worlds, if the majority of the people do not know the right answer.

For example, if a survey question asking participants whether Philadelphia is the capital of Pennsylvania, then the simple majority vote obtained from the answers to the question may be wrong, as most of the participants may give the wrong answer. However, if, in addition to the original question, another question is asked about the respondent's guess of the percentage of the respondents who correctly answer the first question, then a consistent estimator can be found by combining the answers to both questions, and by giving more weight to the answers by the minority of respondents, who answer the first question correctly but think that the majority of respondents will get the first question wrong.

Dai et al. (2021a) study a related problem of the wisdom of the crowd, namely making a prediction or judgment using prediction markets. In particular, they attempt to answer the following question: Can one design a prediction market, such that a consistent estimator can be found using data in this market? The answer is yes, if additional individual level data is available. Indeed, they give two estimators, a market-adjusted estimator based solely on market inputs, including individual trading positions and a hybrid estimator based on both market and inputs from one survey question (instead of two questions), such that both estimators are consistent under the BTS setting, but under different regularity conditions.

Furthermore, using a real data set of sports betting, which include trading positions of three groups (people betting on home team, people betting on away team and the bookmakers), Dai et al. (2021a) find that the market-adjusted estimator can significantly improve the accuracy of prediction, if the number of participants is large so that the market has enough liquidity.

3.1 The Bayesian Learning Setup

In this section, we shall first review the Bayesian learning setup in Prelec et al. (2017). Consider humans live in a world where there are unknown events. Denote the value of such an unknown event to be Ω, which is a random variable in the human world. Without loss of generality, we assume there are only $K + 1$ possible outcomes in Ω, i.e. $\Omega \in \{0, 1, \cdots, K\}$. The human world is different from a conceptually omniscient and eternal world, which is referred to as an objective world. In the objective world, there is no time elapse and the truth (e.g. the outcome of any future events, or correct answers to any questions) is revealed. Thus, the value of the event in the objective world, denoted by $\Omega_{obj} \in \{0, 1, \cdots, K\}$, is a deterministic constant but unknown to the human world.

People can learn about the objective world in certain ways by engaging in cognitive learning, e.g. observations, experiments, or reasoning. The consequence of such behavior leads to a signal. More precisely, suppose that there are N people learning about Ω, each receiving a signal S_i, $1 \leq i \leq N$, which is a sequence of independent identically distributed (i.i.d.) random variables with $S_i \sim f_{obj}(\cdot|\Omega_{obj})$ i.i.d., where $f_{obj}(\cdot|\Omega_{obj})$ is known as the objective likelihood function. For simplicity, we also assume $S_i \in \{0, 1, \cdots, K\}$. Thus, the joint objective likelihood function is $S_1, \cdots, S_N \sim \prod_{i=1}^{N} f_{obj}(S_i|\Omega_{obj})$. For simplicity, here we only consider the case with binary outcome, that is, Ω_{obj} and Ω take a value of either 0 or 1.

After learning about the unknown event, an agent forms his subjective view about Ω based on the Bayesian updating rule. More precisely, the agent i updates the prior distribution $\mathbb{P}_i(\Omega = k) = \pi_i^k$, by using the private signal S_i, to get a posterior probability $\Omega|S_i \sim p_i(\cdot|S_i)$, according to the Bayesian formula and the subjective likelihood function $f_i(S_i|\Omega)$. Subjective probabilities are denoted by p_i and $1 - p_i$, where

$$p_i := p_i(1|S_i) = \frac{\pi_i f_i(S_i|\Omega = 1)}{\pi_i f_i(S_i|\Omega = 1) + (1 - \pi_i) f_i(S_i|\Omega = 0)}, \quad 1 - p_i = p_i(0|S_i). \tag{7}$$

Here $\pi_i = \mathbb{P}_i(\Omega = 1)$, as we omit the superscript for simplicity in the binary case. Then the joint subjective likelihood function is $S_1, \cdots, S_N|\Omega \sim \prod_{i=1}^{N} f_i(S_i|\Omega)$. Note that different likelihood function means that different agents can interpret the signal in subjective and distinct ways.

For example, suppose in the human world, Ω takes values in 0 (lose) or 1 (win) with common prior probability 0.6 for winning. In the objective world, $\Omega_{obj} = 0$ (doomed to lose) and the objective likelihood function $f_{obj}(1|0) = 0.4$. Then approximately 40% of people would receive signal 1 and the rest receive signal 0. Suppose that all agents have the same subjective likelihood function $f_i(\cdot|0) = f_{obj}(\cdot|0)$, which happens to be the same as the objective likelihood, and that $f_i(1|1) = 0.8$, then those who receive signal 1 would have subjective posterior probability

$$p_i(1|1) = \frac{0.6 * 0.8}{0.6 * 0.8 + 0.4 * 0.4} = 0.75,$$

while those who receive signal 0 would have subjective posterior probability

$$p_i(1|0) = \frac{0.6 * 0.2}{0.6 * 0.2 + 0.4 * 0.6} = 1/3.$$

3.2 *The Estimator in Prelec et al. (2017)*

In contrast to a traditional survey question asking a particular question, an innovative idea in Prelec et al. (2017) is that an additional question is designed to ask respondents to predict the others' answers. For example, if the first survey question is "which one is more likely to happen, $\Omega = 1$ or $\Omega = 0$", then the additional question can be "what is the proportion of the whole population who think $\Omega = 1$ is more likely to happen".

Assume that giving a signal $S_i = 1$ agent i will report $\Omega = 1$. Then for agent i the answers to the above two questions are equivalent to $\mathbf{1}_{\{S_i=1\}}$ and $\mathbb{P}(S_j = 1|S_i)$, respectively. Therefore, the survey designer can estimate $\mathbb{P}_{obj}(S_i = 1|\Omega_{obj})$ and $(M_{kl})_{k,l=0,1} = \mathbb{P}(S_j = l|S_i = k)$, known as metaknowledge matrix. An important observation made by Prelec and Seung (2006) and Prelec et al. (2017) is that $M_{kl}\mathbb{P}(S_i = k) = M_{lk}\mathbb{P}(S_i = l)$. Therefore, the ratio of two marginal probabilities $\mathbb{P}(S_i = k)/\mathbb{P}(S_i = l)$ can be estimated from M without knowing prior distribution or likelihood function via

$$\frac{\mathbb{P}(S_i = k)}{\mathbb{P}(S_i = l)} = \frac{M_{lk}}{M_{kl}}. \tag{8}$$

To summarize, we denote agent i's answer to the first question by $v_i \in \{0, 1\}$ and answers to the second question by $\xi_i^k = \mathbb{P}(S_j = k|S_i)$, $k = 0, 1$.

Now we make another assumption on the posterior distribution

$$\mathbb{P}(\Omega = 1|S_i = 1) > \mathbb{P}(\Omega = 0|S_i = 1), \quad \mathbb{P}(\Omega = 0|S_i = 0) > \mathbb{P}(\Omega = 1|S_i = 0).$$

Note that this assumption is a mild one, as it is neither about the prior distribution nor about the likelihood. From this assumption, we get

$$\mathbb{P}(\Omega = 1|S_i = 1) > \frac{1}{2} > \mathbb{P}(\Omega = 1|S_i = 0), \quad \mathbb{P}(\Omega = 0|S_i = 0) > \frac{1}{2} > \mathbb{P}(\Omega = 0|S_i = 1).$$

Therefore,

$$\mathbf{1}_{\left\{\mathbb{P}(\Omega=\Omega_{obj}|S_i=1)>\mathbb{P}(\Omega=\Omega_{obj}|S_i=0)\right\}}$$

equals 1 if $\Omega_{obj} = 1$, and 0 otherwise. Hence the quantity of interest, which is $\mathbf{1}_{\{\Omega_{obj}=1\}}$, can also be written as

$$\mathbf{1}_{\left\{\mathbb{P}(\Omega=\Omega_{obj}|S_i=1)>\mathbb{P}(\Omega=\Omega_{obj}|S_i=0)\right\}} = \mathbf{1}_{\left\{\frac{\mathbb{P}_{obj}(S_i=1|\Omega_{obj})}{\mathbb{P}(S_i=1)} > \frac{\mathbb{P}_{obj}(S_i=0|\Omega_{obj})}{\mathbb{P}(S_i=0)}\right\}}, \tag{9}$$

where we have used the fact that $\mathbb{P}(A|B) = \mathbb{P}(B|A)\mathbb{P}(A)/\mathbb{P}(B)$.

Using (8) the estimator proposed by Prelec et al. (2017) for (9) is

$$\mathbf{1}_{\left\{\frac{\sharp\{i:v_i=1\}}{m_{01}} > \frac{\sharp\{i:v_i=0\}}{m_{10}}\right\}}, \tag{10}$$

where $m_{kl} = \dfrac{1}{\sharp\{i : v_i = k\}} \displaystyle\sum_{i:v_i=k} \xi_i^l$. It is worthwhile pointing out that Prelec and Seung (2006) also construct a more complicated estimator. We refer to (10) as the BTS survey estimator.

Here is an illustration of the BTS estimator. We ask people two questions. (1) Is Philadelphia the capital of Pennsylvania? (2) What is your guess of the percentage of the respondents who answer yes to the first question? Suppose 70% of respondents say yes to the first question; among the people who answer yes to the first question the average of their answers to the second question is 70%, and among the people who answer no to the first question the average of their answers to the second question is 75%. Assume that $\Omega = 1$ means that Philadelphia is the capital of Pennsylvania and $\Omega = 0$ means otherwise. Note that

$$\sharp\{i : v_i = 1\} = 0.7N, \quad \sharp\{i : v_i = 0\} = 0.3N,$$

$$m_{01} = 0.75, \ m_{10} = 0.3.$$

Thus, the estimator is $\mathbf{1}_{\left\{\frac{0.7}{0.75} > \frac{0.3}{0.3}\right\}} = 0$. In short, the estimator indicates that Philadelphia is not the capital of Pennsylvania.

3.3 Market Prices in Prediction Markets

We restrict attention to a one-period prediction market, in which two winner-take-all securities (H and T) are traded, to bet on the outcome of a coin tossing. The holder of one share of security H (T) gets 1 dollar if the head (tail) shows up, and gets nothing otherwise. We first assume that market is an ideal market, e.g. with no transaction costs, no bid-ask spread, and short sale being allowed (unless otherwise specified). The case of transaction costs will be discussed later when we discuss sports betting markets.

Suppose that there are N agents in the market with their subjective probabilities (p_i, q_i), $i = 1, 2, \cdots, N$, where $p_i + q_i = 1$ and $p_i, q_i > 0$. We also assume that the amount of money traded in the prediction market is small enough so that an agent will not link the payoff of the market to his overall investment strategy. In other words, the prediction market is not a hedging market, and the agents only maximize the expected utility within this prediction market. In addition, without loss of generality, the interest rate in this prediction market is zero.

Given the market price $\bar{p}$ for H and $\bar{q}$ for T and initial wealth w_i, agent i needs to determine the number of shares (x_i, y_i) in H and T respectively, to maximize the expected utility, with respect to his subjective probability, namely,

$$\max_{x_i, y_i:\ \bar{p}x_i + \bar{q}y_i = w_i} p_i U_i(x_i) + q_i U_i(y_i). \tag{11}$$

To preclude arbitrage, we need $\bar{p} + \bar{q} = 1$. Also, we can exchange one dollar for one share of H and one share of T, and holding the same shares of two securities is equivalent to investing in a risk-free money market account with zero interest rate.

A competitive equilibrium is defined as a pair $(\bar{p}, \bar{q})$ and (x_i, y_i), $i = 1, \cdots, N$ such that $\bar{p} + \bar{q} = 1$, (x_i, y_i) is an optimal solution to (11), and market clears. Here, the market clearing condition means that the total positions in both assets are the same. That is, $\sum_{i=1}^N x_i = \sum_{i=1}^N y_i$. This clearing condition holds because in the prediction market investors initially buy the unit share consisting of one H and one T and then unbundle either H or T shares via trading in the market. This indicates that the payment to the winners can be fully covered by the total asset value, i.e. $\sum_{i=1}^N x_i = \sum_{i=1}^N y_i = \sum_{i=1}^N (\bar{p}x_i + \bar{q}y_i)$, no matter what the outcome is.

Dai et al. (2021a) find the equilibrium market price and optimal trading positions:

(i) If agent i has utility function $U_i(c) = 1 - \frac{1}{\gamma_i} e^{-\gamma_i c}$, for $1 \le i \le N$, then the equilibrium price $\bar{p}$ exists and is the unique solution to

$$\sum_{i=1}^N \frac{1}{\gamma_i} \log \frac{p_i}{1-p_i} = \log \frac{\bar{p}}{1-\bar{p}} \sum_{i=1}^N \frac{1}{\gamma_i}, \tag{12}$$

and optimal strategy (x_i, y_i) for agent i satisfies

$$x_i - y_i = \frac{1}{\gamma_i}(\log \frac{p_i}{1-p_i} - \log \frac{\bar{p}}{1-\bar{p}}). \tag{13}$$

(ii) If agent i has utility function $U_i(c) = \frac{c^{1-\gamma_i}}{1-\gamma_i}$ and initial wealth w_i, for $1 \le i \le N$, then the equilibrium price $\bar{p}$ exists and is the unique solution to

$$\sum_{i=1}^N \frac{1 - (\frac{\bar{p}(1-p_i)}{(1-\bar{p})p_i})^{\frac{1}{\gamma_i}}}{[(\frac{\bar{p}(1-p_i)}{(1-\bar{p})p_i})^{\frac{1}{\gamma_i}} - 1]\bar{q} + 1} w_i = 0, \tag{14}$$

and optimal strategy (x_i, y_i), $x_i > 0$ and $y_i > 0$, for agent i satisfies

$$\log x_i - \log y_i = \frac{1}{\gamma_i}(\log \frac{p_i}{1-p_i} - \log \frac{\bar{p}}{1-\bar{p}}). \tag{15}$$

If all agents are in continuum type, then the above results coincide with Wolfers and Zitzewitz (2006) and Ottaviani and Sørensen (2015). Here we focus on finite number of agents, because our goal is to develop a statistical estimator based on market prices. The above results indicate that the trading position of a risk averse agent depends on subjective probability, and becomes more aggressive when the subjective probability deviates further from market price. Therefore, the net position of an agent contains additional information about the subjective probability. If all agents have CRRA preference, it is remarkable that short selling and leveraging are always not optimal.

3.4 A Market-Adjusted Estimator in Dai et al. (2021a)

In a prediction market, risk averse agents have to evaluate the chance of the future event to make an optimal investment, so that the subjective probabilities may be embedded in trading positions, which overcomes the difficulty in surveys. Dai et al. (2021a) propose a market-adjusted estimator defined as

$$\hat{\Omega}_{obj,market} = \mathbf{1}_{\{\hat{p}_N > \frac{1}{2}\}},$$

and $\hat{p}_N$ can be computed via

$$\frac{1}{N} \log \frac{\hat{p}_N}{1-\hat{p}_N} = \log \frac{\bar{p}}{1-\bar{p}} + \frac{\bar{\gamma}}{N} \sum_{i=1}^{N} z_i, \tag{16}$$

where $\sum_{i=1}^{N} z_i = 0$ for CARA utility, $\sum_{i=1}^{N} z_i = \sum_{i=1}^{N} (\log x_i - \log y_i)$ for CRRA utility,

$$\bar{\gamma} := \frac{1}{\mathbb{E}_{obj}[\frac{1}{\gamma_i}]}.$$

The estimator $\hat{p}_N$ is based on market price and trading positions, with proper information on the average risk aversion coefficient and average prior distribution. The market price arises in our estimator as a natural way to aggregate subjective probabilities. The second term consists of both average risk aversion (harmonic average) and average position (arithmetic or geometric average) as an adjustment for risk preferences. Indeed, the equilibrium market price can be driven by agents who are wealthy and less risk averse. Therefore, to obtain a consistent estimator, it is necessary to consider correcting the bias caused by risk aversion and heterogeneous wealth effect.

The second term vanishes if agents have CARA utility functions but remains nonzero in general if agents have CRRA utility functions. Note that for CARA utility functions

$$\frac{\hat{p}_N}{1-\hat{p}_N} = \left(\frac{\bar{p}}{1-\bar{p}}\right)^N,$$

where the market equilibrium price $\bar{p}$ depends on N.

Note that $\hat{p}_N$ can be interpreted as the "posterior probability" conditioned on all signals, i.e. $\mathbb{P}(\Omega = 1 | S_1, \cdots, S_N)$, with a suitable prior distribution and the likelihood function. However, this is not the standard posterior in Bayesian statistics. In fact, under the objective probability, the joint distribution of Ω and $S_1, \cdots, S_N$ is the product probability between a point mass Ω_{obj} and the joint objective likelihood function $\prod_{i=1}^{N} f(S_i | \Omega_{obj})$. Thus, such "posterior probability" can be regarded as the subjective probability of a fictitious agent who can observe all signals and whose prior is the population average.

Under suitable regularity conditions, Dai et al. (2021a) prove that the market-adjusted estimator using (16) is consistent, i.e.

$$\hat{\Omega}_{obj,market} \to \Omega_{obj}, \quad \mathbb{P}_{obj}\text{-a.s., as } N \to \infty.$$

Moreover, a central limit theorem for the market-adjusted estimator is also provided, yielding the convergence rate of the consistent estimator

It is difficult to elicit people's risk preference and probabilistic beliefs simultaneously under our framework, since an observed individual action is the joint consequence of his/her belief and risk preference. Here the average risk aversion is needed as an input. In the empirical implementation, we can take a typical value (e.g. from 1 to 3) used by the existing literature. The robustness of such choice is confirmed empirically. Dai et al. (2021a) also gives a proposition to provide a theoretical justification of the robustness by showing that a wrong average risk aversion $\tilde{\gamma}$ may still lead to a consistent estimator.

3.5 An Extension with Transaction Fees in Sports Betting Markets

The sports betting market is a multi-billion business that is legalized in some states, e.g. Nevada thanks to the Professional and Amateur Sports Protection Act of 1992. Birge et al. (2018) study this market from the perspective of bookmakers. To demonstrate the efficiency of the estimator empirically, the data set Dai et al. (2021a) used is downloaded from Pregame.com. The website, run by a professional sports betting media company, has betting records of the U.S. sports games such as NFL, college football, NBA, MLB, and NHL. The records for each game include time series about dollar amount of cash bet, number of participants, and point spreads.

In such a betting market, bookmakers charge large proportional transaction costs. For example, bookmakers usually offer an 11–21 payout for a game between two teams X and Y, which means if one bets 11 dollars and wins, then he gets 21 dollars back, earning 10 dollars; and if he loses, then he gets nothing back. Moreover, bookmakers provide liquidity by clearing the market. Hence, the market differs from the frictionless, equilibrium prediction market assumed in the previous standard model. As a result, we have to extend the standard model.

Without loss of generality, consider the betting market with 11–21 payout as given above. Let $R = \frac{10}{11}$. Given the initial endowment w_i, agent i decides how much money to bet by solving the following expected utility maximization problem

$$\max_{(x,y)\in K} p_i U_i(w_i - y + Rx) + q_i U_i(w_i - x + Ry),$$

where the interest rate is assumed to be 0, and x and y are the dollar amounts bet for teams X and Y, respectively. Here the optimization is done for a CARA type utility over the set $K = \{x \geq 0, y \geq 0\}$, and for a CRRA type utility $K = \{x \geq 0, y \geq 0, w_i - y + Rx \geq 0, w_i - x + Ry \geq 0\}$.

The prediction scores are defined by

$$\begin{cases} score^X = \bar{\gamma}(1+R) \sum\limits_{i \text{ bets on team } X} x_i - \log R \, \sharp\{i : i \text{ bets on team } X\} \\ score^Y = \bar{\gamma}(1+R) \sum\limits_{i \text{ bets on team } Y} y_i - \log R \, \sharp\{i : i \text{ bets on team } Y\} \end{cases} \tag{17}$$

for CARA utility, and

$$\begin{cases} score^X = \bar{\gamma} \sum\limits_{i \text{ bets on team } X} \log \dfrac{1 + Rx_i/w_i}{1 - x_i/w_i} - \log R \, \sharp\{i : i \text{ bets on team } X\} \\ score^Y = \bar{\gamma} \sum\limits_{i \text{ bets on team } Y} \log \dfrac{1 + Ry_i/w_i}{1 - y_i/w_i} - \log R \, \sharp\{i : i \text{ bets on team } Y\} \end{cases} \tag{18}$$

for CRRA utility, where x_i and y_i are the amounts of cash that agent i bets on team X and Y, respectively, and w_i is agent i's initial endowment. To be more precise, X is predicted to win if $score^X > score^Y$, and to lose if otherwise.

Dai et al. (2021a) prove that under suitable conditions the estimator is consistent, namely

$$\mathbf{1}_{\{score^X > score^Y\}} \to \mathbf{1}_{\{\text{team } X \text{ wins the bet}\}}, \quad \mathbb{P}_{obj}\text{-a.s.},$$

as the size of participants $N \to \infty$. The intuition of the prediction scores is as follows: If one bets more, then he contributes more to the corresponding prediction score. There are two terms in each prediction score: The first term reflects the amount of money bet and the second term the number of participants. Clearly, our estimator accounts for the trade-off between these two, weighted by the risk aversion coefficient.

3.6 An Empirical Study

We collect the side betting data of NBA basketball games in season 2018–2019 available at Pregame.com, where one bets on whether the difference between two teams is larger than a given number set by the bookmaker. First of all, we need to define the meaning of winning the bet. The difference in points is called the point spread, which is determined by the bookmaker and can be adjusted dynamically. For example, if the point spread for the home team is $+2$, then people who bet on the home team will win if home team point $+2 >$ away team point, lose if home team point $+2 <$ away team point, and get their money refunded if home team point $+2 =$ away team point. At the same time, the point spread for the away team is -2.

The data consists of betting records at different time stamps of 1312 NBA game matches. At each time, data records current point spread (point spread), accumulated amount of money bet (CASH), and accumulated number of people (TICKET) for away team and home team, respectively. On average, there are about 500 participants who bet around 100,000 dollars in total for each match. In the data set, records for away team and home team are not always synchronized, so we first match two groups of data at synchronized time stamps, resulting in a total of 101,332 records for full sample. Note that there are not many participants for synchronized data in most cases, for example, there are only about 150 participants in more than 75% cases and there are only 20 participants in 25% cases. Thus, later we have to focus on the cases with many participants to ensure market liquidity so that our market based estimator can work well.

Ideally, to implement (17) and (18), we need to observe each agent's betting amount. However, we only have accumulated betting records at different timestamps. Thus, Dai et al. (2021a) make additional assumptions to approximate the quantities of individual's betting amount, so that the estimator can be constructed dynamically at any synchronized timestamps.

The performance will be evaluated by two criteria, the accuracy of bet based on our estimator and the return of resulting betting portfolio. To ensure fairness, we test our results by randomly bootstrapping 1000 samples in data set with replacement and repeating this procedure for 1000 times to compute the mean, standard deviation, and quantiles. We need to focus on those cases with a reasonably large number of participants, to ensure there is enough liquidity for the market equilibrium to work. Therefore, we select subsets of samples with at least 200, 400, or 800 participants, respectively.

The prediction accuracy is reported in Table 1. Note that the prediction accuracy is significantly improved for our estimator as the number of participants increases, while there are no such patterns for the majority following strategy. For example, when the number of participants is at least 400, the prediction accuracy is at least 55.7%.

Betting return is presented in Table 2. Our betting strategy is to bet 11 dollars on the team with the higher estimator, and no bet if the two estimators are the same. When the number of participants is at least 400, our return is at least 6.4% with Sharpe ratio at least 2.2, as reported in Table 2. Return is also significantly raised for our estimator as the number of participants increases, in contrast to the naive majority cash following strategy. It is also worthwhile pointing out that we achieve such a good result because a small improvement in prediction accuracy can lead to a substantial amount of profit.

Table 1 The summary statistics of prediction accuracy of sports betting. We sample $n = 1000$ records in our data set conditioned on different numbers of participants. And we repeat this procedure for $B = 1000$ times to calculate standard deviation and quantile. In the implementation, $w = 1.5$

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
(a) Panel A: conditioned on the total participation number $N > 200$ *with 18,140 observations*						
major.cash	0.514	0.016	0.468	0.502	0.524	0.569
crra $\bar{\gamma} = 1$	0.533	0.015	0.484	0.523	0.544	0.581
crra $\bar{\gamma} = 2$	0.532	0.015	0.484	0.522	0.542	0.586
crra $\bar{\gamma} = 3$	0.540	0.016	0.498	0.530	0.551	0.583
(b) Panel B: conditioned on the total participation number $N > 400$ *with 5741 observations*						
major.cash	0.497	0.016	0.439	0.486	0.507	0.542
crra $\bar{\gamma} = 1$	0.557	0.015	0.511	0.546	0.568	0.598
crra $\bar{\gamma} = 2$	0.561	0.015	0.514	0.551	0.570	0.613
crra $\bar{\gamma} = 3$	0.569	0.015	0.518	0.558	0.580	0.618
(c) Panel C: conditioned on the total participation number $N > 800$ *with 1177 observations*						
major.cash	0.520	0.015	0.472	0.510	0.530	0.567
crra $\bar{\gamma} = 1$	0.578	0.015	0.537	0.568	0.588	0.634
crra $\bar{\gamma} = 2$	0.596	0.015	0.542	0.586	0.606	0.642
crra $\bar{\gamma} = 3$	0.612	0.015	0.562	0.602	0.622	0.664

There are three things worth noting here: (1) The naive majority following strategy yields negative return; this is not surprising as there is a significant amount of transaction fee paid to the bookmaker and a random guess would lead to 5% loss on average. (2) Our betting strategy has fewer betting opportunities, as the number of time stamps is 5741 out of 101,332 when the number of participants is above 400. (3) Our focus here is to demonstrate the prediction accuracy, not to make real money even with high Sharpe ratios. This is mainly due to the limit size of the market, with average only 100,000 dollars in total for each NBA game match.

Although the estimator is developed theoretically from a static model, in practice it can be used to generate dynamic forecasting, as illustrated in our empirical test. To make a prediction, the estimator based on adjusted market prices only need to aggregate current judgment from a group of people and do not use any historical data. This is different from forecasting using time series model or other econometric models that rely on historical data.

It is an interesting open problem to compare empirically the performance of our estimator based on adjusted market prices and conventional time series estimators. We should point out that there is an extensive literature comparing predictions based on unadjusted market prices and time series models, especially in the case of political elections; see, e.g., the references in Kou and Sobel (2004).

Table 2 The summary statistics of returns in sports betting market using different betting strategies. We sample $n = 1000$ records in our data set conditioned on different numbers of participants, and bet 11 dollars each time. And we repeat this procedure for $B = 1000$ times to calculate standard deviation and quantile. In the implementation, $w = 1.5$

Statistic	Mean	St. Dev.	Sharpe ratio	Min	Pctl(25)	Pctl(75)	Max
(a) Panel A: conditioned on the total participation number $N > 200$ with 18,140 observations							
major.cash	−0.019	0.030	−0.633	−0.106	−0.040	0.001	0.087
crra $\bar{\gamma} = 1$	0.019	0.029	0.655	−0.076	−0.001	0.038	0.110
crra $\bar{\gamma} = 2$	0.017	0.029	0.586	−0.089	−0.002	0.036	0.133
crra $\bar{\gamma} = 3$	0.033	0.030	1.100	−0.049	0.013	0.053	0.114
(b) Panel B: conditioned on the total participation number $N > 400$ with 5741 observations							
major.cash	−0.051	0.030	−1.700	−0.161	−0.072	−0.031	0.035
crra $\bar{\gamma} = 1$	0.064	0.029	2.207	−0.024	0.043	0.084	0.142
crra $\bar{\gamma} = 2$	0.071	0.029	2.448	−0.018	0.053	0.090	0.171
crra $\bar{\gamma} = 3$	0.087	0.029	3.000	−0.009	0.066	0.108	0.180
(c) Panel C: conditioned on the total participation number $N > 800$ with 1177 observations							
major.cash	−0.005	0.029	−1.724	−0.097	−0.025	0.014	0.084
crra $\bar{\gamma} = 1$	0.105	0.028	3.750	0.027	0.086	0.123	0.211
crra $\bar{\gamma} = 2$	0.139	0.029	4.793	0.035	0.119	0.158	0.226
crra $\bar{\gamma} = 3$	0.160	0.029	5.517	0.076	0.151	0.189	0.268

4 Conclusion

At the intersection of finance, technology, and statistics, FinTech econometrics is a fast-growing research field. In this paper, I only scratch the surface of FinTech econometrics by reviewing two research problems related to information flows, namely privacy issues and how to incorporate additional market information at the individual level in the problem of the wisdom of the crowd.

Technology has made a significant impact on finance. For example, the introduction of credit cards in 1950 and ATM machines in 1967 by banks fundamentally changed consumer finance. However, these technology developments in the past were mainly driven by internal changes. What is interesting about the current FinTech development is that it brings outside information technology tailored to the need of finance, rather than internal changes within the financial system. In particular, technology companies outside of finance, such as Google and Facebook, play an important role in the current FinTech development, leading to interesting interdisciplinary research problems.

Bear in mind two attractive features that FinTech has, namely interesting non-trivial academic research problems and industrial jobs for students. These two might well make FinTech the future financial engineering 2.0.

References

Abbe, E. A., Khandani, A. E., & Lo, A. (2012). Privacy-preserving methods for sharing financial risk exposures. *The American Economic Review: Papers and Proceedings, 102*, 65–70.

Acquisti, A., Brandimarte, L, & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science, 347*, 509–514.

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature, 54*, 442–492.

Babich, V., & Birge, J. (2021). Interface of operations and finance: A tutorial. In V. Babich, J. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations*. Springer.

Babich, V., & Hilary, G. (2020). Distributed ledgers and operations: What operations management researchers should know about blockchain technology. *Manufacturing & Service Operations Management, 22*, 223–228.

Banaloh, J. (1986). Secret sharing of homomorphisms: Keeping shares of a secret sharing. In *CRYPTO 1986*.

Birge, J. R., Feng, Y., Keskin, N. B., & Schultz, A. (2018). Dynamic learning and market making in spread betting markets with informed bettors. *Available at SSRN 3283392*.

Cai, N., & Kou, S. (2019). Econometrics with privacy preservation. *Operations Research, 67*, 905–926.

Cao, Y., Dai, M., Kou, S., Li, L., & Yang, C. (2018). Designing stable coins. *Working Paper*.

Capponi, A., Olafsson, S., & Zariphopoulou, T. (2020). Personalized robo-advising: Enhancing investment through client interaction. *Working Paper*.

Dai, M., Jia, Y., & Kou, S. (2021a). The wisdom of the crowd and prediction markets. *Journal of Econometrics, 222*, 561–578.

Dai, M., Jin, H., Kou, S., & Xu, Y. (2021b). A dynamic mean-variance analysis for log returns. *Management Science, 67*, 1093–1108.

Dai, M., Jin, H., Kou, S., & Xu, Y. (2021c). Robo-advising: A dynamic mean variance approach. *Digital Finance*, 3, 81–97.

Duchi, J. C., Jordan, M. I., & Wainwrigh, M. J. (2014). Privacy aware learning. *Journal of the ACM, 61*, 1–57.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* (pp. 265–284).

Fan, J., Li, R. L., Zhang, C.-H. Z., & Zou, H. (2020). *Statistical foundations of data science*. Chapman & Hall/CRC.

Galton, F. (1907). Vox populi. *Nature, 75*, 450–451.

Garratt, R. (2016). CAD-coin versus Fedcoin.*R3 Report*.

Gu, S., Kelly, B., Li, L., & Xu, D. (2018). Empirical asset pricing via machine learning. *Working Paper*.

Harvey, C. R. (2016). Cryptofinance. *Working Paper*.

Horel, E., & Giesecke, K. (2020). Significance tests for neural networks. *Journal of Machine Learning Research, 21*, 1–29.

Hou, A. J., Wang, W., Chen, C. Y., & Härdle, W. K. (2020). Pricing cryptocurrency options. *Journal of Financial Econometrics, 18*, 250–279.

Karr, A., Li, X., Sanil, A. P., & Reiter, J. P. (2005). Secure regression on distributed database. *Journal of Computational and Graphical Statistics, 14*, 263–279.

Keppo, J., Phan, T., & Tan, T. (2021). Microbanks in online peer-to-peer lending: A tale of dual roles. In V. Babich, J. Birge, & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations*. Springer.

Kou, S. G., & Sobel, M. E. (2004). Forecasting the vote: a theoretical comparison of election markets and public opinion polls. *Political Analysis, 12*, 277–295.

Lipton, A., Hardjono, T., & Pentland, A. (2018). Pricing cryptocurrency options. *Royal Society Open Science, 5*, 180155.

Ottaviani, M., & Sørensen, P. N. (2015). Price reaction to information with heterogeneous beliefs and wealth effects: Underreaction, momentum, and reversal. *American Economic Review, 105*, 1–34.

Posner, R. A. (1981). The economics of privacy. *The American Economic Review, 71*, 405–409.

Prelec, D., & Seung, H. S. (2006). An algorithm that finds truth even if most people are wrong. *Technical Reports*.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature, 541*, 532–535.

Ritchken, P., & Wu, Q. (2021). The impact of technology choice on capital structure. In V. Babich, J. Birge, and G. Hilary (Eds.), *Innovative technology at the interface of finance and operations*. Springer.

Shamir, A. (1979). Pow to share a secret. *Communications of the ACM, 11*, 612–613.

Strub, M. S., Li, D., & Cui, X. Y. (2020). An enhanced mean-variance framework for robo-advising applications. *Working Paper*.

Trimborn, S., & Härdle, W. K. (2018). CRIX an index for cryptocurrencies. *Journal of Empirical Finance, 49*, 107–122.

Wolfers, J., & Zitzewitz, E. (2006). *Interpreting prediction market prices as probabilities*, Technical Report, National Bureau of Economic Research.

Yao, A. C.-C. (1982). Protocols for secure computations. *FOCS, 82*, 160–164.

Yermack, D. (2017). Corporate governance and blockchains. *Review of Finance, 21*, 7–31.

The Impact of Technology Choice on Capital Structure

Peter Ritchken and Qi Wu

1 Introduction

Operational flexibility embedded in technology provides an important hedge against changes in the state of markets and is a topic of substantial interest. Operational flexibility includes the ability to switch use of raw materials, adjust product mixes, change product lines, or alter production volumes. The degree of flexibility inherent in the technology affects the decision as to when to purchase, and it simultaneously affects the mix of equity and debt used to fund the investment. This chapter investigates the investment timing and financing choice with the goal of identifying how these decisions are affected as the technology becomes more flexible.

Since the seminal work of Modigliani and Miller (1958) that proved in perfect capital markets corporate financing and investment decisions were separable, much attention has focused on the role market imperfections play in the simultaneous design of capital structures and in the execution of investment decisions. With taxes and deadweight bankruptcy costs, this separation no longer exists, and the timing and financing of investments have to be done simultaneously. Flexibility, as an operational characteristic in the investment should lead to a change in investment and financing decisions.

P. Ritchken (✉)
Weatherhead School of Management, Case Western Reserve University, Cleveland, OH, USA

African Institute of Financial Markets and Risk Management, University of Cape Town, Cape Town, South Africa
e-mail: phr@cwru.edu

Q. Wu
Weatherhead School of Management, Case Western Reserve University, Cleveland, OH, USA
e-mail: qxw132@case.edu

V. Babich et al. (eds.), *Innovative Technology at the Interface of Finance and Operations*, Springer Series in Supply Chain Management 13,
https://doi.org/10.1007/978-3-030-81945-3_11

Specifically, we analyze the interaction between investment and financing decisions where the firm with a flexible technology has the ability to dynamically control production decisions of assets in place, while the firm with a rigid technology has limited ability to affect the revenue stream. Both firms have expansion options that can be enacted at any time and can be financed with debt and equity. The expansion options are on investments that also have different degrees of operational flexibility.

Firms with greater operational flexibility are more able to manage costs in distressed markets. In contrast, firms with more rigid, inflexible technologies have less operational flexibility and with less tools to manage risk in market downturns might respond by taking on less initial debt (Mauer & Triantis, 1994). On the other hand, greater operational flexibility provides managers with more tools to utilize and the potential lenders of funds to the firm may be concerned that these tools will be used in ways that will not benefit their interests. As a result, they may rationally charge more for loans, and this may feedback into the firm choosing to use less debt (Mello & Parsons, 1992). Some empirical support for this risk shifting behavior is provided by MacKay (2003) and Eisdorfer (2008), but there are conflicting results (Kuzmina, 2013; Reinartz & Schmid, 2016). Our fundamental focus here is to review a framework for analyzing these types of problems and to summarize recent theoretical results which provide unambiguous results on how volume flexibility affects timing and capital structure decisions.

The framework that we provide follows the real option approach that analyzes investment under uncertainty. The real options approach posits that the opportunity to invest in a project is analogous to an American call option on an investment opportunity. Given the connection, the vast machinery of option pricing theory becomes available. An excellent resource of this approach is Dixit and Pindyck (1994). Perhaps the most well known result in this literature is the invalidation of the rule of investing in any project with a positive net present value. Since the future value is uncertain, there is an opportunity cost of investing right away. This is referred to as the option to wait. The corrected decision rule therefore is only to invest when the asset value exceeds the investment cost by a potentially large option premium (McDonald & Siegel, 1986). The value of waiting for more information increases as the uncertainty of the revenue stream increases.

Much of the operations literature on timing and financing of large irreversible projects typically assumes all-equity financing. In practice, however, large irreversible investments are funded with debt. Elsas et al. (2006), for example, study how firms actually pay for these large investments, and find that most are externally financed, with new debt providing at least half the required funds in the year of the investment, and less than 20% are financed with equity. Therefore, establishing how the use of debt alters investment timing decisions, and the value of waiting, is of interest. With capital market imperfections, the funding of large investments interacts with the nature of the underlying technology, with its embedded flexibility, to influence the timing of the expansion. A key feature of real option models is that they allow for the cost of borrowing in a creditor-firm equilibrium to be endogenous.

We follow *the tradeoff theory* developed in classical corporate finance literature/textbooks (e.g., see Section 16.4 of Berk & DeMarzo, 2007). This theory is framed in an environment that trades off the *tax benefit of debt* against *deadweight bankruptcy costs* where all decisions in the firm are made so as to optimize the value of equity. It is important to note that the investment timing and the financing should be done contemporaneously. If one constrains the investment time or equivalently ignores the investment timing decision, any result on the financing decision would be incomplete. In theory, we could use n-period or infinite time horizon discrete time models. However, a discrete time n-period model would not likely lead to clean analytical solutions. For the same reason, almost all existing real option and dynamic capital structure models that we are aware of that involve optimal timing decision are built off continuous time models.

With debt on the balance sheet, equityholders may invest earlier than when they would invest if the funding was with equity. This is referred to as the under-investment problem. It also is possible, that with debt in place, equityholders defer investments. This is the over-investment problem. We will see that the under- or over- investment is very much dependent on the degree of operational flexibility. Indeed, in our setting, we will find that greater operational flexibility mitigates to some extent these issues.

We also will consider a firm with an existing capital structure in place, that is considering a second round of debt financing as part of the funding mechanism for an expansion. In this case, conflicts of interest arise between the shareholders and the existing bondholders with possibilities emerging that the shareholders will time their investment and choose debt financing so as to maximize their welfare which is not necessarily best for the firm. That is, the shareholder strategies may expropriate wealth from the existing bondholders if it serves their interest. Ex ante the original bondholders recognize this problem, and they charge a fair credit spread to reflect these future rational decisions made by equityholders. So ultimately, equityholders bear this cost. The resulting timing and financing decisions that are made when shareholders maximize their welfare are termed *second best* solutions since they do not maximize firm value. Decisions made on the basis of firm maximization are termed *first best*. However, it typically is not possible to achieve first best outcomes because managers, assumed to carry out the policies of the owners (i.e. shareholders) will focus on equity maximization. Our setting is rich enough that we can explore the magnitude of these distortions created by these conflicts of interest and examine how they are affected by operational flexibility.

Traditional capital structure literature studies the tradeoff between debt and equity financing. A much deeper issue is why these securities exist in the first place. The financial contracting literature addresses this question. In our setting, perhaps contracts can be designed so that first best policies can indeed be accomplished. In most settings it is not possible for the creditors to be able to fully control the actions of the equityholders. Indeed, flexible technology provides managers with tools they can use, such as stopping and restarting operations, conducting maintenance, and impacting quality, that would be very difficult to monitor and hence contract on. As a result, a great deal of technological flexibility empowers the equityholders with possible mechanisms of counteracting contracts. Creditors, aware of this possibility, again, will demand higher credit spreads for their loans. But perhaps there are financial mechanisms whereby financial contracts could be designed to mitigate these effects and to accomplish first best policies. We illustrate this idea and show how such contracts could be designed. The idea is not really to promote the use of our solution. Indeed, we do not observe contracts with our proposed design that are in regular use. Rather, and indeed consistent with this entire chapter, our goal is to illustrate how the design process can be identified and made operational.

Throughout, we assume the firm is a price taker, unable to influence its value. This assumption is not insignificant. The option to wait depends critically on the lack of competition. If firms fear preemption, then the option to wait becomes less valuable.[1]

This chapter proceeds as follows. In Sect. 2 we provide a brief literature review, In Sect. 3 we very formulate typical continuous time investment problems in a contingent claims framework, and illustrate the basic pricing mechanism with two simple problems that highlight stopping and the value of the option to wait. Section 4 is devoted to investigating how growth options are affected by market imperfects and by the nature of technology. We consider optimal financing and investment decisions when the technology is inflexible, then repeat the analysis for a technology with flexibility and compare results and draw some conclusions. In Sect. 5 we add to the problem by considering the impact of an existing capital structure. In this case future decisions are impacted by the magnitude of debt on the current balance sheet. When the expansion option is exercised, the balance sheet consists of two tranches of debt. The impact of operational flexibility on capital structure in this dynamic setting is explored. We compare first best with second best solutions and see how the nature of technology impacts agency costs. We investigate whether the absolute priority rule used in bankruptcy is the source for our main differences attributable to flexibility, and conclude that changing the rule has little impact. We investigate whether financial contracts can be constructed to induce equityholders to implement first best solutions. In Sect. 6 we extend our model by allowing ongoing fixed costs to exist. Prior to this, all costs were variable costs. However, different technologies may involve different degrees of fixed versus variable costs. How different degrees of operational leverage, combined with operational flexibility affect financial decisions is explored. All the extensions of the basic model in Sect. 3 lead to similar consistent conclusions. Section 7 concludes.

2 Literature Review

Our modeling approach follows Ritchken and Wu (2021) and is also closely related to Hackbarth and Mauer (2012). The latter extend the original Leland (1994) model in several very interesting dimensions. In particular they begin with a firm whose

[1]For discussions of investments in industries with competitive pressure, a game theoretic analysis as in Grenadier (2002), becomes essential. For further discussions see Huberts et al. (2015).

existing capital structure influences future investment decisions in that there is stockholder- bondholder conflict not only over the *timing* of the future investment but also over the *financing* of the new investment. The new financing introduces the possibility of a new tranche of debt causing additional agency issues. In their analysis, the assets in place, as well as the growth options, are based on a technology that has no operating flexibility whatsoever. As a result, they cannot study how technological flexibility affects timing and financing decisions, but they can investigate financial mechanisms for better aligning the interests of creditors and equityholders when there is no operational flexibility. Ritchken and Wu (2021) extend Hackbarth and Mauer (2012) by incorporating operational flexibility into the assets. They also allow for negative cash flows. Since most firms encounter negative cash flows prior to bankruptcy, this extension adds significant realism. This chapter emphasizes and summarizes Ritchken and Wu (2021), and also provides additional extensions including investigating the role operating leverage plays when investigating operating flexibility in investments, the impact of operational flexibility when considering alternative priority rules associated with different tranches of debt in bankruptcy, and discussions of how to design contracts that provide first best outcomes. This involves the use of putable bonds.

Our modeling framework also relates to Mello and Parsons (1992) who extend Brennan and Schwartz (1985) by including operational and capital structure issues for a commodity-based firm in a competitive market. However, these models focuses on techniques that add significant realism into the model. Similarly Mauer and Triantis (1994) analyze the interaction among a firm's dynamic investment, operating, and financing decisions in a model with operating adjustment and recapitalization costs, but due to the complexity, they are unable to get any clean theoretical results, and hence their economic insights are based on numerical simulations.

Like Hackbarth and Mauer (2012), our model could easily include transaction costs of issuing debt and equity. The introduction of fixed costs has been important. Without fixed costs, continuous time trade off models will continuously readjust leverage so as to maintain the delicate balance between high leverage, with its advantageous tax benefits and bankruptcy, with its large deadweight costs. From the empirical literature, however, it has been observed that large liquid profitable firms with low bankruptcy costs do not issue as much debt as these trade off models suggest.[2] However, with the existence of large fixed costs, continuous leverage adjustments are not possible, and rebalancing towards a target is done infrequently, or at times when larger investments are required.[3]

Within the operations literature, several studies focus on inventory and production decisions under financial constraints. Birge and Xu (2011), for example, study joint production and financing decisions of a capital-constrained firm in the presence of demand uncertainty. Li et al. (2013) use a dynamic model to maximize the

[2]See Myers (1993), Graham (2000) and the literature review of Frank and Goyal (2005).

[3]Examples, highlighting this include Goldstein et al. (2001), Strebulaev (2007), and Frank and Goyal (2014).

expected value of dividends by simultaneously choosing operational and financial decisions. Many studies consider various interactions between the firm and suppliers taking into account rational behavior by financial institutions lending funds. Lederer and Singhal (1994) highlight the importance of jointly considering financing and technology choices when making manufacturing investments. Fine and Freund (1990) demonstrate how the capacity-pooling benefit of flexible technology acts as an effective hedge against demand uncertainty when there are no financial constraints. Recent studies have incorporated budget constraints (Boyabatli & Toktay, 2011; Boyabatli et al., 2016). Chod and Zhou (2014) examine how investment in the capacity of flexible and nonflexible resources is affected by financial leverage, and conversely, how a firm's resource flexibility affects the firm's optimal capital structure. They demonstrate that capacity flexibility can mitigate agency issues between shareholders and bondholders.[4] Our study deviates from this interesting field in that we focus on volume flexibility rather than product flexibility.

In many studies where production volume decisions are made, the manufacturer, assumed to be a monopolist facing a downward sloping demand curve, determines the capacity, and once it is installed, determines the ongoing production quantities. Dangle (1999) considers such a problem when the firm is an all-equity firm. Hagspiel et al. (2016) compare the timing and investment strategies of two all-equity firms that are identical in all aspects except one has greater ability to adjust production volume and conclude that when demand is sufficiently uncertain, greater operational flexibility results in delaying investment. Our study also compares two firms with different degrees of operating flexibility. Similar to their study, we also examine timing decisions, but we focus on how the optimal mixes of equity and debt used to finance the investment alter the timing of the investment.

Operating flexibility is fundamentally different from *operating leverage*. Operational leverage is defined as the proportion of the fixed costs to total costs. In contrast, our operational flexibility refers to the ability to control production quantities. In particular, for the firm using flexible technology, costs are incurred only if they can be covered by revenues, which implies that operating leverage fluctuates over time. A large literature exists on the impact of operating leverage on capital structure. The operating leverage trade-off hypothesis (Van-Horne, 1977) suggests that firms with low operating leverage could display capital structures with higher leverage. Empirical studies that provide some support for this include Ferri and Jones (1979) and Mandelker and Rhee (1984). However, other studies, including Huffman (1983), Dugan et al. (1994), and Lord (1996), have provided mixed or even negative findings.[5]

[4]Related studies that focus on the merits of flexibility in dealing with uncertain demand include Aivazian and Berkowitz (1998), Van-Mieghem (1998), Huberts et al. (2015), and the references cited in these works.

[5]Kahl et al. (2014) argue that high fixed cost firms might alternatively benefit from larger cash reserves to hedge their greater risk.

We demonstrate that operational flexibility can impact the relationship between operational leverage and financial leverage and could be a factor that explains the seemingly inconsistent findings in the operational leverage literature. Specifically, increasing fixed costs for firms with inflexible technologies is analogous to adding in a fixed coupon bond obligation. As a result, increasing operating leverage crowds out debt. However, we demonstrate that increasing fixed costs in a firm with a flexible technology does not compete with debt issuance and in fact may encourage more use of debt. As a result, the link between operating leverage and financial leverage depends very critically on the degree of operating flexibility.

Operational flexibility, as studied in this paper, should also be distinguished from reversible investment flexibility. A firm with a flexible investment opportunity set (e.g., a firm trading financial assets) can easily increase or decrease its investments. The option to increase or decrease investment induces skewness in the distribution of future cash flows (see, e.g., Titman et al., 2004). In our model, once growth options are exercised, the investments are *irreversible*. In this regard, except for the timing of the investment, there is no other source of investment flexibility. So, for example, our framework precludes the sale of assets outside bankruptcy.

3 Contingent Claim Formulation and Risk Neutral Measure

Consider a firm that produces a commodity. The firm is a price taker and the output price, x_t follows the process

$$dx_t = \alpha x dt + \sigma x dw_t \tag{1}$$

where α and σ are the drift and volatility terms and dw_t is a standard Wiener increment. The commodity provides a convenience yield, q, proportion to the price. For the moment, assume the firm has no operational flexibility, in that once production starts, it continues indefinitely, unless the equityholders decide to default. Let $F(x)$ be the value of any financial claim such as equity, debt, or the value of the levered firm itself. If the market is sufficiently complete such that the commodity risk can be fully hedged, then, to avoid riskless arbitrage opportunities, each one of these claims must satisfy the fundamental pricing equation (Merton, 1974):

$$\tfrac{1}{2}\sigma x^2 F_{xx}(x) + (r - q)x F_x(x) - rF(x) + F_t + \pi(x, t) = 0 \tag{2}$$

where subscripts denote partial derivatives and $\pi(x, t)$ is the total cash flow awarded to the claim over the time increment. When the claims are independent of time, then this partial differential equation becomes an ordinary differential equation

$$\tfrac{1}{2}\sigma^2 x^2 F_{xx}(x) + (r - q)x F_x(x) - rF(x) + \pi(x) = 0 \tag{3}$$

Further, if the cash flow is of the form $\pi(x) = CF_0 + hx$, then the solution has the form

$$F(x) = \frac{CF_0}{r} + \frac{h}{q}x + a_1 x^{\alpha_1} + a_2 x^{\alpha_2} \tag{4}$$

$$\alpha_1 = \frac{1}{2} - \frac{(r-q)}{\sigma^2} + \sqrt{\left(\frac{1}{2} - \frac{r-q}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}} > 1$$

$$\alpha_2 = \frac{1}{2} - \frac{(r-q)}{\sigma^2} - \sqrt{\left(\frac{1}{2} - \frac{r-q}{\sigma^2}\right)^2 + \frac{2r}{\sigma^2}} < 0.$$

Notice that, unlike classical dynamic programming formulations, the solution does not depend on the drift term of the underlying statistical process driving uncertainty, nor any explicit discount factor. Indeed the results are free of preferences. The Feynman Kac theorem provides a connection between certain partial differential equations and expectations of stochastic differential equations, and it turns out that the solutions of these claims can be expressed as the riskless discounted values of their expected cash flows computed under an equivalent risk neutralized process.

As an example, if the claim is equity with $\pi(x_t) = x_t$, then if the firm has no debt, the claim is priced as

$$EQ(x_t) = E^Q\left[\int_t^\infty e^{-ru} x_u du\right] \tag{5}$$

where the commodity price evolves under a risk neutral measure, $\mathbb{Q}$, as:

$$dx_t = (r-q)x_t dt + \sigma x_t dw_t, x_0 \text{ given.} \tag{6}$$

Here the total rate of return r consists of price appreciation at rate $r - q$ together with the convenience yield of q.[6]

If the firm has some operational flexibility, then the above ODE must be replaced by a system of connected ODEs. For example, if production can be temporarily halted and restarted depending on whether output prices, x_t exceed production costs, say K, then we have two possible connected ODEs of the form

$$\frac{1}{2}\sigma^2 x^2 F_{xx}(x) + (r-q)F_x(x) - rF(x) + \pi_1(x) = 0 \text{ for } x \geq K$$
$$\frac{1}{2}\sigma^2 x^2 F_{xx}(x) + (r-q)F_x(x) - rF(x) + \pi_2(x) = 0 \text{ for } x \leq K.$$

Here $\pi_1(x)$ and $\pi_2(x)$ are the cash flows in the two different regions. The solutions in each region take the form in Eq. (4) with now typically a richer requirement of

[6]For further discussion on the Feyman-Kac theorem and the connection between stochastic calculus and partial differential equations, see Shreve (2004).

matching and smooth pasting conditions that not only occur at bankruptcy but also at the switching point, K.

3.1 Simple Examples

3.1.1 Valuation of a Contract That Receives \$1 Contingent on Hitting $x = U$

To illustrate, we consider a contract that pays \$1 when the state variable hits $x = U$ for the first time from below. Let $AD^u(x)$ represent its value for $x < U$. Since this claim satisfies Eq. (3) its value can be computed as an expectation as in Eq. (5) under the risk neutral measure, Q, as

$$AD^u(x) = E^Q\left[\int_0^\infty e^{-r\tau} g(\tau) d\tau\right],$$

where $\tau = inf\{t \geq 0 | x_t \geq U\}$, and $g(\tau)$ is its density.

The value of the claim can be more readily computed from Eq. (4). Since there are no intermediate cash flows $\pi(x) = 0$. Further since $\lim_{x \to 0} = 0$, $a_2 = 0$, so the form of the solution is $AD^u(x) = a_1 x^{\alpha_1}$. Now to determine a_1, consider the matching condition at $x = U$. We require

$$AD^u(U) = a_1 U^{\alpha_1} = 1,$$

from which $a_1 = U^{-\alpha_1}$ and $AD^u(x) = \left(\frac{x}{U}\right)^{\alpha_1}$. Using similar logic the value of a claim that pays \$1 when x hits D from above for the first time, $AD^d(x)$, say is given by

$$AD^d(x) = \left(\frac{x}{D}\right)^{\alpha_2}.$$

This two claims, which we refer to as Arrow-Debreu claims are quite important and we will use them to price risky debt which pays out in full unless a default barrier is hit.

3.1.2 Valuation of a Simple Investment Option

As a second simple example, assume x is the net revenue stream from a project. The present value of the revenue stream under the risk neutral measure is PV where

$$PV = \int_0^\infty e^{-rt} E_0[x_t] dt = \int_0^\infty e^{-rt} x_0 e^{(r-q)t} dt = \frac{x}{q}.$$

The NPV of the project if initiated today is $\frac{x_0}{q} - I$ where I is the investment cost. So $NPV > 0$ if and only if $x > Iq$. Let $x^{(NPV)} = qI$ above which the NPV is positive.

Now assume we have the option to invest. Until the option is exercised there is no cash flow. Let U be the endogenously determined point where the option is exercised. The form of the value of the claim from Eq. (4) is $C(x) = a_1 x^{\alpha_1}$. The matching and smooth pasting conditions that determine a_1 and U are

$$a_1 U^{\alpha_1} = \frac{U}{q} - I$$
$$\alpha_1 a_1 U^{\alpha_1 - 1} = \frac{1}{q},$$

from which $a_1 = (\frac{U}{q} - I)U^{-\alpha_1}$ and $U = \frac{\alpha_1 q I}{\alpha_1 - 1}$.

The value of the investment option, $C(x)$, can be rewritten as

$$C(x) = (\frac{U}{q} - I)\left(\frac{x}{U}\right)^{\alpha_1}. \tag{7}$$

Note that since $\frac{U}{q}$ is just the present value of the revenue stream at U, the above expression is just the net payout $(\frac{U}{q} - I)$ at U multiplied by the Arrow Debreu price, $AD^u(x) = (\frac{x}{U})^{\alpha_1}$. Clearly, with $\alpha_1 > 1$, $U > x^{(NPV)}$. This example, highlights the notion that just because a project has a positive NPV, it does not mean it should start right away. There is value in waiting. Indeed, it is easily shown that as uncertainty, i.e. σ, increases, the value of α_1 is affected, and that leads to a higher value of U, a higher value in waiting.

4 Valuation of an Investment Option

In the above problem we did not consider the impact of any market frictions such as taxes and bankruptcy costs. As a result, the nature of financing the investment of I was irrelevant. We now consider the impact of decisions when such frictions exist. We also assumed that the earning stream from the investment followed a Geometric Wiener process. So, there was no possibility of shortfalls. In this section, we assume that production costs are fixed at rate K per unit of capacity per year, and that sales follow a Geometric Wiener Process. As a result the earning before interest and taxes per unit of capacity at date t is $x_t - K$ which now can be negative. A firm with the rigid technology has to bear any operational losses, whereas a firm with flexible technology can avoid these losses by temporarily shutting down operations. In this section we compare the investment timing and financing of these two different technologies.

The objective of the firm is to maximize shareholder value. We start by assuming the firm has no assets in place and only has this growth option. Later we extend our analysis to the case where a firm has existing assets in place that are funded with a mix of debt and equity. We assume the debt takes the form of a perpetuity paying a coupon of c. The coupon payments are tax deductible with the tax rate being τ. Let $\gamma = 1 - \tau$. The optimal capital structure at the investment date consists of a mixture of equity and debt financing. Let $E_2(x)$ and $B_2(x)$ represent the value of the equity and debt after the investment point. The subscript 2 reflects the fact that these values are after the investment has taken place and r indicates this is a firm using rigid technology. Let $E_1(x)$ be the value of the growth option prior to the investment date. This value also reflects the full value of the firm, since prior to investment there is no debt outstanding.

After investment, and outside of bankruptcy the firm earns the effective tax shields of τc. Unfortunately, if prices fall below operating costs, operating losses are incurred. These losses, together with obligatory coupon payments have to be made by the existing shareholders. If these losses are significant, then rather than make them the equityholders will declare bankruptcy. Let P_2 be the level of prices below which the firm defaults. Upon default, the firm is liquidated and the bondholders receive residual values according to their priority. Bankruptcy costs include the loss of tax shields, and the loss of a fraction, b of the assets in place. The value of the levered firm after investment is $V^L(x) = E_2(x) + B_2(x)$. This value would exceed the value of the assets of the firm because of the tax shield. However higher funding with debt, increases the likelihood of incurring bankruptcy costs. The idea then is to choose the equity-debt mix in such a way as to optimally trade off these two forces.

As before the commodity or output price x per unit, evolves in accordance with a geometric Brownian motion under a risk-neutral measure, $\mathbb{Q}$ given in Eq. (6). Notice that at the investment point, $X = u$, the firm has no debt and so chooses the coupon so as to maximize the firm's value. The cash raised from the debt offering is then passed onto the equityholders as a special dividend and the capital structure of the firm now consists of debt and equity claims. The option to invest could be exercised immediately, but, as we have seen, the optimal policy could involve waiting until the commodity price rises to some critical level U.

4.1 Valuation of a Growth Option Without Operational Flexibility

Assume the growth option has been exercised and the coupon is c. The equity of the levered firm after the growth option has been exercised has the form in Eq. (4) with

$$E_2(x) = -\left(\tfrac{mK}{r} + \tfrac{c}{r}\right)\gamma + \tfrac{m\gamma}{q}x + A_2x^{\alpha_2} \text{ for } x \geq P_2. \tag{8}$$

Using the boundary conditions at the default point, P_2 we have, $E_2(P_2) = 0$ or:

$$-\left(\tfrac{mK}{r} + \tfrac{c}{r}\right)\gamma + \tfrac{m\gamma}{q}P_2 + A_2P_2^{\alpha_2} = 0,$$

from which

$$A_2 = \left(\frac{mK}{r} + \frac{c}{r} - \frac{mP_2}{q}\right)\gamma P_2^{-\alpha_2}.$$

To obtain the default point, P_2, we incorporate the smooth pasting condition at P_2. Differentiating Eq. (8) at P_2 leads to

$$\frac{m\gamma}{q} + \alpha_2 A_2 P_2^{\alpha_2 - 1} = 0.$$

Substituting for A_2 and simplifying yields:

$$P_2 = \frac{(mK+c)q\alpha_2}{mr(\alpha_2-1)}. \tag{9}$$

The value of an unlevered firm of size m can be obtained from the above equity equation by putting $c = 0$. This leads to:

$$V^U(x) = \left(\frac{m}{q}x - \frac{mK}{r}\right)\gamma + A^* x^{\alpha_2} \text{ for } x \geq P^*, \tag{10}$$

where A^* and P^* are just A_2 and P_2 with $c = 0$.

Let $B_2(x)$ be the values of the debt after investment has taken place. The value can readily be obtained given the Arrow-Debreu price of a claim that pays \$1 if P_2 is hit from above for the first time. Let $AD^d(x)$ be the price of such a claim. We have:

$$B_2(x) = \frac{c}{r} - \frac{c}{r}AD^d(x) + (1-b)V^U(P_2)AD^d(x) \tag{11}$$

The first term represents the price of a riskless perpetuity. The second term represents the price of a perpetuity that stops paying when x hits P_2 from above. So collectively, the first two terms represent the present value of a coupon stream that is extinguished when x hits D_2 for the first time. If there was no salvage value (i.e. if $b = 1$), this would be the value of the risky debt. But upon hitting P_2 the bondholders get some fraction, $(1-b)$, of the value of the assets of the unlevered firm. The third term represents the present value of this claim.

Substituting $AD^d(x) = (\frac{x}{P_2})^{\alpha_2}$ and for P_2 as well as the expression for $V^U(P_2)$ into the bond pricing equation, and simplifying eventually leads to

$$B_2(x) = \frac{c}{r}\left[1 - \left(\frac{x}{P_2}\right)^{\alpha_2}\right] + (1-b)V^U(P_2)\left(\frac{x}{P_2}\right)^{\alpha_2}. \tag{12}$$

The value of the levered firm $V^L(x)$ after investment is just the sum of the equity and debt values. This can be expressed as

$$V^L(x) = \frac{\tau c}{r} + \frac{-mK\gamma}{r} + \frac{m\gamma}{q}x + \eta x^{\alpha_2}, \tag{13}$$

where $\eta = ((1-b)V^U(P_2) - \frac{c}{r})P_2^{-\alpha_2} + A_2$.

We now consider valuation before the investment is made. We need to determine the investment time, U, and the optimal coupon c. Since decisions are determined by the equityholders we first investigate the equity value before investment, $E_1(x)$. The form of the value function for equity before the growth option is exercised is given by Eq. (4) with:

$$E_1(x) = M_1 x^{\alpha_1} \text{ for } x \leq U. \tag{14}$$

The unknown variables U and M_1 are fully determined by the following matching and smooth pasting conditions at U:

$$\begin{aligned} M_1 U^{\alpha_1} &= \tfrac{\tau c}{r} - \tfrac{mK\gamma}{r} + \tfrac{m\gamma}{q} U + \eta U^{\alpha_2} - I \\ \alpha_1 M_1 U^{\alpha_1} &= \tfrac{m\gamma}{q} U + \alpha_2 \eta U^{\alpha_2}. \end{aligned}$$

Substituting the first equation into the left hand side of the second equation and rearranging leads to identification of U as the solution to

$$\alpha_1(\tfrac{\tau c}{r} - \tfrac{mK\gamma}{r} - I) + \tfrac{m\gamma}{q}(\alpha_1 - 1)U + (\alpha_1 - \alpha_2)\eta U^{\alpha_2} = 0.$$

Given U, M_1 is

$$M_1 = \left(\tfrac{\tau c}{r} - \tfrac{mK\gamma}{r} + \tfrac{m\gamma}{q} U + \eta U^{\alpha_2} - I\right) U^{-\alpha_1}.$$

The optimal U and $E_1(x)$ values will depend on the coupon c. The optimal coupon is chosen so as to maximize the firm value at U. So optimal c and optimal U needs to be solved simultaneously. For general $K > 0$, numerical methods are required to solve for U.[7]

To illustrate the results, consider the valuation of the growth option for the case where the parameters of the risk neutralized commodity price are $r = 0.06$, $q = 0.05$, $\sigma = 0.25$, with $x = 20$. Assume the cost of production is $K = 20$ and $m = 1$. The investment cost $I = 100$, and the tax rates and bankruptcy costs are $\tau = 0.15$ and $b = 0.50$.

If we restrict attention to all equity financing, then the value of the growth option is 96.65 and the optimal point at which to invest would be when the commodity price reaches $U = 47.24$. In this case, the value of the all equity firm right after investment would be 541.37. In contrast, if debt is allowed and is taken optimally, then the investment point *decreases* to 46.67, the value of the growth option *expands*

[7]Notice that if $K = 0$, it can be shown that the optimal coupon is linear in U. This observation simplifies the expression for U and indeed the optimal U to be made explicit. However, with $K > 0$ numerical methods are required to solve for U. For the case where $K = 0$, if we further restrict financing so that no debt can be used, then the value of the growth option, $E_1(x)$, reduces to our earlier equation, (7).

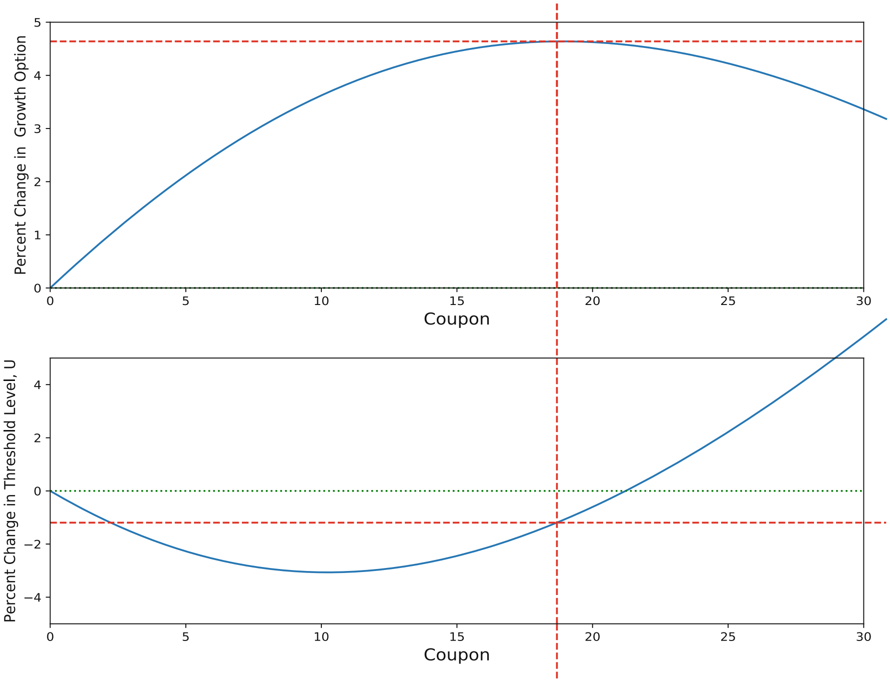

Fig. 1 Impact of optimal financing decisions compared to all equity financing. The top (bottom) graph shows how the value of the growth option (investment timing decision) is affected by increasing the use of debt financing from its base case of $c = 0$. The results are reported in percentage terms relative to the values for an all-equity financed firm

to 101.14 and the optimal capital structure after investment consists of debt worth 220.19, equity worth 331.93 and an increased firm value of 552.11. The growth option with optimal financing enhances the value over pure equity financing by 4.64%.

The top panel of Fig. 1 shows the percentage improvement in the value of the growth option as the coupon increases. The bottom panel shows how the threshold value changes (relative to the all equity threshold) as the coupon increases. Notice that when low coupons are used the threshold decreases from all equity financing. This property is referred to as the over-investment problem. Specifically, the use of a low amount of debt results in earlier investment than if all funding is provided by equity. This property results from the fact that, with debt, equityholders have less skin in the game and therefore are incentivized to commit to a project earlier.

When debt becomes excessive, however, then default risk becomes dominant and the rational response is to extend the waiting to a higher threshold before investing. The figure shows that under-investment occurs if the coupons in excess of 20 are used. Notice that at the optimal coupon of $c^* = 18.68$, we still have over investment, with investment occurring earlier than with all equity financing.

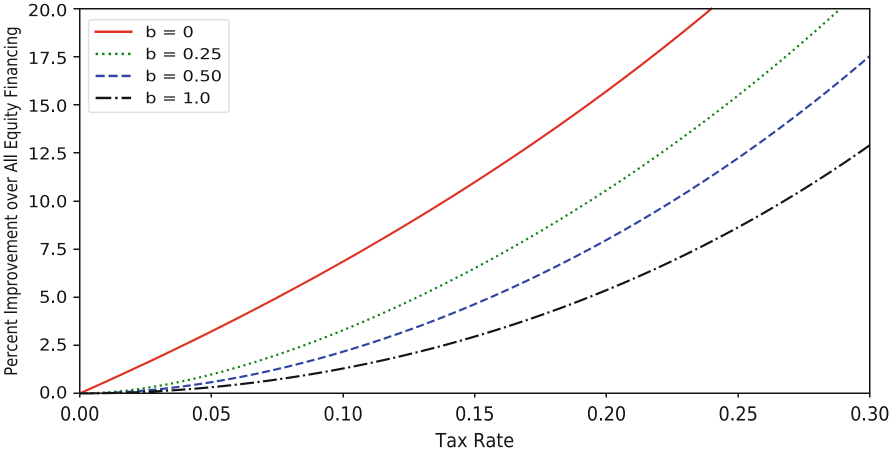

Fig. 2 Impact of frictions on optimal financing decisions compared to all equity financing. The graph shows how the value of the growth option is affected by increasing taxes for different bankruptcy costs relative to the all equity financing case for a firm with the same tax rate and bankruptcy cost. As taxes increase the advantage of debt financing over equity financing increases regardless of the bankruptcy cost. The standard case parameters are used

Financing affects operating decisions because of the frictions introduced by taxes and deadweight bankruptcy costs. Figure 2 shows how financing affects the value of the growth option as tax rates increase from 0 to 30% for different levels of deadweight bankruptcy costs. When b and τ are zero, i.e. when there are no frictions, financing decisions create no value. As tax rates increase, the benefit from debt due to the tax shield increases and this is reflected in a greater percentage improvement. As bankruptcy costs increase, the optimal coupon decreases as the consequences of bankruptcy increases. This decreases the value of the tax shield reducing the percentage improvement.

4.2 Valuation of a Growth Option with Operational Flexibility

So far we have assumed the technology provided by the investment had no operational flexibility. In other words, operational managers, representing the interests of shareholders had no ability to shut production down temporarily when prices fell below operational costs. As technology improvements have occurred, in many industries the ability to open and shut down has been reduced. In this section we consider the case when managers have a technology that they can use to adaptively respond to market conditions. Specifically, we assume the technology allows for costlessly shutting down and reopening. The idea is to investigate how flexibility in the underlying technology affects the timing and financing of investments. This system is studied extensively in Ritchken and Wu (2021), hence we will only briefly sketch the main procedure of solving this system.

We proceed is a somewhat similar way to our previous discussion on a rigid technology. We first examine the value of an unlevered firm, then establish the value of the equity and debt after investment and finally we get the value of the growth option, prior to investment. To the extent that we need to differentiate the value of claims for the flexible case with our earlier rigid case, we will add superscripts of F, for flexible, to distinguish from the rigid case we discussed before.

As for the rigid system, Let $E_2(x)$ be the value of the equity after the expansion, and let $B_2(x)$ be the corresponding value of the debt with coupon c. The instantaneous payoff for equity holder is $(\max(x_t - K, 0)m - c)\gamma$ till bankruptcy and $h = m\gamma$ if $x \geq K$; and $CF_0 = 0$ and $h = 0$ if $x < K$. Specifically, with $V(x) = E_2(x)$, we have:

$$\begin{array}{ll} \frac{1}{2}\sigma^2 x^2 V^F_{xx}(x) + (r-q)xV^F_x(x) - rV^F(x) + m(x-K-c)\gamma & = 0 \text{ for } x \geq K \\ \frac{1}{2}\sigma^2 x^2 V^F_{xx}(x) + (r-q)xV^F_x(x) - rV^F(x) - mc & = 0 \text{ for } P^F_2 < x < K, \end{array} \tag{15}$$

with boundary condition $V^F = 0$ and smooth pasting condition $V^F_x = 0$ holding at the default point, $x = P^F_2$. The general solution takes the form

$$E_2(x) = \begin{cases} -\frac{mK\gamma}{r} - \frac{c\gamma}{r} + \frac{m\gamma}{q}x + A_{12}x^{\alpha_2} & \text{for } x \geq K \\ -\frac{c\gamma}{r} + A_{21}x^{\alpha_1} + A_{22}x^{\alpha_2} & \text{for } P_2 \leq x < K \\ 0 & \text{for } x < P_2. \end{cases} \tag{16}$$

P_2 is the bankruptcy threshold chosen to optimize the value of E_2. Similar to the calculation for the unlevered firm case, the four unknowns (A_{12}, A_{21}, A_{22}, and P_2) in the system above can be solved by the four matching and smooth pasting conditions at P_2 and K. The matching condition at P_2 requires that the equity value is zero at P_2, and the matching condition at K requires the two equity function values be equal at K, so that there is indifference between operating and being idle.

In setting up the matching and smooth pasting conditions, we have assumed that the coupon is sufficiently small such that default occurs while the system is idle. For this to happen $P_2 < K$. Since P_2 increases in coupon used, there exist $\bar{c}$, such that if the coupon exceeds $\bar{c}$, then default will occur when the firm is operating. In this case, the solution for equity takes on the form:

$$E_2(x) = \begin{cases} -(c + mK)\frac{\gamma}{r} + \frac{m\gamma}{q}x + L_{22}x^{\alpha_2} & \text{for } x > P_2 \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Once the default barrier is established, the price of the bond immediately follows using Eq. (12), where $V^U(P_2)$ for the value of all equity firm, can be obtained simply by setting $c = 0$ in $E_2(x)$ in Eq. (17). The value of the levered firm is then the sum of the value of the equity and total debt. If the coupons are chosen optimally, then this value should exceed the value of the unlevered firm.

We establish the coupon at state variable U that maximizes the value of equity. In this case, since there is no existing debt in place, the equityholders choose the debt that maximizes the value of the levered firm. Then given a fixed U, $c^*(U) = \arg\max_c V_1^L(U; c)$ with $V_1^L(U; c) = E_2(U; c) + B_2(U; c)$. Interestingly, although the system of ODEs for the flexible case is more complicated than the rigid case, the optimal coupon c and the default barrier P_2 takes simpler analytical form. c is a power function in the investment point U and P_2 is a linear function in U. Since the coupon c is a simple function of U, the bond and equity values at investment can be written as solely a function of U. For the rigid case this was not possible.

To solve for the optimal investment threshold U, we need to consider the firm's value before and after the investment. Just before the investment, the firm is an all-equity firm's instantaneous cash flow is 0, and the equity value E_1 satisfies

$$\tfrac{1}{2}\sigma^2 x^2 V_{xx}^F(x) + (r-q)xV_x^F(x) - rV^F(x) = 0 \text{ for } x \geq K. \tag{18}$$

The general solution of the above ODE is θS^{α_1}, where θ and U are constants that need to be determined. Once U is identified the optimal coupon is also known. The values for U and θ are identified by the matching and smooth pasting conditions at $x = U$. The value matching condition requires the equity value before investment to equal the equity value after investment less the contributions from the equity holders, $I - B_2(U)$.

Using the same parameters as for the inflexible technology, and assuming all equity financing of the flexible (rigid) technology, the value of the growth option is 99.42 (96.65) and the threshold at which investment takes place is $U = 44.18$ (47.24). At the investment point, the all equity value of the firm is 503.26 (541.37). If optimal financing using debt and equity is allowed then the value of the growth option expands to 102.88 (101.14) with the threshold investment level being 44.18, (46.67) which is the same (different) as the all equity financing case. At the investment point the optimal coupon is $c = 14.77$ (18.69) and the capital structure at the investment point consists of equity worth 332.62 (331.93) and debt worth 184.71 (220.19) for a levered firm value of 517.33 (552.11).

The top panel in Fig. 3 shows the percentage improvement in the value of the growth option as the coupon increases, while the bottom panel shows the percentage change in threshold level as more debt is used. Notice that as the coupon increases from zero, optimal timing takes place at lower threshold levels then if the investment were fully financed with equity. As for the rigid case, this is consistent with the over-investment problem that debt induces. However, beyond some coupon the threat of bankruptcy becomes dominant and investment is delayed. In this example, we see that at the optimal coupon, investment takes place at exactly the same time as if the investment were fully funded by equity! This result can be shown analytically in the following proposition.

Proposition 1 *Consider a firm with a growth option to buy a flexible technology. The optimal time to exercise this option and finance it with optimal amounts of debt and equity occurs at the same time for a firm that elected to fund the investment*

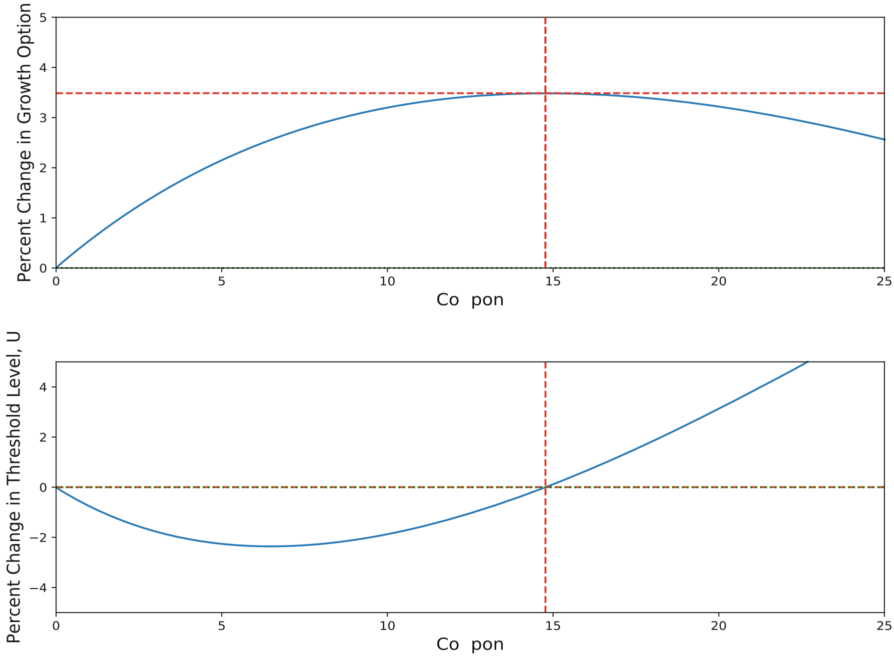

Fig. 3 Impact of optimal financing decisions compared to all equity financing. The top (bottom) graph shows how the value of the growth option (investment timing decision) is affected by increasing the use of debt financing from its base case of $c = 0$. The results are reported, in percentage terms, relative to the values for an all-equity financed firm with the same tax rate and bankruptcy cost. Here the technology is flexible while in Fig. 1 the technology is rigid

only with equity. This statement does not hold if the underlying technology to be purchased had no flexibility.

Proof (See Ritchken and Wu (2021)) In other words, with a flexible technology, using optimal financing there is no longer an over- or an under-investment issue. Timing decisions are exactly unaltered by financing frictions such as taxes and bankruptcy deadweight costs.

4.3 Comparison of the Investment and Financing Decision for Flexible and Rigid Technology

The examples of a rigid and flexible technology illustrate another point. In particular, not surprisingly, all things being equal, the flexible technology leads to a greater value. But more interesting is that with optimal financing the adoption of the flexible technology takes place earlier than the rigid technology ($U^F = 44.18$ versus

$U^R = 46.68$) and the optimal coupons used for financing are lower ($c^F = 14.77$ versus $c^R = 18.69$). At first glance this seems surprising. One could imagine that adoption of the flexible technology allows management to avoid operational losses by temporarily shutting down. And because of this, it allows the firm to take on more debt. However, such an argument is incomplete. Indeed, because the firm with the rigid technology cannot avoid operational losses, they may find it more efficient to push up the default barrier, P_2, so as to decrease the region where losses are experienced. Of course, the lenders of the firm with the rigid technology bear more downside risk and hence they fairly pass this cost onto the firm, in the form of higher credit spreads. While the above example suggests that the rigid firm may take on more debt, it turns out that this is generally true. In particular, Ritchken and Wu (2021) compare two firms with identical characteristics, except along the dimension of technology. The following theorem answers these questions.

Theorem 1 *Assume there are two firms with identical characteristics, except one firm has a growth option on a flexible factory, while the other has a growth option on a rigid factory. Both factories have no recovery value from bankruptcy. Let U^F and U^R be the levels of the commodity price at which the firms exercise their growth options and c^F and c^R be the optimal coupons used to finance their investments. Then:*

(i) Given $c^F = c^R = c_2$, the optimal exercise time of the growth option satisfies $U^{F,*}(c_2) < U^{R,*}(c_2)$.
(ii) Given $U^F = U^R = U$, the optimal coupon payments of the two firms satisfy $c^{F,*}(U) < c^{R,*}(U)$.
(iii) The optimal investment threshold U^* and the coupon payment c^* satisfy $U^{F,*} < U^{R,*}$ and $c^{F,*} < c^{R,*}$.

Proof (See Ritchken and Wu (2021)) The first part of Theorem 1 states that if the coupons are the same, then the firm with the option on the flexible technology would exercise earlier than its rigid counterpart. This seems rather intuitive: The flexible technology is more profitable, so the marginal cost of deferring exercise is larger. Therefore, if the coupons are constrained to be equal, it seems reasonable that the firm holding the option on the more profitable flexible factory would be exercised earlier.

The second part of Theorem 1 states that if the investment dates were constrained to be equal, then the optimal coupon for the firm with the inflexible technology would be *higher* than the optimal coupon used to finance the purchase of the flexible technology. As discussed earlier this is less intuitive since at first glance, it seems plausible that since the firm with the more flexible technology can better manage risk by temporarily shutting down in bad states of nature, it could avoid operational losses, and as a result has the potential to take on more debt by having a higher coupon. However, this line of reasoning is incomplete. It ignores the strategic purpose of using debt to reduce operational losses when flexibility is absent.

In particular, as the revenue falls below the costs, the net cash flow from operations for the rigid firm is negative. As the coupon increases, the bankruptcy threshold rises. For the firm adopting the rigid technology, this increase shrinks the region where operating losses are realized. Thus, for the equityholders of rigid firms, increasing coupons not only provides tax benefits in good states but also reduces the loss region in bad states. These two features have to be traded off against increasing the present value of the bankruptcy costs associated with the rising threshold. In contrast, for the firm adopting the flexible technology, when prices fall below production costs, production is temporarily stopped, so operating losses are not incurred. Hence, for firms with flexible technologies, while increasing debt provides tax benefits, that have to be traded off against bankruptcy risks, it does not reduce operating losses. Compared to a firm with rigid technology, these relatively diminished benefits have to be traded off against the increased cost of default from raising the bankruptcy threshold. This tension results in the firm with an inflexible technology choosing a *higher* coupon than the firm with a flexible technology.

The third part of Theorem 1 builds on the first two parts. The flexible firm has a profit process that stochastically dominates that of the rigid firm, i.e., is more profitable. Theorem 1 states that this more profitable flexible firm will invest earlier and will use less debt than the rigid firm. Further, based on Proposition 1, if the investment is financed optimally, the timing of the investment is coincides with the investment time if the firm were financed fully with equity. That is, there are no under- or over-investment agency costs.

Although Theorem 1 is derived based on $b = 1$, i.e., zero recovery rates, the result can be extended to the general case where $b < 1$. In Fig. 4, we compare the optimal coupon c, and the optimal investment timing U for different values of b for the flexible and rigid investments. We find that the relationship in Theorem 1 continues to hold for general b. Flexibility always leads to a lower coupon issued, and an earlier investment time, compared to the rigid technology. The bottom graphs shows with decreasing coupon usage as default costs increase, the default barrier P_2 deceases as does the value of the levered firm.

5 Extensions

An important take away from our analysis so far is that operational decisions, such as when to purchase a technology, depends very much on the nature of the technology and in particular, the embedded flexibility associated with the technology. The results hint at the fact that the greater the flexibility, all things being equal the earlier the investment will be made, the less debt that will be used, and the lower the issues will be with under or over investment. So far, our results is based the simplest setting of such models, and quite stylized in several dimensions. For example, we have assumed that prior to investment the firm had no existing assets in place and further that the firm had no existing debt in place. In his section, we shall explore how the existence of assets and debt affect the timing and financing of both rigid and flexible technologies.

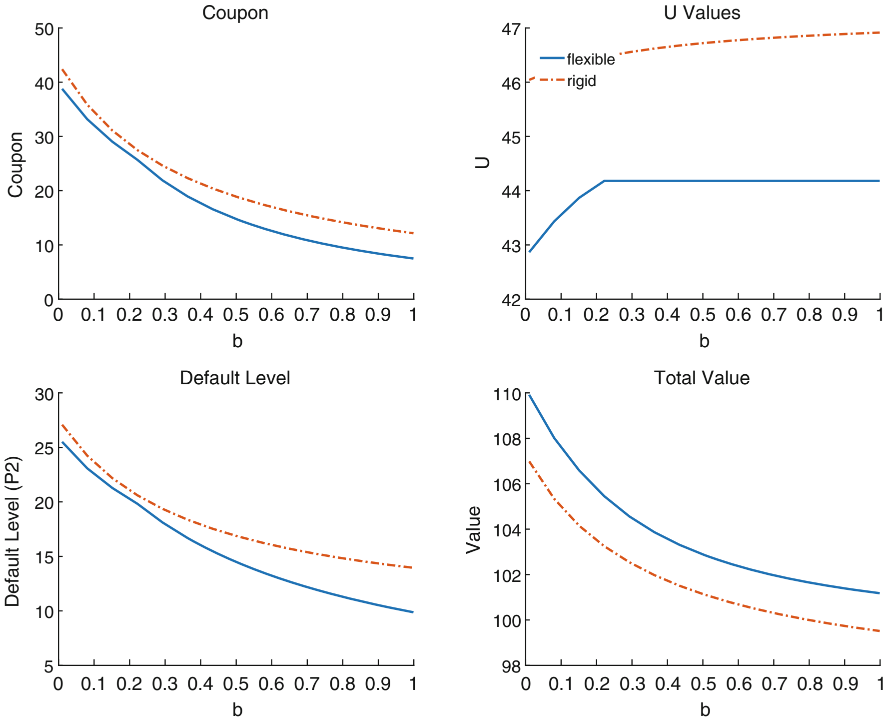

Fig. 4 Comparison of firms with flexible vs rigid technology for general b. The top two graphs show the optimal coupons, c_2, and the optimal exercise points, U, are compared for flexible and rigid firms for the increasing bankruptcy costs, b. The lower two graphs show the corresponding default barrier P_2 and the total value of the firm at time 0

The time line of events we consider here is as follows. At time 0, the firm decides its initial capital structure consisting of equity and debt, with the debt taking the form of a perpetuity. Therefore, determining the initial capital structure is equivalent to determining the coupon c_1. Since, at this point in time, the firm does not have any debt, maximizing the equity value is equivalent to maximizing the firm's value. Since then, the capital structure consists of equity priced at $E_1(x)$ and debt priced at $B_1(x)$. At U, the investment, I is funded with a mix of equity and debt paying c_2. The new default barrier after investment is now P_2. The capital structure after investment consists of equity priced at $E_2(x)$ and two tranches of debt with the original debt now priced at $B_{21}(x, c_1)$ and the new debt priced at $B_{22}(x, c_2)$. However, at U, when the firm/manager now acts in the benefit of shareholders (equityholder), the initial bond holder's benefit is not taken into consideration. And this leads to agency cost in this setting with multiple trench of debt. The objective of this firm is always to maximize shareholder value, i.e., to maximize the firm's equity value. Such decisions are referred to as second best solutions; while decisions made to maximize the value of the firm are referred to as first best solutions. The difference between the first best and second best is what we refer to as the agency cost.

In the rest of the section, we lay out the details for the rigid technology. Discussion of results for the flexible technology is omitted since the solving procedure is similar, and details are provided in Ritchken and Wu (2021).

5.1 Investing in a Rigid Technology with Asset in Place

We now consider the case of a firm with assets of size m_1 in place and an existing capital structure consisting of equity and debt. The debt consists of a perpetuity paying a coupon of size c_1 per year. The firm also has an investment option that allows them to invest I at any future point in time and expand capacity from m_1 to $m = m_1 + m_2$. Let U be the price of the commodity at which point expansion will take place. At that point, the firm can finance with a portfolio of debt and equity. The debt takes the form of a perpetuity paying coupons at rate c_2. So the total coupons being paid after expansion is $c = c_1 + c_2$. As before, we assume the tax rate is τ, and the deadweight bankruptcy proportion is b.

Using this setting, the equity value has the same form as in Eq. (8),with the only distinction that c and m being $c_1 + c_2$ and $m_1 + m_2$. For the bond value, we still assume that an absolute priority rule is in place so that when bankruptcy takes place, equityholders receive nothing. Upon bankruptcy, the firm is liquidated, and the bondholders receive residual values according to their priority. When the bankruptcy occurs after the investment, there are two trenches of debt (the original tranche issued at 0 and the second, issued at U). When bankruptcy occurs, the residual value is distributed according to the contractually specified priority rule. We assume, as our benchmark case, equal priority in bankruptcy between the two rounds of debt. This assumption is consistent with the fact that most firms issue debt sequentially at one seniority level only (Bris et al., 2009; Billett et al., 2007). Under equal priority, the liquidation proceeds of the firm are shared according to their contributions, with the original (first-round) bondholders receiving the fraction c_1/c, and the second-round bondholders receiving the fraction c_2/c. Therefore the bond value is

$$B_{2j}(S) = \frac{c_j}{r}\left(1 - \left(\frac{S}{P_1}\right)^{\alpha_2}\right) + F_j(SV(P_2))\left(\frac{S}{P_2}\right)^{\alpha_2} \text{ for } j = 1, 2. \tag{19}$$

where $F_j(SV(x))$ for $j = 1, 2$ is defined according to the absolute priority rule. Specifically, under our benchmark assumption of equal priority, we have $F_j(SV(x)) = (1 - b)SV(x)\frac{c_j}{c_1+c_2}$. In contrast, if the first bond issue is senior, we have $F_1(SV(x)) = \min[\frac{c_1}{r}, (1 - b)SV(x)]$ and $F_2(SV(x)) = (1 - b)SV(x) - F_1(SV(x))$.

The optimal coupon value, c_2, depends on the level of the state variable, x, when the growth option is exercised, and it is determined by the equityholders. Assume the growth option is exercised when $x = U$. The optimal coupon, also will depend on c_1. Therefore, c_2 is dependent on both U and c_1 is chosen so as to maximize the equity value, of $E_1(U; c_1, c_2)$, which equals to $E_2(U; c_1, c_2)$ plus the second tranche of debt $B_{22}(U; c_1, c_2)$ minus the investment cost, i.e.,

$$c_2^* = \max_{c_2} \left(E_2(U; c_1, c_2) + B_{22}(U; c_1, c_2) - I \right). \tag{20}$$

At the point U, the shareholders make decisions in their best interest. The new bondholders will always demand fair value. But the original bondholders may get a raw deal. Of course, they recognize that this could happen at the expansion date, and so ex ante, when pricing the bond at date 0, they take this into account by requiring an appropriate fair credit spread. So equity holders bear this agency cost. This solution is called a second best solution, since it does not maximize the value of the firm. If all decisions were based on firm maximization, then we would have a first best solution. In the earlier section, we actually did not have this problem since with only one round of financing the optimal strategy for equityholders is to maximize firm value. However, with these two rounds of debt issuance, this agency problem does arise.

Before the investment, the existence of debt implies that there is a threshold, P_1, such that the equityholders will rationally declare bankruptcy if the price falls below the threshold. If coupon c_1 is sufficiently small such that the default barrier is below K, i.e., $0 < P_1 \leq K$, then the general form of the equity given in Eq. (4) is:

$$E_1(x) = \begin{cases} -\frac{(m_1 K + c_1)\gamma}{r} + \frac{m_1 \gamma}{q} x + M_1 x^{\alpha_1} + M_2 x^{\alpha_2} & \text{for } P_1 \leq x < U \\ 0 & \text{for } x < P_1. \end{cases}$$

To identify the equity value, we first need to compute the two coefficients M_i with $i = 1, 2$. These values are established by the two value matching conditions at P_1 and U. together with the two smooth pasting conditions:

$$\frac{dE_1(x)}{dx}\Big|_{x=P_1} = 0 \tag{21}$$

$$\frac{dE_1(x)}{dx}\Big|_{x=U} = \frac{dE_2(x)+B_{22}(x}{dx}\Big|_{x=U} \tag{22}$$

Although the analysis is more complex the overarching principles remain similar to what we covered when there was only one round of financing.

Now, the optimal capital structure at date 0, and the optimal investment time, U, will obviously depend on the magnitude of m_1 and m_2. Moreover, how c_2 is determined will have an impact on the choice of c_1. For example, firms with large growth options may reserve debt capacity for the later expansion. The choice of capital structure and the timing of expansion is also very dependent on the nature of the technology.

$$
\begin{aligned}
&Max_{c_1}\left(E_1(x_0;c_1,c_2)+B_1(x_0;c_1,c_2)\right)\\
\text{where}\quad & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(P1)}\\
&c_2 = argmax_c\left(E_2(U;c_1,c)-(I-B_{22}(U;c_1,c)\right),
\end{aligned}
$$

where the value of the coupon bond, $B_1(S)$ is given by

$$B_1(S) = \frac{c_1}{r}(1-p_u(S)-p_d(S)) + p_d(S)(1-b)V^U(P_1;m_1) + p_u(S)B_{22}(U),$$

with $p_u(S) = -\frac{P_1^{\alpha_2}}{\Sigma}S^{\alpha_1} + \frac{P_1^{\alpha_1}}{\Sigma}S^{\alpha_2}$ and $p_d(S) = -\frac{U^{\alpha_2}}{\Sigma}S^{\alpha_1} - \frac{U^{\alpha_1}}{\Sigma}S^{\alpha_2}$ where $\Sigma = P_1^{\alpha_1}U^{\alpha_2} - P_1^{\alpha_2}U^{\alpha_1}$. For derivation of this expression, see Goldstein et al. (2001).

In summary, the owners of the firm choose c_1 so that the value of the levered firm is maximized, taking into account that at some future date, established by the endogenously determined threshold U, the future optimal financing packages, i.e., equity and debt via c_2, will be used to fund the future investment. Unfortunately, due to the complexity of this system, there is no clean analytical solutions. All the decisions, c_1, c_2, P_1, P_2 and U need to be solved numerically.

5.2 Numerical Results

Table 1 reports the optimal policy variables for the second best strategies for the firm with flexible and rigid assets. The firm has assets in place with size, $m_1 = 1$ and a growth option of size $m_2 = 1$. The policy variables are the optimal coupons, c_1 and c_2, the investment threshold, U, and the default barriers before and after investment, P_1, and P_2.

When $K = 0$, negative earnings streams are not possible. In this case, there is no value in flexibility, and the results for the rigid and flexible firms are the same. These results are identical to Hackbarth and Mauer (2012). As K increases toward the current commodity price, negative earning states become more likely, and the impact of operational flexibility on the optimal decisions becomes more apparent. As operating costs increase, the flexible and rigid firms exercise their growth options later and use larger coupons (c_2) to finance this expansion. However, as K increases, the differences in timing U and debt issuance c_2 due to the nature of the assets, also increase. Consistent with our theory, the growth option is exercised earlier by the flexible firm, and the flexible firm uses less debt.

Table 1 shows that the initial optimal coupon, c_1, first increases and then decreases in the operating costs, K. To explain this, first note that as K increases from 0, the firm exercises its growth options later and uses larger coupons, c_2, to finance the investment. This has two competing effects. On the one hand, the longer the time to second-round financing, the greater the incentive to use more initial debt. On the other hand, the larger the coupon used at U, the lower the need to use debt at date 0. For both the rigid firm and the flexible firm, this trade-off works in the same overall direction. Very profitable firms will tend to use little initial debt, since the

Table 1 Comparison of leverage with increasing ongoing fixed costs. The table compares the policies of a flexible with a rigid firm when there are assets in place, $m_1 = 1$, and a growth option that could double the scale of the firm. The variables are the optimal coupons, c_1 and c_2, the investment threshold price, U, the default thresholds, P_1 and P_2, the total value of the levered firm, V_0

Second best policies and values								First best policies and values						
Flexible							Agency cost	Flexible						
K	C_1	C_2	U	P_1	P_2	V_0		K	C_1	C_2	U	P_1	P_2	V_0
0	0.00	28.17	20.0	0.00	6.11	616.7	0.00%	0	0.00	28.17	20.0	0.00	6.11	616.7
8	2.82	31.32	25.3	2.91	10.88	411.2	0.34%	8	6.79	26.09	27 1	4.79	10.61	412.6
16	3.96	43.38	37.1	4.53	17.22	293.3	0.22%	16	5.06	40.19	39.5	5.21	16.76	294.0
24	3.68	52.85	48.4	5.07	22.65	225.2	0.13%	24	4.11	49.47	508	5.41	21.98	225.5
32	3.29	59.23	59.0	5.32	27.17	185.5	0.08%	32	3.50	56.46	61.35	5.52	26.54	185.60
Rigid							Agency cost	Rigid						
K	C_1	C_2	U	P_1	P_2	V_0		K	C_1	C_2	U	P_1	P_2	V_0
0	0.00	28.17	20.0	0.00	6.11	616.7	0.00%	0	0.00	28.17	20.0	0.00	6.11	616.7
8	3.78	32.06	25.8	4.22	11.25	405.7	0.41%	8	8.16	25.28	27.5	5.57	10.73	407.3
16	4.76	44.71	38.2	7.31	17.68	269.6	0.19%	16	5.93	40.12	40.1	7.67	16.94	270.1
24	3.94	56.73	50.5	9.83	23.58	170.0	0.08%	24	4.17	53.86	52.23	9.90	23.01	170.2
32	3.17	68.59	62.7	12.37	29.46	95.9	0.10%	32	3.25	66.64	64.30	12.39	29.05	95.9

second debt issuance is close to the first issuance (in fact, when $K = 0, U = x_0$). As profitability decreases due to K increasing, U is pushed further away, and the firms respond by taking on more initial debt. However, beyond some operating cost, K, the advantage of using more initial debt falls below the marginal issuing debt later at U.

6 The Impact of Technology on Agency Costs

When a firm has assets in place and initial debt, as well as new investment opportunities financed with new debt, conflicts of interest arise between the equity holders and original debt holders. Regarding this, two natural questions arise. First, how do the priority rules between different issuances of debt affect the firm's capital structure? Second, what is the value loss due to the conflict of interest? In other words, what is the difference between the first best and second best solutions.

6.1 Impact of Different Absolute Priority Rules

So far we have assumed an absolute priority rule in which the two bond issues have equal priority (EQ) in default. In Fig. 6, for our three types of firms, we compare initial coupons, c_1, leverage, $L_m(S)$, credit spreads, cs_1, and the investment threshold, U, when the absolute priority rule is equal (EQ) to the case in which the priority rule requires the first bond issued to be senior (FS). For this comparison, we assume the firm has assets in place with $\pi_1 = 1$, and a growth option with $\pi_2 = 1$, with an expansion costing $I = \$100$. We make these comparisons for several different operating costs.

Figure 5 shows that, although the priority of the bond issue impacts the magnitude of the policy variables, the overall conclusions from the equal priority rule carry over when we compare the flexible firm to the rigid firm. Specifically, for highly profitable firms (i.e., firms with lower K), the initial coupon for the flexible firms is lower than for rigid firms and the time to exercise is earlier. As operating costs, K, increase, flexible firms may have higher coupons, but the investment timing remains earlier.

In comparing FS to EQ policies, we see that at low operating costs, higher coupons, c_1, are used under FS, with the difference decreasing as profitability decreases. Similarly, the difference in leverage increases with profitability. Highly profitable firms will take on significantly more initial debt if the debt is senior rather than equal. For example, for the flexible firm, when the operating cost is $K = 16$, the optimal coupon when senior (equal) debt is issued first is 4.43 (3.96). In addition, issuing senior debt first results in delaying investment with $U = \$40.84$, compared with $U = \$37.05$ under equal priority. In sum, while absolute priority rules affect

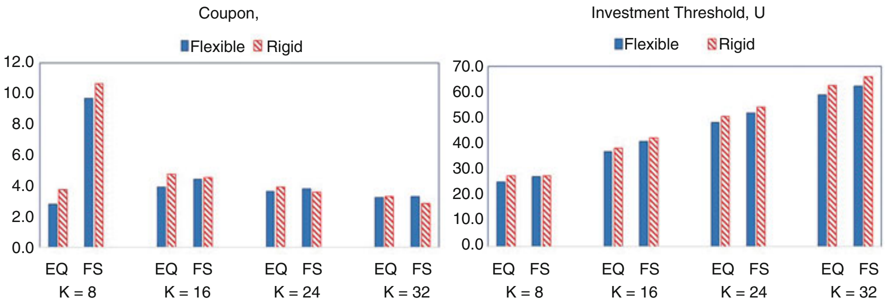

Fig. 5 Impact of different absolute priority rules. The graphs compares equal priority rules with first-senior priority rules. The left panel compares coupons and the right panel compares investment threshold levels

capital structure and investment timing, the direction of the impact of technological flexibility remains unchanged.

6.2 Technology Choices, First Best Solutions, and the Magnitude of Agency Costs

In our model, agency costs arise because after the initial bonds are issued, equityholders choose U and c_2 to maximize their own interests. While the new bondholders will require fair compensation, the original bondholders cannot influence these decisions. However, because they can anticipate the equityholders' actions, they certainly take this into consideration in determining the fair price of bonds when issued at date 0.

The right part of Table 1 shows the first best solutions, which maximize the firm's value. For both flexible and rigid firms, the initial coupons for first best policies are higher than their second best counterparts. For example, for the flexible (rigid) firm with $K = 16$, the first best coupon is 5.06 (5.93), while the second best coupon is lower at 3.96 (4.76). When profitability is greater (i.e., when K is smaller), differences between the first best and second best initial coupons are even larger.

The first best strategies also require delaying exercise until a higher state variable is realized, and they require using lower second-round coupons. For example, for the flexible firm, with $K = 16$, the first (second) best strategy has $U = 39.50$ (37.05) and a coupon of $c_2 = 40.19$ (43.38). Under second best policies, equityholders over-invest and exploit the original bondholders.

Since higher initial coupons are used under first best financing, the default levels, P_1, before investment are higher than for second best strategies. However, after financing of the growth option has occurred, the default barrier under first best policy, P_2, is lower than under second best financing. This is due to the fact that the second coupon, c_2, is smaller for the first best solution.

In Table 1, the agency cost of debt, expressed as a percentage, is given by the ratio of the value of the levered firm under the first best policy minus the value under the second best policy, relative to the value under the second best policy. When $K = 0$, the optimal policy is to exercise the growth option immediately. For this case, c_1 and c_2 are bundled together into one issue and there are no agency costs. With $K > 0$, agency costs increase as profitability increases. For our case parameters, the magnitude of the agency costs are quite small.

While the agency costs are quite small, the difference in optimal policies are quite distinct. For flexible (rigid) technologies, moving from second best to first best coupons in the first round would results in enormous increases of coupons. For example, if $K = 16$ the increase in coupons in the first round is 28% (25%). The increased use of debt in the first round also results in waiting longer before initiating the expansion (U increases by about 7%). Due to the existence of first round debt and in spite of the higher U values, the use of second round debt for first best policies is much lower, especially for rigid technologies. Moving from second best to first best policies when $K = 16$ results in 7.5% (10.3%) less debt.

6.3 *Financial Contracts to Achieve First Best Outcomes*

Is it possible to design financial contracts that induce the equityholders to adopt first best policies? For this to happen we need a financial mechanism whereby first best behavior naturally arises when second round debt is secured. One way to achieve this is for the first round debt to be set up so that ex ante they are fully protected when second round financing is needed. This might happen if the original bondholders can put their bonds back on the firm and receive appropriate fair compensation in such a way that the equityholders would adopt first best behavior. The following proposition shows that such a financial contract could in principle be designed to achieve this goal.

Proposition 2 (Putable Bonds and First Best Outcomes) *Consider an all-equity firm that issues a putable bond that allows the bondholders to return the bond only if new financing takes place. The buyback price at any refinancing point, U, for instance, $G(U)$, is chosen so that $G(U) = B_{21}^{(FB)}(U)$, which is the first-round debt value after investment under the first best policy. Then, equityholders will find it optimal to execute the first best policy.*

Proof The first best policy involves issuing a bond with coupon c_1^{FB}. The value of the bond at date 0 is $B_1(S)$. The default barrier is P_g^{FB} which is completely determined by the coupon c_1^{FB}. Now at price U^{FB}, refinancing occurs and a coupon c_2^{FB} is used, the bond prices are $B_{21}^{FB}(U^{FB}; c_1^{FB})$ and $B_{22}^{FB}(U^{FB}; c_2^{FB})$. The new default barrier is $P_2^{FB} = Q_2(c_1^{FB} + c_2^{FB})^{1\alpha}$.

We will show that there exists a putable bond that results in the shareholders being able to replicate the first best strategy, with any deviation from that strategy being suboptimal for the shareholders.

Clearly, we require $B_1(S) = B_1^{FB}(S)$ with $P_g = P_g^{FB}$. For the latter to be true we must have $c_1 = c_1^{FB}$. Now at time U^{FB} we require the payout of the putable bond, $G(U^{FB})$ to be B_{21}^{FB}. In other words at time U^{FB} the bondholders can redeem their bonds for $G(U^{FB}) = B_{21}^{FB}$. This being the case, then clearly at date 0 the price of the putable bond would be $B_1^{FB}(S)$.

Now the second best firm, after paying off the putable bond, would be an all equity firm. The first best firm after investment, will consist of two tranches of debt, with a default barrier P_2^{FB}. So if the all equity firm can create one bond with the same total coupon $(c_1^{FB} + c_2^{FB})$ that has the same default barrier, then the replication is complete. Clearly, $c_2 = c_1^{FB} + c_2^{FB}$ suffices. Note that the shareholders have no incentive to deviate from this policy, since any deviation, only hurts them. Specifically bond are always priced fairly, so the equityholders are forced into maximizing the value of the firm. Any destruction in firm value, only diminishes their take. ■

The policy is quite intuitive, in that equityholders now face exactly the same decision problem as in the first best. And, it is clear that the bondholders cannot be taken advantage of, since they can redeem at the investment point of the first best policy. At that time, the bondholders receive the same payout as the value of the original bond under the first best policy. Moreover, the equityholders find it optimal to issue new debt to finance the investment, I, with a total coupon of $c_1^{(FB)} + c_2^{(FB)}$. In other words, at the first best refinancing date, we require the original bondholders be compensated fairly, and then the equityholders rationally choose the same total coupon for their new capital structure as the first best capital structure.

7 Operating Leverage and Operating Flexibility

As a final extension of our basic model, we investigate how increasing operational leverage combined with operational flexibility has a feedback effect on the choice of capital structure.

Specifically, we extend our study on the impact of flexibility to firms with different operational leverages. For both the flexible and rigid firm, the production cost is $K \cdot Q$, where Q is the production quantity. In our setting, the flexible firm can control $Q = m$ when $S > K$ and 0 otherwise, while the rigid firm always produces at the full capacity, m. So far we have assumed no ongoing fixed costs. To study how flexibility affects the capital structure for firms with different operational leverage, we add an ongoing fixed cost, F, for both flexible and rigid systems. When the fixed cost increases, the firm's operational leverage increases. Intuitively, as the ongoing

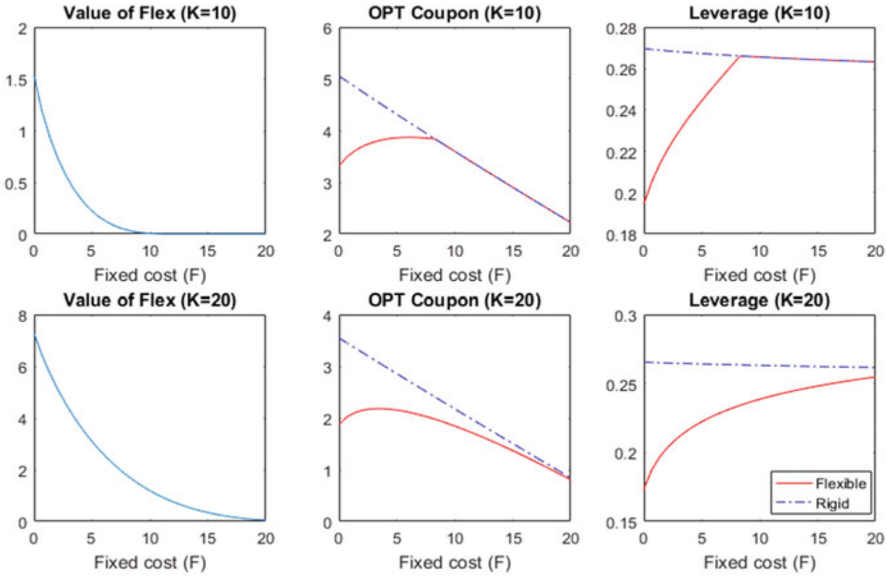

Fig. 6 Comparison of leverage with increasing ongoing fixed costs. The first graph shows the value of technological flexibility defined as the value of the firm with flexible technology minus the value of the firm with rigid technology given all equity financing for an ongoing fixed cost, F, ranging from 0 to 20. The second graph compares the optimal coupon for the flexible and rigid firms as the ongoing fixed cost increases, and the third graph shows the behavior of the resulting leverage of the firms as the ongoing fixed cost increases

fixed costs increase, the advantages of operational flexibility may decrease. Further, since ongoing fixed costs can be likened to ongoing debt obligations, one may think that high fixed costs crowds out debt financing of the investment (Fig. 6).

With this fixed cost F, the earnings process before interest and taxes (EBIT) for the flexible and rigid firms are $EBIT^F = -F + Max[(x-K)m, 0]$ and $EBIT^R = -F + (x - K)m$. We compare the two firms in this setting to show how the effect of flexibility relates to operational leverage. For the rigid firm, Fig. 4 shows that the optimal coupon decreases as the ongoing fixed cost, F, increases, and at the same time, the leverage declines. This result is consistent with the predictions of the operating leverage hypothesis, which states that high fixed costs does indeed crowd out debt.

Now consider the flexible firm. As F increases, the flexible firm becomes less flexible, in that more of its costs are ongoing fixed costs. Just like the rigid firm, the EBIT process of a low F firm stochastically dominates a higher F firm. Yet, as we see from Fig. 4, as F increases, i.e., as operational leverage increases, the coupons *increase* at first, before reaching a point at which the ongoing costs are so high that the flexible firm acts like a rigid firm and leverage starts to decrease.

To highlight the important role that flexibility plays on capital structure, the top left panel computes the difference in value between an unlevered firm using a flexible technology with an otherwise identical firm using a rigid technology. This value provides a pure measure of asset flexibility. Note that as the fixed cost F increases, the pure value of flexibility decreases. In the range where asset flexibility has value (i.e., F is small), financial leverage increases with F; the effect reverses when the value of the asset flexibility decreases (i.e., when F is large). This result stems from the fact that as the ongoing fixed cost, F, increases, the flexible firm loses flexibility and thus has incentives to take on more debt. However, beyond some threshold, the ongoing fixed cost is so high that it competes with ongoing coupon payments, just like a rigid firm.

These result may help explain the fact that different empirical studies arrive at different conclusions about the linkages between operating leverage and capital structure. Our results suggest operating flexibility should be considered as one of the control variables in studying the effects of operational leverage on capital structure.

8 Conclusion

This chapter has investigated how the degree of operational flexibility embedded in technologies affect investment timing and financing when the investments are large and irreversible. Such investments are typically made with significant use of debt. Taxes and deadweight bankruptcy costs have a huge impact on these decisions. In general, we have found that greater operational flexibility is accompanied with less use of debt and earlier investment. The results, for the most part carry over to the case where technology is in place and an existing capital structure is in place. We saw that the under- over-investment problem is somewhat mitigated when the investment is in flexible technology. We investigated how different bankruptcy laws alter decisions and found that they do not appear to change the direction of the results. We also investigated how operating leverage combined with operating flexibility to alter decisions and saw that operating flexibility may counteract operating leverage, while operating rigidity competes with operating leverage. Finally, we investigated how financial contracts could be designed so that first best outcomes could be achieved.

Of course, every model has its limitations. It remains for future studies to investigate how these results would hold up if costly switching costs existed. It also remains for future work to extend this analysis to the case where there is a downward sloping demand curve, as is the case where there is a monopoly. Our overall goal, however, was to provide an overview of how these problems could be addressed through the lens of real option models.

References

Aivazian, V., & Berkowitz, M. (1998). Ex post production flexibility, asset specificity, and financial structure. *Journal of Accounting, Auditing and Finance, 13*(1), 1–20.

Berk, J. B., & DeMarzo, P. M. (2007). *Corporate finance*. Pearson Education.

Billett, M., King, T., & Mauer, D. (2007). Growth opportunities and the choice of leverage, debt maturity, and covenants. *Journal of Finance 6*(62), 697–730.

Birge, J., & Xu, X. (2011). Firm profitability, inventory volatility and capital structure. Working Paper, University of Chicago.

Boyabatli, O., Leng, T., & Toktay, L. (2016). The impact of budget constraints on flexible versus dedicated technology choice. *Management Science, 62*, 225–244.

Boyabatli, O., & Toktay, L. B. (2011). Stochastic capacity investment and flexible versus dedicated technology choice in imperfect capital markets. *Management Science, 57*, 2163–2179.

Brennan, M., & Schwartz, E. (1985). Evaluating natural resource investment. *Journal of Business, 58*, 135–157.

Bris, A., Ravid, S., & Sverdlove, R. (2009). Conflicts in bankruptcy and the sequence of debt issues. Working Paper, Yale International Center for Finance.

Chod, J., & Zhou, J. (2014). Resource flexibility and capital structure. *Management Science, 60*, 708–729.

Dangle, T. (1999). Investment and capacity choice under uncertain demand. *European Journal of Operations Research, 117*, 415–428.

Dixit, A., & Pindyck, R. (1994). *Investment under uncertainty* (1st ed.). Princeton: Princeton University Press

Dugan, M., Minyard, D., & Shriver, K. (1994). A re-examination of the operating leverage-financial leverage tradeoff hypothesis. *The Quarterly Review of Economics and Finance, 34*, 327–334.

Eisdorfer, A. (2008). Empirical evidence of risk shifting in financially distressed firms. *Journal of Finance, 63*, 609–637.

Elsas, R., Flannery, M. J., & Garfinkel, J. A. (2006). Major investments, firm financing decisions, and long-run performance. SSRN Working Paper.

Ferri, M., & Jones, W. (1979). Determinants of financial structure: A new methodological approach. *Journal of Finance, 34*, 631–644.

Fine, C., & Freund, R. (1990). Optimal investment in product-flexible manufacturing capacity. *Management Science, 36*, 449–466.

Frank, M., & Goyal, V. (2005). *Handbook of Corporate Finance: Empirical Corporate Finance, Chapter 7: Tradeoff and Pecking Order Theories of Debt* (1st ed.). Elsevier/North Holland.

Frank, M., & Goyal, V. (2014). The profit Leverage Puzzle Revisited. *Review of Finance, 24*, 1–39.

Goldstein, R., Ju, N., & Leland, H. (2001). An EBIT-based Model of dynamic capital structure. *Journal of Business, 74*, 483–512.

Graham, J. (2000). How big are the tax benefits of debt? *Journal of Finance, 55*, 1901–1941.

Grenadier, S. (2002). Option exercise games: An application to the equilibrium investment strategies of firms. *The Review of Financial Studies, 15*, 691–721.

Hackbarth, D., & Mauer, D. (2012). Optimal priority structure, capital structure, and investment. *Review of Financial Studies, 25*, 747–796.

Hagspiel, V., Huisman, K. J., & Kort, P. M. (2016). Volume flexibility and capacity investment under demand uncertainty. *International Journal of Production Economics, 178*, 95–108.

Huberts, N., Huisman, K., Kort, P., & Lavrutich, M. (2015). Capacity choice in (strategic) real option models: a survey. *Dynamic Games and Applications, 5*, 424–439.

Huffman, L. (1983). Operating leverage, financial leverage and equity risk. *Journal of Banking and Finance, 7*, 197–212.

Kahl, M., Lunn, J., & Nilsson, M. (2014). Operating leverage and corporate financial policies. Working Paper, University of Texas at Austin.

Kuzmina, O. (2013). Operating flexibility and capital structure: Evidence from a natural experiment. Working Paper, New Economic School.
Lederer, P., & Singhal, V. (1994). The effect of financial decisions on the choice of manufacturing technologies. *International Journal of Flexible Manufacturing Systems, 6*, 333–360.
Leland, H. (1994). Corporate debt value, bond covenants, and optimal capital structure. *Journal of Finance, 49*, 1213–1252.
Li, L., Shubik, M., & Sobel, M. J. (2013). Control of dividends, capital subscriptions, and physical inventories. *Management Science, 59*, 1107–1124.
Lord, R. (1996). The impact of operating and financial risk on equity risk. *Journal of Economics and Finance, 20*, 27–38.
MacKay, P. (2003). Real flexibility and financial structure: An empirical analysis. *Review of Financial Studies, 16*, 1131–1165.
Mandelker, N., & Rhee, S. (1984). The impact of the degrees of operating and financial leverage on systematic risk of common stock. *Journal of Financial and Quantitative Analysis, 19*, 45–57.
Mauer, D., & Triantis, A. (1994). Interactions of corporate financing and investment decisions: A dynamic framework. *Journal of Finance, 49*, 1253–1277.
McDonald, R., & Siegel, D. (1986). The value of waiting to invest. *The Quarterly Journal of Economics, 101*, 707–727.
Mello, A., & Parsons, J. (1992). Measuring the agency cost of debt. *Journal of Finance, 47*, 1887–1904.
Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance, 29*, 449–469.
Modigliani, F., & Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American Economic Review, 48*, 261–297.
Myers, S. (1993). Still searching for optimal capital structure. *Journal of Applied Corporate Finance, 6*, 4–14.
Reinartz, S., & Schmid, T. (2016). Production flexibility, product markets and capital structure decisions. *Review of Financial Studies, 29*, 1501–1548.
Ritchken, P., and Wu, Q. (2021). Capacity investment, production flexibility, and capital structure. *Production and Operations Management, 30*(12), 4593–4613.
Shreve, S. (2004). *Stochastic calculus for finance II: Continuous time models*. Springer.
Strebulaev, I. (2007). Do tests of capital structure theory mean what they say? *Journal of Finance, 62*, 1747–1787.
Titman, S., Tompaidis, S., & Tsyplakov, S. (2004). Market imperfections, investment flexibility, and default spreads. *The Journal of Finance, 59*, 165–205.
Van-Horne, J. (1977). *Financial management and policy* (1st ed.). Upper Saddle River, NJ: Prentice Hall.
Van-Mieghem, J. (1998). Investment strategies for flexible resources. *Management Science, 44*, 1071–1078.