Annals of Mathematical Sciences and Applications
Volume 5, Number 2, 317–345, 2020

Statistical models and stochastic optimization in financial technology and investment science

Tze L. Lai*†, Shih-Wei Liao, Samuel P. S. Wong, and Huanzhong Xu

In the decade since the global financial crisis and the Great Recession that followed, the financial technology – or FinTech – revolution has transformed financial markets and services through the automation of trading and risk management, among other things. The "ABCD" of cutting-edge FinTech are AI (Artificial intelligence), Blockchains, Cloud computing and big Data. We first review these technologies of modern FinTech and some of the underlying mathematical foundations. We then describe new statistical models and stochastic optimization methods in FinTech and investment science.

Keywords and phrases: Blockchains, cryptographic hash functions, collision resistance, empirical Bayes, portfolio optimization.

1. Introduction

In this section, we review modern investment science and the ABCD of FinTech. Artificial intelligence refers to the intelligence demonstrated by machines, in contrast with "natural intelligence" displayed by humans and animals. The August 21, 2015 issue of Business Insider of *Financial Times* lists the leaders of AI applications in the global FinTech industry. Most of them are located in the US and China, and focus on the areas of lending, payment, money transfer, and insurance. Ant Financial, with headquarters in Hangzhou of China's Zhejiang Province, is the largest, with a valuation of US$45.5bn at that time and $75bn by late 2016. Its financial services include internet and mobile payment systems, consumer and SME (small and medium-sized enterprise) lending, insurance (with its insurance robo-advisor), personal wealth management (with its wealth robo-advisor), and institutional wealth management, which make it look like a tradional bank even though it is actually a FinTech company with many state-of-the-art

*Corresponding author.
†T.L. Lai is partially supported by NSF DMS-1407828 and DMS-1811818.

technologies deployed to make the services more efficient and far-reaching. AI is assuming an increasingly important role in traditional banking as it provides technologies such as voice recognition, natural language processing, and computer vision for user-account management and fraud detection, machine learning methods and deep learning networks for anti-money-laundering and credit modeling. Mobile and internet payment systems are closely connected to cloud computing. The past ten years have withnessed increasing adoption of cloud computing by financial institutions around the globe. As a highly regulated industry, there are many challenges for the financial industry that handles sensitive personal information to use cloud computing for core business processes such as credit risk management and customer services. Cloud service providers have worked with financial institutions and regulators to address the security and compliance requirements, mitigating the early concerns about privacy and data security in the cloud. PricewaterhouseCoopers (PwC) predicts that the cloud will become a dominant infrastructure model in FinTech 2020.

In the remainder of this section, We describe the mathematical underpinnings of Blockchains and Data in FinTech and provide the background for the developments in Sections 2 an 3.

1.1. Data, models, optimization and algorithms in quantitative trading

A quantitative trading strategy is an investment strategy based on quantitative analysis of financial markets and prediction of future performance that it tries to optimize. The time-scale of an investment strategy defines the horizon of the investment's future return to be considered. Before high-frequency data of reliable quality were available, the daily closing prices of stocks constituted the main data source for investment analysis and strategic decisions. Modeling the daily returns has been a fundamental problem in financial economics and models of "speculative prices", using the terminology of the 1970 Economic Sciences Nobel Prize winner Paul Samuelson [44], have evolved from Brownian motion and geometric Brownian motion (GBM) to stable processes and subordinated processes, culminating in the more general Lévy processes that also allow jumps and have become popular in the recent literature. However, the implicit assumption of i.i.d., or independent but possibly non-identically distributed, returns in these models do not match stylized facts of daily (or other low-frequency) returns of stocks. New mathematical models that relate data to decisions have played an important role in the development of quantitative trading strategies. In

particular, as described in the recent monograph [23], there have been developments that use linear regression with time series regressors, which are lagged or exogenous variables, and martingale difference errors that will be discusses further in Section 2.

Major advances in information technology in the 1980s led to electronic trading platforms that became widely adopted in the decade, beginning in 1986 when the London Stock Exchange moved to electronic trading. Transaction prices are quoted in discrete units or ticks. On New York Stock Exchange (NYSE), the tick size was \$0.125 before June 24, 1997, and \$0.0625 afterward until January 29, 2001, when all NYSE stocks started to trade in decimals. There is a large literature in financial econometrics on models and methods for tick-by-tick transactions data, including market microstructure noise models that have negative lag 1 autocorrelation for (logarithmic) price changes under the efficient market hypothesis (EMH), estimation of the integrated variance of the efficient price process in the presence of market microstructure noise, econometric models of inter-transaction durations, and predictive models relating low-frequency to high-frequency volatilities, as described in [23] that also links the low-frequency time scale (daily, weekly or monthly) to the smaller time scale of seconds to hundredth of a second in algorithmic trading. The two time-scales correspond to the two stages by which traders slice and place their buy and sell orders to electronic exchanges. The first stage, known as optimal execution, optimally slices big orders into smaller ones on a daily basis to minimize the price impact, and the second stage optimally places the orders within seconds. It is known as optimal placement within one exchange, or smart order routing across different exchanges. The sequential nature of these algorithmic trading strategies leads to recursions for updating them as new data arrive.

In high-frequency trading (HFT), a few microseconds can make a significant difference. Therefore the proximity of the server boxes (where the strategy components reside) to the exchange data center (where the matching process takes place and the market data originates) is a significant factor that can affect the overall latency of a strategy. Colocation is a term that refers to the practice of placing trade servers at or close to the exchange data center. Exchanges that provide colocation services are required to offer equal access to all participants. In particular, the physical cables that link all servers to their gateways are required to be of the same length, and hence all colocated clients are subject to the same inbound and outbound latencies. For colocation services, exchanges charge a fee that can depend on the space and power consumption that client servers require. Latency arbitrage in HFT refers to trading assets that are highly correlated and sometimes even equivalent. For example, the exchange-traded fund SPY and the

E-mini S&P 500 futures contract traded on Chicago Mercantile Exchange (CME) under the ticker symbol ES, are both based on the S&P 500 Index, ignoring the dividend and interest rate parts. However, ES is traded on the CME's platform in Chicago and SPY is traded on the platforms of several exchanges based in the East Coast. When ES moves up, SPY should move up too, albeit with a delay. This delay is a function of the speed at which the arbitrageurs can operate. This latency can be in the range of tens of milliseconds, mostly due to the geographic separation of the two exchanges: the distance between Chicago and New York is roughly 714 miles, resulting in delay between the movements of SPY and ES. Latency arbitrage strategies focus on this type of opportunities, and they serve as a facilitator for information transfer between the two trading hubs.

As trading decisions are made at split-second intervals, the task of ensuring that the algorithm and the infrastructure are error-free in a HFT firm pertains to both software engineering and LOB analytics. Trading algorithms are required to undergo multiple stages of testing and certification, with duplicate checks in place. Seemingly inconspicuous bugs such as integer overflow or underflow could have a significant impact on the strategy. Operational risk in this context mainly refers to the risk stemming from infrastructure disruptions and software bugs. A case in point was the 2012 software error of the Knight Capital Group that deployed on August 1 untested software, which contained an obsolete function because a technician forgot to copy the new code to one of its servers for automated routing of equity orders. This caused major disruption in stock prices within 45 minutes and a pre-tax cost of $440 million and subsequent drop of over 70% of the company's stock price. On August 5, 2012, Knight Capital raised about $400 million from several major investors to stay in business. It was subsequently acquired by the Global Electronic Trading Company (Getco LLC).

Is HFT socially useful? This question was asked of James Simons, founder of the highly profitable algo trading company Renaissance Technologies, in 2010 after the Flash Crash. He answered affirmatively because "highly liquid markets are socially useful" and said that with HFT the market "came right back within minutes" after the initial dive, in contrast to 1987 when the stock market "went down 25% in half a day" on October 19 (Black Monday) when a large imbalance between the volume of sell orders and buy orders arose immediately after the opening of NYSE due to pent-up pressure to sell stocks due to general worries of over-valuation of stocks and news of worsening economic indicators, and "didn't recover for six months." Figure 1, which plots the closing prices and traded volumes of the S&P 500 Index in the 12-month periods April 1, 1987–March 31, 1988 and January 1–December 31, 2010, confirms his point.

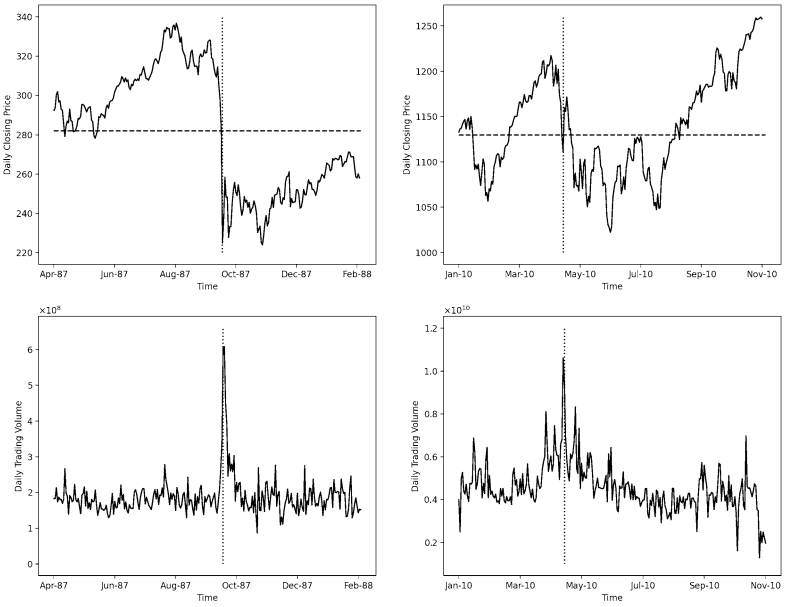

Figure 1: Daily closing price (top) and traded volume (bottom) of the S&P 500 Index for Black Friday 1987 (left) and Flash Crash 2010 (right), with the dotted vertical lines showing the event dates and the dashed horizontal lines the closing prices prior to the events in the top panel.

1.2. Blockchains, cryptography and mathematical foundations

On the occasion of the 80th anniversary of United Overseas Bank in November, 2015, Prime Minister H.L. Lee of Singapore (who graduated with first-class honours from Cambridge University in mathematics in 1974) said: "Blockchains, which are used for bitcoin, can also be used for many other applications like real-time gross settlement, or financial transactions verification, so our banks and our regulators must keep up to date and up to scratch with these developments." Narayanan et al. [41] say in their introductory chapter that Cryptography, which provides a mechanism for security encoding the rules of a cryptocurrency system... "is a deep academic research field using many advanced mathematical techniques." These techniques include hash functions, computational number theory, complexity theory, elliptic curves, digital signatures and public/private key enciphering and deciphering, information theory and probability; see [25] for details. To overcome certain weakness of deterministic encryption which always produces the same ciphertext for a given plaintext and key, Goldwasser and

Micali [22] introduced a new model of probabilistic encryption in 1984. Deterministic encryption is not secure against eavesdropping; the eavesdropper might perform statistical analysis of messages transmitted over a cryptosystem to learn x from $f(x)$ where x is of a special form, or to compute some partial information about x, even though f is a trapdoor function, "which is easy to compute but difficult to invert unless some secret information, the trapdoor, is known." Cryptosystems can be broadly divided into 2 classes: stream ciphers which process the plaintext into small chunks (bits or characters), and block ciphers which act in a combinatorial fashion on large blocks of text, as pointed out by Diffie and Hellman [13]. Goldwasser and Micali [22] "replace deterministic block encryption by probabilistic encryption of single bits, where there are many different encodings of 1 and many different encodings of 0 (and uses) a fair coin to encrypt each message, the encoding (of which) will depend on the message plus the result of a sequence of coin tosses... (hence) there are many possible encodings for each message (but) messages are always uniquely decodable (by) the legal receiver of a message, who knows the trapdoor information, but provably hard for an adversary."

Bitcoin is not the only cryptocurrency. Many difference *altcoins* – alternative to bitcoin – emerged as the value of bitcoin increased. A well-known example is ethereum. Dan Boneh [7] who has been the founding director of the Center for Blockchain Research at Stanford since August 2018 (http://cbr.stanford.edu), says that the swell of excitement in blockchain and cryptocurrencies among coders and computer scientists has not been witnessed since the earliest days of the internet, and that "cryptocurrencies are a wonderful way to teach cryptography." Moreover, to address the computational challenges, bitcoin mining trading is dominated by application-specific integrated circuits (ASICs), which are chips designed, built and optimized for mining bitcoins, yielding arguably the "fastest turnaround time – from specifying a problem to delivering working chips – in the history of integrated circuits."[41] Hence cryptocurrencies have indirectly advanced the blockchain technology.

As *The Economist* explained in its 31 October 2015 issue, financial systems have long operated on the basis of trust, for which banks and governments have served to provide top-down control of monetary value. Now, however, bottom-up "trust machines" are emerging through blockchain technology to provide immutable shared ledgers to exchange information digitally and determine value by consensus, as exemplified by bitcoin and other cryptocurrencies. Similar to debates concerning printed money over fifty years ago in the celebrated work of the Nobel laureate Milton Friedman on the monetary history in the U.S. [19, 20], there is an ongoing debate

over whether bitcoin and other cryptocurrencies actually achieve "trust", or "mischief or mistrust". Economists have questioned whether cryptocurrencies are tangible assets and criticized them as a cause of speculative bubbles. Another Nobel laureate Paul Krugman [30] calls bitcoin "a bubble wrapped in techno-mysticism inside a cocoon of libertarian ideology", saying that it "lacks a tether to reality". He adds: "Although the modern dollar is a fiat currency, not backed by any other asset, like gold, its value is ultimately backed by the fact that the U.S. government will accept it, ... its purchasing power is also stabilized by the Federal Reserve" via monetary policy. In contrast, bitcoin is categorized as a decentralized virtual currency by the U.S. Treasury, as a commodity by the Commodity Futures Trading Commission, and as an intangible asset by Canada, South Africa, the Czech Republic and several other countries. Although bitcoins and altcoins are not expected to provide substitutes for dollar bills (or even $100 bills) in ordinary (or not so ordinary) transactions, they can be useful for payment and settlement in specialized markets. Narayanan et al. [41, Section 9.5] describe the example of "prediction markets" and how altcoins can be used for accepting payments and distributing payouts in a decentralized prediction market with a decentralized order book.

Blockchain technology also much broader applications in FinTech than providing a platform for the bitcoin and altcoin systems, and has been described as the future of the sharing economy [2], the next frontier for online marketplaces [42], and a powerful solution to optimize financial transactions and improve efficiency, security and trust [26]. Moreover, as Boneh [7] points out, "blockchain is rich in possibility" even though the technology is still in its early days. There are already blockchain-based solutions to food and drug traceability, interoperability of electronic health and medical records, which in turn open up new opportunities and challenges to computer science and statistics. In particular, Section 3.2 describes the challenge of "zero-knowledge proofs for fast verification" in blockchain technology and recent breakthroughs in computer science.

The blockchain technology has recently provided a new form of cryptocurrency issued by JP Morgan Chase on 14 February, 2019 dubbed JPM Coin, which has a fixed value redeemable for one US dollar. This digital token is designed to be used by major institutional customers, not by individuals such as bitcoin miners or other cryptocurrency speculators, to transfer cross-border payments or corporate debt services on the blockchain network called Quorum that replaces old technology such as wire transfers.

2. A martingale regression model of equity prices and portfolio optimization under risk constraints

As noted in the first paragraph of Section 1.1, new mathematical models for equity prices have been developed in the past decade to relate financial data to investment decisions such as portfolio optimization and risk management. Section 2.1 describes some recent developments in this direction, beginning with works since the turn of the century to address the "Markowitz" portfolio optimization enigma" – whether 'optimized' is optimal; see Michaud [39]. This section generalizes the mathematical underpinnings of the key stochastic optimization advances and shows how these generalizations can be used for portfolio optimization under convex risk constraints. In particular, we describe an empirical Bayes approach to statistical modeling that is also relevant to blockchain collision analysis in Section 3.1.

2.1. Portfolio optimization enigma and approaches to address it

The 1990 Nobel Prize in Economic Sciences was awarded to Harry Markowitz and William Sharpe for their fundamental contributions to portfolio theory and to Merton Miller for fundamental contributions to the theory of corporate finance. Markowitz's classical single-period, mean-variance portfolio optimization theory in the 1950s [37, 38] gives the efficient frontier in the mean versus volatility plane for portfolios consisting of m stocks with expected returns $\mu_1, \ldots, \mu_m$ and covariance matrix $\mathbf{\Sigma}$ of the stock returns $r_1, \ldots, r_m$. Let $\mathbf{w} = (w_1, \ldots, w_m)^T$ be the vector of portfolio weights such that $\sum_i^m w_i = 1$ and let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$, $\mathbf{r} = (r_1, \ldots, r_m)^T$. While Markowitz derived the efficient frontier from scratch by geometric arguments, advances in the mathematics of optimization show that his problem is basically that of multi-objective optimization, with two objective functions $f_1(\mathbf{w}) = E(\mathbf{w}^T\mathbf{r})$ and $f_2(\mathbf{w}) = -\mathrm{Var}(\mathbf{w}^T\mathbf{r})$, and its Pareto optimal set corresponds to the efficient frontier. The theory assumes known $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, and Markowitz [37] shows how to compute the weight vector $\mathbf{w}_{\text{eff}}$ corresponding to a target value $\mu_* = \mathbf{w}^T\boldsymbol{\mu}$ of a portfolio on the efficient frontier that minimize $\mathrm{Var}(\mathbf{w}^T\mathbf{r}) = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ subject to inequality constraints on the feasible short selling.

In practice, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are unknown and a natural idea is to replace them by the sample mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\mathbf{\Sigma}}$ of a training sample of size n consisting of historical returns. Frankfurther et al. [18] and Jobson and Korkie [27] have reported that these estimated ("plug-in") portfolios can perform worse than the highly inefficient equally-weights portfolio of the m

stocks. The difficulty of estimating $\boldsymbol{\mu}$ well enough for the plug-in portfolio to have reliable performance has then been noted by Jorion [28] who proposes to use for $\hat{\boldsymbol{\mu}}$ a shrinkage estimator similar to the Bayes estimator, while Ledoit and Wolf [35] propose to shrink $\hat{\boldsymbol{\Sigma}}$ towards a structured covariance matrix. Lai et al. [33] carry out an extensive empirical study of other proposals to plug into the efficient frontier and Michaud's [39] proposal to use bootstrap resampling to rectify the plug-in portfolios, and find that their improvements over the classical plug-in portfolio are minimal in comparison with the NPEB (nonparametric empirical Bayes) portfolio developed therein.

2.2. Bayesian generalization of mean-variance portfolio optimization for unknown $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$

A major difficulty with the plug-in efficient frontier (which replaces the unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by their sample counterparts or Bayes/shrinkage estimates based on a sample $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ of current and past return vectors) is that Markowitz's idea of using $\mathrm{Var}(\boldsymbol{w}^T \boldsymbol{r}_{n+1})$ as a measure of portfolio's risk is no longer valid when it and $E(\boldsymbol{w}^T \boldsymbol{r}_{n+1})$ are replaced by estimates that have sampling distributions. To resolve this difficulty, Lai et al. [33] started by generalizing Markowitz's optimization problem to the Bayes decision problem

$$\max_{\boldsymbol{w}} \{E(\boldsymbol{w}^T \boldsymbol{r}_{n+1}) - \lambda \mathrm{Var}(\boldsymbol{w}^T \boldsymbol{r}_{n+1})\} \tag{2.1}$$

with a prior distribution on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ so that the maximum in (2.1) is over weight vectors $\boldsymbol{w}$ whose components sum to 1 and which are functions of the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$. Note that if the prior distribution puts all its mass at the actual parameter value $(\boldsymbol{\mu}^{\boldsymbol{a}}, \boldsymbol{\Sigma}^{\boldsymbol{a}})$, then the Bayes decision problem reduces to Markowitz's optimization problem that assumes $\boldsymbol{\mu}^{\boldsymbol{a}}$ and $\boldsymbol{\Sigma}^{\boldsymbol{a}}$ to be given. The Lagrangian multiplier λ in (2.1) can be interpreted as the investor's risk-aversion index when variance is used to measure risk.

Standard methods to compute Bayes rules maximizing the expected reward $E(X; d)$ over decision rules d by using the law of iterated expectations

$$E(X; d) = E\{E[X; d(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)] | \boldsymbol{r}_1, \ldots, \boldsymbol{r}_n\}$$

are not applicable to (2.1) because $\mathrm{Var}(X) = EX^2 - (EX)^2$ involves the square of an expectation of $X = \boldsymbol{w}^T \boldsymbol{r}_{n+1}$. The approach used in [33] is to convert (2.1) to a family of Bayes decision problems, indexed by a parameter

$\eta \in \mathbb{R}$, such that each problem involves only expecatations. Specifically, (2.1) can be written in the form

$$\max_{\eta}\{E[\boldsymbol{w}^T(\eta)\boldsymbol{r}_{n+1}] - \lambda \mathrm{Var}[\boldsymbol{w}^T(\eta)\boldsymbol{r}_{n+1}]\}, \text{ where} \tag{2.2}$$

$$\boldsymbol{w}(\eta) = \arg\min_{\boldsymbol{w}}\{\lambda E[(\boldsymbol{w}^T\boldsymbol{r}_{n+1})^2] - \eta E(\boldsymbol{w}^T\boldsymbol{r}_{n+1})\}, \tag{2.3}$$

as we now demonstrate. Letting $W = \boldsymbol{w}^T\boldsymbol{r}_{n+1}$ and $W_B = \boldsymbol{w}_B^T\boldsymbol{r}_{n+1}$, where $\boldsymbol{w}_B$ is the Bayes weight vector, Lai et al. [33] write $E(W) - \lambda \mathrm{Var}(W) = h(EW, EW^2)$, where $h(x, y) = x + \lambda x^2 - \lambda y$, and note that

$$\begin{aligned} 0 &\geqslant h(EW, EW^2) - h(EW_B, EW_B^2) \\ &= EW - EW_B - \lambda(EW^2 - EW_B^2) + \lambda((EW)^2 - (EW_B)^2) \\ &= EW - EW_B - \lambda(EW^2 - EW_B^2) + \lambda(EW - EW_B)^2 \\ &\quad + 2\lambda EW_B(EW - EW_B) \\ &\geqslant (1 + 2\lambda EW_B)(EW - EW_B) - \lambda(EW^2 - EW_B^2) \\ &= (\lambda EW_B^2 - \eta EW_B) - (\lambda EW^2 - \eta EW), \end{aligned} \tag{2.4}$$

where $\eta = 1 + 2\lambda EW_B$, and inequality is strict unless $EW = EW_B$ and $EW^2 = EW_B^2$ (i.e., W and W_B have the same mean and variance). This shows that (2.1) is equivalent to minimizing $\lambda E(\boldsymbol{w}^T\boldsymbol{r}_{n+1})^2 - \eta E(\boldsymbol{w}^T\boldsymbol{r}_{n+1})$ over weight vectors $\boldsymbol{w}$ that can depend on $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$. Since η is linear function of EW_B and W_B is yet to be determined, [33] cannot apply the equivalence directly and instead solves a family of Bayes decision problems (2.3) indexed by η and then performs one-dimensional search over η to minimize (2.2).

We next consider computation of $\boldsymbol{w}(\eta)$ in (2.3). Let $\boldsymbol{\mu}_n$ and $\boldsymbol{V}_n$ be the posterior mean and second moment matrix given $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$. Without short selling, the weight vector is given by

$$\boldsymbol{w}(\eta) = \arg\min_{\boldsymbol{w} \geqslant 0 : \boldsymbol{w}^T\mathbf{1} = 1} (\lambda \boldsymbol{w}^T\boldsymbol{V}_n\boldsymbol{w} - \eta \boldsymbol{w}^T\boldsymbol{\mu}_n), \tag{2.5}$$

which can be computed by quadratic programming (e.g., `quadprog` in `MATLAB`). When short selling is allowed up to certain limits, $\boldsymbol{w} \geqslant \mathbf{0}$ in (2.5) is replaced by $\boldsymbol{w} \geqslant \boldsymbol{w}_0$, where $\boldsymbol{w}_0$ is a vector of negative numbers. When there is no limit on short selling, the constraint $\boldsymbol{w} \geqslant \mathbf{0}$ in (2.5) is removed and $\boldsymbol{w}(\eta)$ is given explicitly by

$$\boldsymbol{w}(\eta) = \frac{\boldsymbol{V}_n^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{V}_n^{-1}\mathbf{1}} + \frac{\eta}{2\lambda}\boldsymbol{V}_n^{-1}\left(\boldsymbol{\mu}_n - \frac{\boldsymbol{\mu}_n^T\boldsymbol{V}_n^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{V}_n^{-1}\mathbf{1}}\mathbf{1}\right).$$

2.3. From Bayes to empirical Bayes and then to nonparametric bootstrap

Besides specifying a prior distribution to $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, (2.1) also requires specification of the common distribution of the returns $\boldsymbol{r}_i$ (assumed to be i.i.d. by Markowitz and others who interpret an efficient financial market as a "random walk down Wall Street" [36]) to evaluate $E_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}_n^T \boldsymbol{r}_{n+1})$ and $\mathrm{Var}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}_n^T \boldsymbol{r}_{n+1})$, where $\boldsymbol{w}_n$ is the weight vector for a chosen portfolio and may depend on $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$. Lai et al. [33] use the nonparametric bootstrap, which samples $\{\boldsymbol{r}_{b_1}^*, \ldots, \boldsymbol{r}_{b_n}^*\}$, drawn with replacement from the observed sample $\{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n\}$ for $1 \leqslant b \leqslant B$, to estimate

$$
\begin{aligned}
E_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}_n^T \boldsymbol{r}_{n+1}) &= E_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}_n^T \boldsymbol{\mu}) \\
\mathrm{Var}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}_n^T \boldsymbol{r}_{n+1}) &= E_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}_n^T \boldsymbol{\Sigma} \boldsymbol{w}_n) + \mathrm{Var}(\boldsymbol{w}_n^T \boldsymbol{\mu}).
\end{aligned} \tag{2.6}
$$

They use Bayes or empirical Bayes (EB) estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (2.6); the EB approach allows the prior distribution in the Bayesian model to include unspecified hyperparameters which can be estimated from the training sample by maximum likelihood or generalized method of moments. In particular, using a scale hyperparameter τ for a given distribution of $\tau^{-1}\mathrm{vec}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and letting $\tau \to \infty$ leads to a nonparametric EB (NPEB) implementation of (2.1), for which $\boldsymbol{\mu}_n$ and $\boldsymbol{V}_n$ in (2.5) are the sample mean vector $\bar{\boldsymbol{r}} = n^{-1} \sum_{i=1}^n \boldsymbol{r}_i$ and second moment matrix $n^{-1} \sum_{i=1}^n \boldsymbol{r}_i \boldsymbol{r}_i^T$ since the posterior distribution converges to the empirical distribution of $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ as $\tau \to \infty$.

2.4. Extension of NPEB to incorporate time series features of $\boldsymbol{r}_i$ via martingale regression

Lai and Xing [32] describe diagnostic checks and statistical tests of the i.i.d. assumption on daily (or weekly) asset returns in their Chapters 5 and 6, and Lai et al. [33] show how the NPEB approach described in the preceding subsection can be extended to incorporate these "stylized facts" of the time series of asset returns. The year 1982 marked the publication of Robert Engle's groundbreaking paper [15] on ARCH (autoregressive conditional heteroscedastic) models, for which he was awarded the 2003 Nobel Prize in Economic Sciences. In that year, Lai and Wei [31] published their seminal paper on "stochastic regression", which laid the groundwork for the martingale regression model described in this subsection and which was

developed for the analysis of stochastic input-output systems of the form

$$y_t = \boldsymbol{\beta}^T \boldsymbol{x}_t + \epsilon_t, \tag{2.7}$$

in which y_t represents the output at time t, the regressor $\boldsymbol{x}_t$ is a vector depending on past inputs and outputs and is therefore $\mathcal{F}_{t-1}$-measurable, where $\mathcal{F}_s$ is the σ-field generated by $\{(\boldsymbol{x}_i, y_i), i \leqslant s\}$, and the random disturbance ϵ_t satisfies $E(\epsilon_t|\mathcal{F}_{t-1}) = 0$, i.e., $\{\epsilon_t, t \geqslant 1\}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t, t \geqslant 1\}$. Note that the autoregressive model AR(p) is a special case of (2.7) with $\boldsymbol{x}_t = (y_{t-1}, \ldots, y_{t-p})^T$, and so is the ARX($p, r$) model $y_t = \beta_1 y_{t_1} + \cdots + \beta_p y_{t-p} + \beta_{p+1} u_{t-1} + \cdots + \beta_{p+r} u_{t-r}$, in which u_s is the exogenous input at time s. The assumption of martingale difference (instead of i.i.d. zero-mean) random errors ϵ_t in (2.7) allows Lai et al. [33] to incorporate volatility changes in modeling ϵ_t.

The ARCH model has been generalized to the GARCH(h, k) model of the form

$$\epsilon_t = \sigma_t \zeta_t, \ \sigma_t^2 = \omega + \sum_{i=1}^{h} b_i \sigma_{t-i}^2 + \sum_{j=1}^{k} a_j \epsilon_{t-j}^2, \tag{2.8}$$

in which ζ_t are i.i.d. with mean 0 and variance 1; ARCH(k) corresponds to the case without $\sum_{i=1}^{h} b_i \sigma_{t-i}^2$. Letting $\eta_t = \epsilon_t^2 - \sigma_t^2$, which forms a martingale difference sequence, we can write (2.8) as an ARMA model for ϵ_t^2:

$$\epsilon_t^2 = \omega + \sum_{j=1}^{\max(h,k)} (a_j + b_j)\epsilon_{t-j}^2 + \eta_t - \sum_{i=1}^{h} b_i \eta_{t-i},$$

in which $a_j = 0$ for $j > k$ and $b_j = 0$ for $j > h$; see [32, p. 147]. Lai et al. [33] extend NPEB to the following "martingale regression model" for the returns r_{it} of the ith asset at time t:

$$r_{it} = \boldsymbol{\beta}_i^T \boldsymbol{x}_{i,t-1} + \epsilon_{it}, \ \epsilon_{it} = s_{i,t-1} z_{it}, \ s_{i,t-1}^2 = \omega_i + a_i s_{i,t-2}^2 + b_i r_{i,t-1}^2, \tag{2.9}$$

where the components of $\boldsymbol{x}_{i,t-1}$ include 1, factor variables such as the return of a market portfolio like S&P500 at time $t-1$, and lagged variables $r_{i,t-1}, r_{i,t-2}, \ldots$ The random disturbances ϵ_{it} are assumed to be martingale differences, as in (2.7), that undergo dynamic changes in volatility via the GARCH(1, 1) model for $\epsilon_{it} = s_{i,t-1} z_{it}$, with i.i.d. z_{it} that have mean 0 and variance 1. The regression parameter $\boldsymbol{\beta}_i$ can be estimated by least squares,

whereas the GARCH$(1,1)$ parameter vector (ω_i, a_i, b_i) in (2.9) can be estimated by applying maximum likelihood to the residuals $r_{it} - \boldsymbol{\beta}_i^T \boldsymbol{x}_{i,t-1}$ under the "working model" that z_{it} has a standardized Student's t-distribution t_{ν_i} in which $\nu_i \geqslant 2$ is treated as an unknown parameter; see [32, p. 149–151] where the function `garchfit` in the `MATLAB GARCH` toolbox is used (the `R` function `fGARCH` in `cran.r` can be used as an alternative). It should be emphasized that this GARCH$(1,1)$ parameter estimate is actually QMLE (quasi-maximum likelihood estimate) in the martingale regression model in (2.9) with i.i.d. z_{it} that have mean 0 and variance 1. Assuming a working model of $N(0,1)$ for z_{it} yields $\hat{\boldsymbol{\beta}}_i$ as QMLE of $\boldsymbol{\beta}_i$, and further relaxing the $N(0,1)$ model to a standardized t_{ν_i} parametric model yields QMLE of ν_i and ω_i, a_i, b_i. Despite potential model misspecification, QMLE can still be consistent and asymptotically normal, as shown by White [46] and Lee and Hansen [34].

Since (2.9) produces i.i.d. $\boldsymbol{z}_t = (z_{1t}, \ldots, z_{mt})^T$, the NPEB approach can still be used to determine the optimal weight vector, bootstrapping from the estimated common distribution of $\boldsymbol{z}_t$, as carried out by Lai et al. [33]. The NPEB approach yields the following formulas for $\boldsymbol{\mu}_n$ and $\boldsymbol{V}_n$ in (2.5):

$$\boldsymbol{\mu}_n = (\hat{\boldsymbol{\beta}}_1^T \boldsymbol{x}_{1,n-1}, \ldots, \hat{\boldsymbol{\beta}}_m^T \boldsymbol{x}_{m,n-1})^T, \ \boldsymbol{V}_n = \boldsymbol{\mu}_n \boldsymbol{\mu}_n^T + (\hat{s}_{i,n} \hat{s}_{j,n} \hat{\sigma}_{ij})_{1 \leqslant i,j \leqslant m},$$

in which $\hat{s}_{l,n}^2$ is the QMLE of $s_{l,n}^2$ in (2.9) and $(\hat{\sigma}_{ij})_{1 \leqslant i,j \leqslant m}$ is the sample covariance matrix of the residuals of $r_{in} - \hat{\boldsymbol{\beta}}_{in}^T \boldsymbol{x}_{i,n-1}$ $(i = 1, \ldots, m)$.

2.5. Sharpe's CAPM, information ratio and choice of $\boldsymbol{\lambda}$

In 1964, building on Markowitz's mean-variance portfolio optimization theory, Sharpe who shared the Nobel prize with Markowitz and Miller in 1990 published his foundational paper [45] on the Capital Asset Pricing Model (CAPM) that develops economy-wide implications of the trade-off between return and risk, assuming that the market has a risk-free asset with return r_f (interest rate) besides the m risky assets in Markowitz's theory and that all investors have homogeneous expectations and hold mean-variance optimal portfolios. Allowing lending and borrowing of the risk-free assets at rate r_f, the efficient frontier (Pareto optimal set) in CAPM is a straight line, called the "capital market line", that is tangent to Markowitz's efficient frontier for the m risky assets at a point M, which can be interpreted as an index fund or market portfolio. The "one fund theorem" states that any efficient portfolio can be constructed as a linear combination of the fund and the risk-free asset. Hence the capital market line can be defined by the linear

equation $\mu = r_f + \sigma(\mu_M - r_f)/\sigma_M$, where μ is the mean and σ^2 the variance of the return of an efficient portfolio and μ_M and σ_M^2 are those for the market portfolio. The *Sharpe ratio* of a portfolio whose return has mean μ and standard deviation σ is $(\mu - r_f)/\sigma$, which is the expected excess return per unit of risk. For an efficient portfolio, its Sharpe ratio is the same as that of the market portfolio. Instead of using a risk-free asset (such as U.S. Treasury bond) as the benchmark, the *information ratio* uses a surrogate of the market portfolio (such as S&P500) or another market index such as Dow Jones or Nasdaq Composite as the benchmark with return r_b and is defined by $E(r - r_b)/\sqrt{\mathrm{Var}(r - r_b)}$, which is the expected excess return of a portfolio with return r over the (risky) benchmark per unit of risk measured by the standard deviation of $r - r_b$. The information ratio is often annualized (see [32, p. 66]) and is a commonly used measure of a fund's performance relative to other funds.

As we have noted in the first paragraph of Section 2.2, λ in (2.1) represents an investor's risk-aversion parameter which may be difficult to specify, especially when risk is measured by the variance of a portfolio's return. Since information ratio is often used to measure a portfolio's performance, Lai et al. [33] regard λ in (2.1) as a tuning parameter for the weight vector $\boldsymbol{w}_\lambda$, and choose it over a grid to maximize the information ratio

$$E_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}_\lambda^T \boldsymbol{r} - r_b)/\sqrt{\mathrm{Var}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}_\lambda^T \boldsymbol{r} - r_b)}. \tag{2.10}$$

The mean and variance in (2.10) can be estimated by the bootstrap method, similar to NPEB that estimates $E_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}^T(\eta)\boldsymbol{r}) - \lambda \mathrm{Var}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{w}^T(\eta)\boldsymbol{r})$ by the bootstrap.

2.6. Generalization of variance to convex risks of centered returns and an empirical study

Harvey and Siddique [24], Dittmar [14], Ang et al. [1] and You and Daigler [47] proposed to replace variance as a measure of risk by higher moments of centered returns, with mean centering (which subtracts the mean of a random variable from all observations on the variable). We now show how the martingale regression model (2.9) and the NPEB approach can be extended to convex functions of the centered returns, whose expectations include the higher moments advocated by these authors.

Let $\psi : \mathbb{R} \to [0, \infty)$ be a convex function such that $\psi(0) = 0$ and consider the Bayes decision problem, with prior distribution on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for

$$\max_{\boldsymbol{w}} \{E(\boldsymbol{w}^T \boldsymbol{r}_{n+1}) - \lambda E\psi(\boldsymbol{w}^T \boldsymbol{r}_{n+1} - E\boldsymbol{w}^T \boldsymbol{r}_{n+1})\} \tag{2.11}$$

over weight vectors $\boldsymbol{w}$ whose components sum to 1 and which are functions of the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$. The case $\psi(x) = x^{2k}$ reduces to (2.1) for $k = 1$, and to higher moments of centered returns for $k = 2, 3, \ldots$. Let $W = \boldsymbol{w}^T \boldsymbol{r}_{n+1}, W_B = \boldsymbol{w}_B^T \boldsymbol{r}_{n+1}$, and assume that ψ is continuously differentiable in some neighborhood of EW_B. Let $\eta = 1 + \lambda\psi'(EW_B)$. Then by convexity,

$$\begin{aligned}
&\psi(W - EW) - \psi(W_B - EW_B) \\
=&\{\psi(W - EW) - \psi(W)\} + \psi(W) - \psi(W_B) + \{\psi(W_B) - \psi(W_B - EW_B)\} \\
\geqslant&\psi'(EW_B)(-EW + EW_B) + \psi(W) - \psi(W_B).
\end{aligned}$$

Hence the same argument as that in (2.4) yields

(2.12)
$$\begin{aligned}
0 \geqslant\ & EW - EW_B + \lambda\psi'(EW)(EW - EW_B) + \lambda\{E\psi(W_B) - E\psi(W)\} \\
=\ & (\lambda E\psi(W_B) - \eta EW_B) - (\lambda E\psi(W) - \eta EW).
\end{aligned}$$

Therefore the same ideas as in Sections 2.3, 2.4 and 2.5 to implement NPEB can be used to solve the problem for $\boldsymbol{w}_\lambda^{(n)}(\eta) := \arg\min_{\boldsymbol{w}:\boldsymbol{w}^T\mathbf{1}=1} E\{E[(\psi(W) - \eta W)|\mathcal{F}_n]\}$, where $W = \boldsymbol{w}^T \boldsymbol{r}_{n+1}$ and $\mathcal{F}_n$ is the σ-field generated by $\{(r_{it}, x_{i,t-1}) : i = 1, \ldots, m; t \leqslant n\}$, and then to search for η to minimize (2.11) for given λ, followed by searching λ over a grid to minimize the information ratio (2.10).

We illustrate the method with an empirical study involving the weekly returns of $m = 10$ stocks from January 2010 to December 2013 obtained from Yahoo Finance (Wells Fargo, JP Morgan, Apple, Microsoft, Google, IBM, Walmart, AIG, General Electric). The convex function ψ is the *conditional value at risk* $\mathrm{CVaR}_\alpha(X)$, also called *expected shortfall*, which is defined as $E(X|X > \mathrm{VaR}_\alpha)$ for a short position and can be computed by using the function `hHistoricalVaRES` in the `MATLAB` Risk Management toolbox, with X replaced by $-X$ for a long position; see [32, Section 12.1.2 and 12.1.3] on regulatory capital requirements based on VaR_α and CVaR_α and their respective definitions. We use sliding windows of $n = 120$ weeks of training data to construct portfolios for the subsequent week. Performance of a portfolio is measured by its excess returns $e_t = r_t - u_t$ over the benchmark portfolio S&P500 Index u_t. As t varies over the weekly test periods from January 2010 to December 2013, we add up the realized excess returns to give the cumulative realized excess return $\sum_{l=1}^{t} e_t$ up to time t. For the NPEB portfolios we follow the model in Section 2.4 and use GARCH(1, 1) for ϵ_{it} in (2.9). We choose λ which can maximize the information ratio over

the grid $\lambda = \{2^i, i = -3, -2, \cdots, 6\}$. Figure 2 plots the time series of cumulative realized excess returns over the S&P500 Index during the test period of 200 weeks. Also given in Figure 2 for comparison is the corresponding time series for the plug-in portfolio.

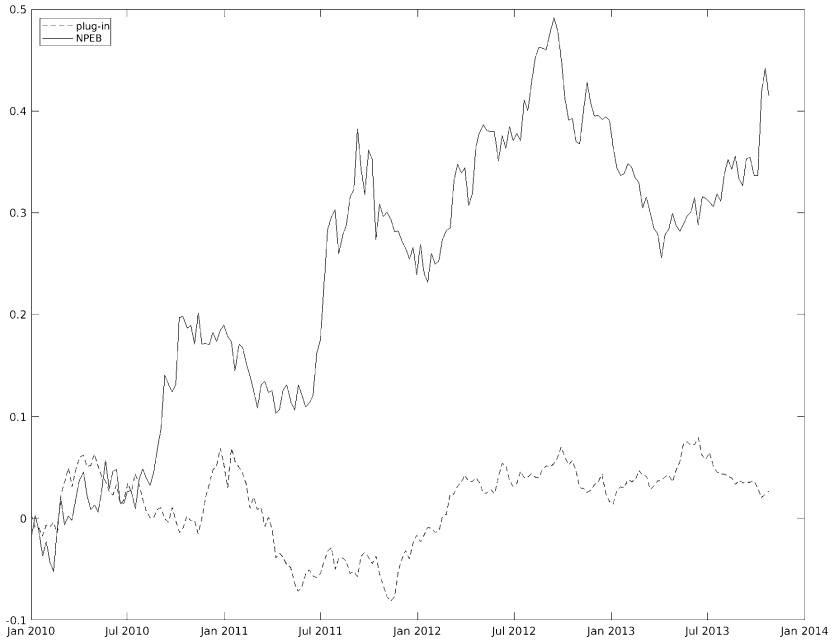

Figure 2: Cumulative excess returns of NPEB portfolio (solid curve) and plug-in portfolio (dashed curve) over the S&P500 Index.

3. Secure hash algorithms and stochastic models, collision and verification problems in blockchains

A *hash function* takes a string x is a string of any length and returns a bit string $H(x)$ with 0 and 1 as its entries such that the computation of $H(x)$ is fast and easy, roughly linear time, but inversion of $H(x)$ is difficult, hence it is a *trapdoor function*. It is called cryptographic if it is (i) *collision-resistant*, (ii) *hiding* and (iii) *puzzle-friendly*. Bitcoin uses one such function called SHA-256, in which SHA stands for *Secure Hash Algorithm*. A collision of H occurs at $x \neq y$ if $H(x) = H(y)$, hence H being collision-resistant means that it is computationally infeasible to find x and y such that collision occurs. Since the range of H is a finite proper subset of its domain of infinite size,

its collision is an absolute certainty. However, for a bit string of length 256 as in SHA-256, and for any given x, locating y by random sampling over the range of H that collides with x requires $2^{256}+1$ inversions of H in the worst case and 2^{128} inversions on average. Even if a computer could invert H $10,000$ times per second, computing 2^{128} inversions of H requires 10^{27} years. On the other hand, if $H(x)$ is regular (say, of the form of x mod 2^{256}), one may find collisions quite easily by detecting patterns. Although no function has been proven to be collision-resistant, the commonly used hash functions in cryptography are believed to be collision-resistant because no collisions have been found (or reported) so far. Cryptologists customarily perform the collision analysis of a hash function via the birthday problem in combinatorial probability, which will be described and further developed in Section 3.1.

To ensure a hash function H to be able to keep the input secrecy, $H(x)$ needs to be "hiding" in the sense that for a 256-bit long y, it is computationally infeasible to find x such that $H(r||x)=y$, where r is sampled from a maximum-entropy distribution [10] and $||$ is the string concatenation operator. Puzzle-friendliness refers to the constraint that given a 256-bit long hashed value $y=H(r||x)$ and its secret key r, the number of operations needed to determine x cannot be significantly less than 2^{256}. Hence, it is computationally infeasible to invert a puzzle-friendly hash function significantly faster than searching over its range. Although its input is restricted to be 768-bit long, SHA-256 can still be used repeatedly to fabricate a cryptographic hash function. Dividing a long x into a sequence of blocks such that each block is of length 512-bit, it concatenates the first block of x to an initialization vector of length 256-bit, which is called the *genesis block*, and is passed to SHA-256 whose output is again of length 256-bit and is concatenated with the second block of x as the input of the second call of SHA-256. Such process repeats until the last block of x is reached. Assuming SHA-256 is collision-resistant, the resulting composite function can be proved to be also collision-resistant. In fact, SHA-256 is known as the *compression function* while the terminology of the cryptographic hash function H is reserved for the composite function; the method of computing cryptographic hash function by applying the compression function (that requires fixed-length input) repeatedly to a sequence of blocks is called the Merkle–Damgård transform. Hence a blockchain is simply a linked list of consecutive data, each of which is hashed by H. Because of the hiding property of H, all data recorded in the chain are kept confidential from those who do not know the corresponding secret key. Yet the collision-resistant property allows every

user of the chain, irrespective of whether the user knows the key, to detect if any data have been tampered.

In his overview of the development of cryptographic systems and eventually bitcoin and other cryptocurrencies, Clark [9] says: "To create a free-floating digital currency that is likely to acquire real value, you need to have something that is scarce by design. In fact, scarcity is also the reason gold or diamonds have been used as a backing for money. In the digital realm, one way to achieve scarcity is to design the system so that minting money requires solving a (difficult) computational puzzle." He says that bitcoin mining integrates this computational puzzle idea with the use of blockchain as a ledger in which all transactions are securely recorded and which "doesn't require trusted timestamping and merely tries to preserve the relative order of blocks and transactions." He also explains why Satoshi Nakamoto [40] used this pseudonym to maintain anonymity in publishing this seminal white paper in the domain that was registered 2 months earlier, concluding that "Bitcoin was able to build up a (vibrant supporting) community of passionate users as well as developers willing to contribute to open-source technology, (unlike) previous attempts at digital cash, which were typically developed by a company." The "peer-to-peer" in the title of this white paper for the proposed electronic cash system refers to that the responsibility of ensuring no double spending is carried by all users rather than by a centralized party (e.g., bank). Bitcoin users, known as *miners*, conduct transactions by using Public Key Encryption to broadcast each encrypted transaction over the network of miners who verify the transactions coded in the blockchain data structure as follows [41, Section 5.1]. Taking as input the block solution s' at the head of the current version of the blockchain and denoting concatenation of strings by $+$, solve for s is the hash $s = h(s' + x + n)$ such that s has at least a specified number ($\sim$64) of leading zeros, where h denotes the SHA-256 hash function, x is the string that is intended to be incorporated into the next block, and n is "nonce" which is a random value that can be updated to make the valid block below a "difficulty target". Note that s' is itself an output of the hash function, hence the term "double SHA-256", and that x contains information of new "transactions on the network (which the miner has to) validate by checking that (digital) signatures are correct and that the outputs being spent haven't already been spent."

Nakamoto [40] points out the importance of (i) incentives for the nodes (miners) of the bitcoin network "to stay honest" and (ii) "vanishingly small" probability that an attacker can use CPU power to "catch up from a deficit" as the number of blocks the attacker has to catch up increases. He uses analogy with the Gambler's Ruin problem (see Feller [17, p. 347]) to show

that the probability π_z the attacker will ever catch up from z blocks behind is $(\tilde{p}/p)^z$, where p (respectively, $\tilde{p}$) is the probability that an honest node (respectively, attacker) finds the next block. It is assumed that $\tilde{p} < p$, hence π_z decreases exponentially with z. If $\tilde{p} \geq p$, then $\pi_z = 1$. Rosenfeld [43] uses a Markov chain model to "clarify and expand on this work", noting that "bitcoin transactions are grouped into blocks, with every block referencing an earlier block by including the uniquely identifying hash of this earlier block in its header" except for the genesis block and that "the blocks form a tree, with the genesis block as the root and each block being a child of the block it references, (and with) a branch being a path from the leaf block to the genesis block (and) representing one version of the history of two conflicting transactions." Göbel et al. [21] further incorporate the difference in communication delay between a "pool of bitcoin miners" and "the rest of the community" in their Markov chain model assuming that the pool mines honestly and in another Markov chain that assumes the pool games the network by using Eyal and Sirer's selfish mining strategy [16].

A classical problem in combinatorial probability is the "birthday problem" on the probability p_n that at least two of n randomly chosen people have the same birthday, assuming 365 (or 366, in a leap year) possible birthdays. For $n \leqslant m = 365$, $p_n = 1 - \prod_{i=2}^{n-1}(1 - \frac{i}{m})$ under the assumption that (A) each individual is born equally likely over every single day of the year and (B) all births are independent of each other. A well-known cryptographic application of the probability p_n, with $m = n^{256} \approx 10^{77}$ for SHA-256 used by the bitcoin blockchain, is the probability of finding a collision for the "secure" hash function. For such large m and sufficiently large n, the Poisson approximation [11, Theorem 1] yields $p_n \approx 1 - \exp[-n(n-1)/2m]$. Recent collision analysis questions the validity of the assumption (A). In particular, Boneh and Shoup [8, Corollary B.2] have shown that the Poisson approximation $1 - \exp[-n(n-1)/2m]$ is only a lower bound of p_n when assumption (A) does not hold.

In Section 3.1 we address this issue by using a new empirical Bayes (EB) model for birthday probabilities. Section 3.2 describes recent breakthroughs in the problem of "zero-knowledge proofs for fast verification" mentioned in the penultimate paragraph of Section 1.2.

3.1. EB model for birthday probabilities in collision analysis

For the birthday problem with $m = 365$ and its generalization to general m under the equivalence between the birthday probability p_n and the probability of at least a match when n balls are dropped randomly into m

boxes, Diaconis and Holmes [12] introduce a general Dirichlet prior distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ for the probability vector $(\pi_1, \ldots, \pi_m)$, when π_i is the probability of a ball dropping into box i. The Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ has probability density function $f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \big(\prod_{i=1}^m x_i^{\alpha_i - 1}\big)/B(\boldsymbol{\alpha})$ for $\boldsymbol{x} = (x_1, \ldots, x_m)$ with $x_i \geqslant 0$ such that $\sum_{i=1}^m x_i = 1$ (i.e., $\boldsymbol{x}$ belongs to the m-dimensional simplex), where $B(\boldsymbol{\alpha}) = \big(\prod_{i=1}^m \Gamma(\alpha_i)\big)/\Gamma\big(\sum_{i=1}^m \alpha_i\big)$ is the multivariate beta function expressed in terms of the gamma function $\Gamma(\cdot)$. Section 2.2 of [12] shows that the Dirichlet prior distribution includes the uniform prior distribution on the m-dimensional simplex, which can be analyzed by Polya's urn scheme and corresponds to the symmetric Dirichlet distribution with $\alpha_1 = \cdots = \alpha_m = c$ and for which Propositions 2.2 and 2.3 of [12] give the Poisson approximation for the probability of at least one match:

$$p_n = 1 - \prod_{i=1}^{n-1} \frac{(m-i)c}{mc+i} \approx 1 - e^{-\lambda} \tag{3.1}$$

as $n^2/m \to \lambda$. For general $\boldsymbol{\alpha}$, (3.1) still holds if

$$\binom{n}{2} \sum_{i=1}^m \alpha_i(\alpha_i + 1) \Big/ \left\{ \left(\sum_{i=1}^m \alpha_i \right) \left(1 + \sum_{i=1}^m \alpha_i \right) \right\} \to \lambda; \tag{3.2}$$

see Remark 1 on Proposition 2.4 (which is about the number of boxes with two or more balls) of [12] that uses this proposition to solve the Bayesian extension of the classical coupon collector's problem, one of "the three principal examples of Feller's Volume I".

We now address the issue with traditional collision analysis of hash functions in blockchains via the birthday problem, mentioned in the first paragraph of this subsection, by using am empirical Bayes approach to the specification of $\boldsymbol{\alpha}$ in the prior distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ of the parameter vector $(\pi_1, \ldots, \pi_m)$. Here m is an astronomical number, e.g., $m \approx 10^{77}$ for SHA-256, and n is also large but manageably smaller than m. Similar to the NPEB approach to portfolio optimization in Section 2.3, we choose hyperparameter $c\sqrt{m}$ for $(\alpha_1, \ldots, \alpha_n)$ and δ for $(\alpha_{n+1}, \ldots, \alpha_m)$ so that $(\pi_1, \ldots, \pi_n)$ yields the estimated probability of collision:

$$\hat{p}_k^{(m)} \approx 1 - \exp\left\{ -\left(\frac{1+\delta^{-1}}{2} \right) \frac{k(k-1)}{m-n} - \frac{k}{\sqrt{m}} \frac{nc}{\delta} \right\}, k = n+1, \ldots, m, \tag{3.3}$$

given that no collision is observed up to time n. To derive (3.3), let $\gamma = c\sqrt{m}$ and consider the $\mathrm{Dir}(\boldsymbol{\alpha})$ prior with $\alpha_1 = \cdots = \alpha_n = \gamma, \alpha_{n+1} = \cdots = \alpha_m = \delta$,

hence the prior distribution of $(\sum_{i=1}^{n}\pi_i, \pi_{n+1}, \cdots, \pi_m)$ is $\mathrm{Dir}(n\gamma, \delta, \cdots, \delta)$. Let $\eta = \sum_{i=1}^{n}\pi_i$. An argument similar to that in the proof of Propositions 2.2 and 2.3 in [12] using Polya's urn shows that analogous to (3.1),

$$P(\text{no collision at times } n{+}1, \ldots, n{+}k|\eta) = (1{-}\eta)^k \prod_{i=1}^{k-1} \frac{\delta(m-n-i)}{(m-n)\delta + i}. \tag{3.4}$$

Moreover, since $\eta \sim \mathrm{Beta}(n\gamma, (m-n)\delta)$, $E\{(1-\eta)^k\}$ is equal to

$$\begin{aligned} &\int_0^1 (1-x)^k \frac{\Gamma(n\gamma + (m-n)\delta)}{\Gamma(n\gamma)\Gamma((m-n)\delta)} x^{\eta\gamma-1}(1-x)^{(m-n)\delta-1}dx \\ &\quad = \prod_{i=1}^{k-1} \frac{(m-n)\delta + i}{(m-n)\delta + n\gamma + i}. \end{aligned} \tag{3.5}$$

Combining (3.4) and (3.5) yields

$$P(\text{no collision at times } n+1, \ldots, n+k) = \prod_{i=1}^{k-1} \frac{\delta(m-n+1)}{(m-n)\delta + n\gamma + i},$$

from which an argument similar to the proof of Propositions 2.1 and 2.4 shows that for $\gamma = c\sqrt{m}$ and $k/\sqrt{m}$ converging to a positive limit,

$$\begin{aligned} &P(\text{collision occurs at some time } i \in \{n+1, \ldots, n+k\}) \\ &\quad = 1 - \exp\left\{ -\left(\frac{\delta+1}{2}\right)\binom{k}{2} \Big/ (m-n) - \frac{k}{\sqrt{m}}\frac{cn}{\delta} \right\}, \end{aligned} \tag{3.6}$$

leading to the NPEB estimate (3.3) when there is no collision up to time n.

Figure 3 displays the collision probability $\hat{p}_k^{(m)}$ given by (3.3) with $\delta = 1, n = 2^{15} = 32768$, and k in the range 2^{110} to 2^{130}, applied to the hash function SHA-256 wih $m = 2^{256} \approx 10^{77}$. This is the dashed curve in Figure 3, showing $\hat{p}_k^{(m)} \approx 0.5$ for $k = 2^{121.5} \approx 10^{36}$. In comparison, the classcial Poisson approximation $p_{n+k} \approx 1 - \exp\big(-\binom{n+k}{2}/m\big)$, represented by the solid curve in Figure 3, increases sharply from 0.0019 to 0.999 as k increases from 2^{124} to 2^{130} and reaches 0.5 for $k = 2^{128.25}$.

3.2. Zero-knowledge proofs for fast verification

As emerging trend in FinTech is digital cash for payment of goods and services. Just like the cash in the real economy, one needs to prove that

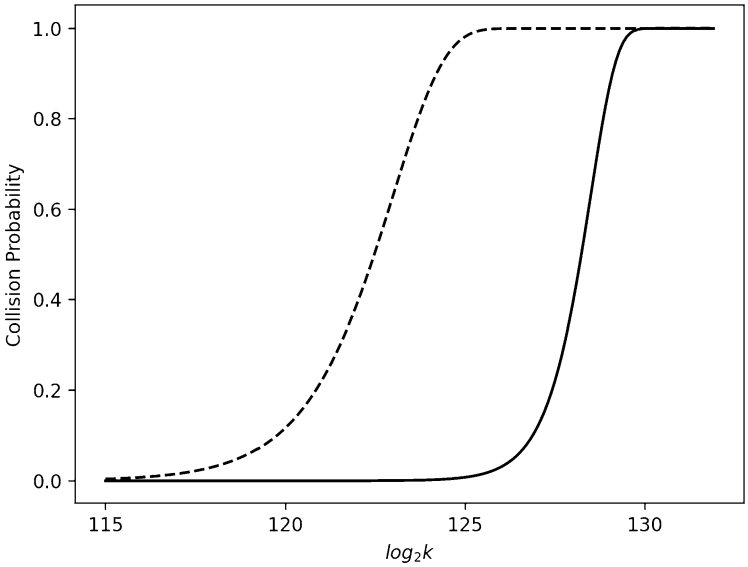

Figure 3: Collision probabilities for SHA-256 using empirical Bayes estimate (dashed curve) and classical Poisson approximation (solid curve).

there is a high probability that one can prevent double spending attacks without granting everyone the right to track his or her cash flow, and that he or she has enough cash on hand to spend. Furthermore, relying on zero-knowledge proofs, there have been several implementations that can protect the privacy of all information about transactions, including payment targets, amounts, and times. In recent years, with the development of regulatory technology, auditing of funds has gradually gained attention and traction. To be precise, one can design a zero-knowledge statement which is capable of supporting selective disclosure: one can provide some specific keys to a third party so that he or she can track whether the specific cash flow he or she provided is correct without revealing unrelated information. Such a technique is universal, for example, one can use similar techniques to prove that funds at a specified address exist.

The invention of Bitcoin ushers in the era of large-scale decentralized cryptocurrency and of blockchains. By using a blockchain network, anonymous participants can reach consensus without trusting each other. This is accomplished by the careful design of consensus algorithms such as Proof-of-Work. Although a proper balance has been struck between Bitcoin's transparent disclosure of addresses and protection of privacy, the state transition

script invented by Bitcoin is still very primitive. It can only verify digital signatures and perform money transfers. On the other hand, a more sophisticated state transition script called "smart contract" was recently introduced to enable a blockchain to handle more complex programming logic. This more general-purpose programmability has transformed the landscape of blockchains. Blockchains today typically include smart contracts in their design, which some researchers name Blockchain 2.0. However, unlike traditional client-server architecture, there is no private storage in these blockchains. This shortcoming is the direct consequence of their design. The verification mechanism is performed by each participant, hence all participants need to have full access to all the information used by the state transitions, making it impossible to verify a secret without knowing it. Thanks to the recent breakthroughs due to Eli Ben-Sasson and his collaborators (see [3], [4] and [5]), we can addresses this shortcoming by zero-knowledge proof in Blockchain 3.0 that is described below.

Digital signatures are important in modern cryptography, especially in finance. Specifically, digital signature is a building block in cryptography in order to verify a message to the public that a particular person, who holds the corresponding private key, has signed to the message without revealing the private key itself. Zero-knowledge proof technology further extends digital signature to prove a wide range of facts without revealing specific information. For all calculations performed by someone who is not completely trusted by us, such as banking, election voting or cloud services, we often need to personally verify that the results are correct. If the calculation process is clear, verification of it is usually simpler and faster than calculation, because the calculator can collect enough evidence to prove that each step of the calculation process is correct. In this scenario, we call the calculator here a "prover". To prove "statement", the simplest verification method is to recalculate, especially when it is a generalized statement, just like what the blockchain's decentralized ledger does, but this suffers from the dilemma between efficiency and privacy. Considering a public function F as a statement, the zero-knowledge prover with data $\mathbf{x}$, which consist of public data $\mathbf{x}_{\mathtt{pub}}$ and private data $\mathbf{x}_{\mathtt{priv}}$, wants to produce a specific result $\mathbf{y}$, keeping the confidentiality of $\mathbf{x}_{\mathtt{priv}}$ and using a verification function V (which returns a bit with 1 denoting "accept" and 0 "reject") to match F, so that one can generate a valid $\mathbf{w}$ such that $V(\mathbf{x}_{\mathtt{pub}}, \mathbf{y}, \mathbf{w}) = 1$ if and only if one has $\mathbf{x}_{\mathtt{priv}}$ such that $F(\mathbf{x}_{\mathtt{pub}}, \mathbf{x}_{\mathtt{priv}}) = \mathbf{y}$. Using the aforementioned breakthroughs in [3], [4] and [5], it is now possible to carry out this verification within $O(1)$ time (instead of polynomial time), about several hundred bytes, comparable

to the most common ECDSA (Elliptic Curve Digital Signature Algorithm) with 64 bytes, regardless of how complex F is.

Non-interactive zero-knowledge proofs are zero-knowledge proofs in which no interaction is needed between the prover and verifier, in contrast to interactive proofs that require multiple communications between the two parties and are therefore inefficient. There have been major developments in the past few years, such as scalable systems that compile a subset of widely used programming languages (C, Java, Python, Javascript, etc.) into a form suitable for proofs and generating the keys, and multiple optimizations to produce a faster process. The verification function can provide a mathematical description model for the target function. In particular, the protocol zk-SNARK [6] (which stands for zero-knowledge Succinct Non-interactive ARguments of Knowledge) now has a QAP (Quadratic Arithmetic Program) for its verification function that has the capability of verifying all NP languages; a language L is said to be NP ifit is possible to check in polynomial time whether an element belongs to L or not. We can describe the QAP computation problem by using a rank-1 constraint system (R1CS) $S := \left((\mathbf{v_i}, \mathbf{w_i}, \mathbf{y_i})_{i=1}^{d}, m\right)$ consisting of d constraints, where each $\mathbf{v_i}, \mathbf{w_i}$ and $\mathbf{y_i}$ is a vector of length m and with entries belonging to a finite field $\mathbb{F}_p$ of order p, in which p is a sufficiently large prime number so that the discrete logarithm problem associated with the group of integers over $\mathbb{F}_p$ is difficult enough to ensure security. It has been shown by Koblitz [29] and others that similar security can be achieved by using a finite group of smaller order from ECDLP (Elliptic Curve Discrete Logarithm Problem) which is described below, e.g., the security of $\mathbb{F}_p$ (with p of 3072 bits) is almost equivalent to the security of 256 bits for the abelian group of elliptic curves.

The cryptographic basis of ECDLP is the computational intractability of finding the discrete logarithm of a random elliptic curve element with respect to a publicly known base; powers b^k can be defined for integers k, with $b^{-k} = (b^{-1})^k$, and elements b of a group for which the discrete logarithm of $x \in G$, with respect to base b, is an integer k such that $x = b^k$. An elliptic curve is the set of plane curves over a finite field satisfying the equation $y^3 = x^2 + ax + b$; a plane curve is a curve in the (x, y) plane which may be an affine or projective plane. This set, together with an ideal point ∞ and the group structure inherited from the divisor group of the underlying algebraic variety, is an abelian group.

References

[1] Ang, A., Chen, J., Xing, Y. (2006). Downside risk. *The Review of Financial Studies*, 19, 1191–1239.

[2] Barzilay, O. (2017). Why blockchain is the future of the sharing economy. https://www.forbes.com/sites/omribarzilay/2017/08/14/why-blockchain-is-the-future-of-the-sharing-economy/.

[3] Bén-Sasson, E., Chiesa, A., Genkin, D., Tromer, E., Virza, M. (2013). Snarks for c: Verifying program executions succinctly and in zero knowledge. *Annual Cryptology Conference*, 90–108. MR3126471

[4] Ben-Sasson, E., Chiesa, A., Tromer, E., Virza, M. (2017). Scalable zero knowledge via cycles of elliptic curves. *Algorithmica*, 79(4), 1102–1160. MR3707355

[5] Ben-Sasson, E., Chiesa, A., Riabzev, M., Spooner, N., Virza, M., Ward, N.P. (2019). Aurora: Transparent succinct arguments for R1CS. *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 103–128. MR3964537

[6] Bitansky, N., Canetti, R., Chiesa, A., Tromer, E. (2012). From extractable collision resistance to succinct non-interactive arguments of knowledge, and back again. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 326–349. MR3388399

[7] Boneh, D. (2019). Still in early days, Blockchain is rich with possibility. https://engineering.stanford.edu/magazine/article/dan-boneh-still-early-days-blockchain-rich-possibility.

[8] Boneh, D., Shoup, V. (2017). *A Graduate Course in Applied Cryptography*. https://toc.cryptobook.us.

[9] Clark, J. (2016). *The long road to Bitcoin. Foreword in Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction.* Princeton University Press.

[10] Cover, T.M., Thomas, J.A. (2006). *Elements of Information Theory, 2nd edition.* Wiley. MR2239987

[11] DasGupta, A. (2005). The matching, birthday and the strong birthday problem: a contemporary review. *Journal of Statistical Planning and Inference*, 130, 377–389. MR2128015

[12] Diaconis, P., Holmes, S. (2002). A Bayesian peek at Feller Volume I. *Sankhýa, Ser. A, Special Issue in Memory of D. Basu*, 64, 820–841. MR1981513

[13] Diffie, W., Hellman, M.E. (1976). New directions in cryptography. *IEEE Trans. Information Theory*, 22, 644–654. MR0437208

[14] Dittmar R.F. (2002). Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *The Journal of Finance*, 57(1), 369–403.

[15] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007. MR0666121

[16] Eyal, I., Sirer, E.G. (2013). Majority is not enough: Bitcoin mining is vulnerable. *Financial Cryptography and Data Security*, 436–454.

[17] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications. Vol. I, 3rd edition.* Wiley. MR0228020

[18] Frankfurter, G.M., Phillips, H.E, Seagle, J.P. (1976). Performance of the Sharpe portfolio selection model: A comparison. *J. Financial and Quantitative Analysis*, 11(02), 195–204.

[19] Friedman, M., Schwartz, A. (1963). *A Monetary History of the United States 1867–1960.* Princeton University Press. MR3887658

[20] Friedman, M. (1992). *Money Mischief: Episodes in Monetary History.* Harcourt.

[21] Göbel, J., Keeler, H.P., Krzesinski, A.E., Taylor, P.G. (2016). Bitcoin blockchain dynamics: The selfish-mine strategy in the presence of propagation delay. *Performance Evaluation*, 104, 23–41.

[22] Goldwasser, S., Micali, S. (1984). Probabilistic encryption. *J. Computer and System Sciences*, 28(2), 270–299. MR0760548

[23] Guo, X., Lai, T.L., Shek, H., Wong, S.P. (2017). *Quantitative Trading: Algorithms, Analytics, Data, Models, Optimization.* Chapman & Hall/CRC Press. MR1988884

[24] Harvey, C. R., Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, 55(3), 1263–1295.

[25] Hoffstein, J., Pipher, J., Silverman, J.H. (2014). *An Introduction to Mathematical Cryptography, 2nd edition.* Springer. MR3289167

[26] IBM Blockchain (2018). Blockchain for financial services means more trust for all: Bring new transparency, simplicity and efficiency to every financial transaction. https://www.ibm.com/blockchain/industries/financial-services.

[27] Jobson, J.D., Korkie, B. (1980). Estimation for Markowitz efficient portfolios. *J. Amer. Statist. Assoc.*, 75(371), 544–554. MR0590686

[28] Jorion, P. (1986). Bayes-Stein estimation for portfolio analysis. *J. Financial and Quantitative Analysis*, 21(03), 279—292

[29] Koblitz, N. (1987). Elliptic curve cryptosystems. *Mathematics of Computation*, 48, 203–209. MR0866109

[30] Krugman, P. (2018). Bubble, bubble, fraud and trouble. https://www.nytimes.com/2018/01/29/opinion/bitcoin-bubble-fraud.html.

[31] Lai, T.L., Wei, C. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Stat.*, 10(1), 154–166. MR0642726

[32] Lai, T.L., Xing, H. (2008). *Statistical Models and Methods for Financial Markets*. Springer-Verlag. MR2434025

[33] Lai, T.L, Xing, H., Chen, Z. (2011). Mean-variance portfolio optimization when means and covariances are unknown. *Ann. Appl. Statist.*, 5(2A), 798–823. MR2840176

[34] Lee S.W., Hansen, B. (1994). Asymptotic theory for the garch(1,1) quasi-maximum likelihood estimator. *Econometric Theory*, 10(1), 29–52. MR1279689

[35] Ledoit, O., Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empirical Finance*, 10(5), 603–621.

[36] Malkiel, B.G. (2003). *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*. W.W. Norton.

[37] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77–91. MR0103768

[38] Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*. Wiley. MR0103768

[39] Michaud, R.O. (1989). The Markowitz optimization enigma: is 'optimized' optimal? *Financial Analysts Journal*, 45, 31–42.

[40] Nakamoto, S. (2018). A peer-to-peer electronic cash system. https://nakamotoinstitute.org/bitcoin.

[41] Narayanan, A., Bonneau, J., Felten, E., Miller, A., Goldfeder, S. (2016). *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press.

[42] Roger, A. (2017). What's the future of online marketplaces & blockchain's technology impact? https://www.forbes.com/sites/rogeraitken/2017/10/24/whats-the-future-of-online-marketplaces-blockchains-technology-impact/#2b91de26630.

[43] Rosenfeld, M. (2014). Analysis of hashrate-based double-spending. arXiv:1402.2009.

[44] Samuelson, P.A. (1973). Mathematics of speculative price. *SIAM Review*, 15(01), 1–42. MR0323315

[45] Sharpe, W.F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3), 425–442.

[46] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25. MR0640163

[47] You, L., Daigler, R.T. (2010). Using four–moment tail risk to examine financial and commodity instrument diversification. *Financial Review*, 45(4), 1101–1123.

Tze L. Lai
Department of Statistics
Stanford University
CA 94305
USA
E-mail address: lait@stanford.edu

Shih-Wei Liao
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, 10617
Taiwan
E-mail address: liao@csie.ntu.edu.tw

Samuel P. S. Wong
Department of Statistics
The Chinese University of Hong Kong
Hong Kong
and
Department of Statistics and Actuarial Science
The University of Hong Kong
Hong Kong
E-mail address: mail@samuelposhingwong.com

Huanzhong Xu
ICME
Stanford University
CA 94305
USA
E-mail address: xuhuanvc@stanford.edu

Received June 23, 2020