Monopoly without a Monopolist: An Economic Analysis of the Bitcoin Payment System*

Gur Huberman† Jacob D. Leshno‡ Ciamac Moallemi§

January 27, 2021

Abstract

Bitcoin provides its users with transaction-processing services which are similar to those of traditional payment systems. This paper models the novel economic structure implied by Bitcoin's innovative decentralized design, which allows the payment system to be reliably operated by unrelated parties called miners. We find that this decentralized design protects users from monopoly pricing. Competition among service providers within the platform and free entry imply no entity can profitably affect the level of fees paid by users. Instead, a market for transaction-processing determines the fees users pay to gain priority and avoid transaction-processing delays. The paper derives closed-form formulas of the fees and waiting times and studies their properties; compares pricing under the Bitcoin Payment System to that under a traditional payment system operated by a profit-maximizing firm; and suggests protocol design modifications to enhance the platform's efficiency. The appendix describes and explains the main attributes of Bitcoin and the underlying blockchain technology.

*This paper was originally circulated in August 2017 and has also appeared with the title "An Economic Analysis of the Bitcoin Payment System". We are grateful to Eric Budish, Alex Frankel, Campbell Harvey, Refael Hassin, Hanna Halaburda, Tammuz Huberman, Emir Kamenica, Seth Stephens-Davidowitz, Jessica Mantel, Canice Prendergast, Bernard Salanie, Ran Snitkovsky, Aviv Zohar, Adam Szeidl (the editor) and the referees for helpful conversations and suggestions, and seminar participants at the Central Bank of Finland, Columbia, EIEF, MSR-NYC, Northwestern, NY Computational Economics, NYU, NYU-IO day, Tel Aviv University, Central Bank of Italy, LUISS, University of Turin, Bocconi, the Paul Woolley Conference, the CEPR conference on Money in the Digital Age, and Stanford for helpful comments. The authors advise FinTech companies. This work is supported by the Robert H. Topel Faculty Research Fund at the University of Chicago Booth School of Business.

†Columbia Business School

‡University of Chicago Booth School of Business

§Columbia Business School

1 Introduction

The 2018 revenue of the global payment industry was $1.9 trillion (McKinsey & Company 2019). The recipients of this revenue – payment-processing firms – enjoy network effects and economies of scale, and therefore limited competition and barriers to entry (Rosenbaum et al. 2017, Morningstar 2019). Multiple lawsuits against payment-processing firms accuse them of abusing their market power and harming welfare.[1] Moreover, regulators worldwide impose restrictions on payment-processing firms, in particular, capping the fees charged to users.[2]

The Bitcoin Payment System (BPS), a platform that provides payment services, shows the feasibility of an alternative, decentralized design. It has been operating reliably since its early 2009 inception. It is not controlled by any entity, governed by a computer protocol, and obtains the required computer infrastructure from anonymous, independent profit-maximizing parties called "miners". Anyone with the required computational power and an internet connection can become a miner and compete with other miners to provide transaction-processing services to the platform and collect the associated rewards.

We model this novel economic structure and show that the BPS's decentralized design offers a prototype of a payment system in which users are protected from monopoly harm even if the payment system were a monopoly.[3] Free entry and competition of service providers *within platform* renders the service providers (i.e., the miners) unable to profitably affect the fees users pay. Even a miner who controls a large fraction of the computational power cannot profitably affect fees. Moreover, the fees users pay do not increase if users lose their alternative payment methods.

[1] For example, see concerns discussed by Herkenhoff & Raveendranathan (2020), and Table 5 therein which provides a list of antitrust lawsuits against credit card payment networks and banks. In a congressional testimony, Aaron Klein (2020) argues that payment systems adopt fee structures that disadvantage the poor. See Evans & Schmalensee (2005) for a detailed description of the payment cards industry.

[2] Hayashi & Maniff (2019) provide a long list of regulatory actions limiting credit card fees in countries around the world. Wright (2012) provides support for the concerns of a long list of public authorities and economists that the fee structure in debit and credit cards leads to inefficiency. In fact, according to Visa Inc. Fiscal 2019 Annual Report, "An increasing number of jurisdictions around the world regulate or influence debit and credit interchange reimbursement rates in their regions. For example, the Dodd-Frank Wall Street Reform and Consumer Act (Dodd-Frank Act) in the U.S. limits interchange reimbursement rates for certain debit card transactions, the European Union's (EU) IFR limits interchange rates in Europe (as discussed below) and the Reserve Bank of Australia and the Central Bank of Brazil regulate average permissible levels of interchange."

[3] The attribution of monopoly power to the BPS is a thought experiment, not an empirical claim.

Standard economic arguments suggest that weak competition among monopolistic firms calls for regulation to mitigate monopoly harm. Under the BPS, users are protected from abuses of monopoly power even without competition from other payment systems. Thus, the BPS addresses potential antitrust concerns in a novel, even revolutionary, way.

In the absence of a price-setting firm, the BPS relies on a market mechanism encoded in its protocol to determine prices and infrastructure. Our analysis of the protocol reveals inefficiencies in this market. Among them is the lack of a mechanism that drives the level of resources acquired and deployed to an efficient level, however defined. We provide design suggestions to address these concerns.

The model elaborates on the observation that the blockchain design makes the BPS a two-sided platform whose constituencies are: (i) miners who collectively provide the system's infrastructure in return for payment; (ii) users who make transactions and pay fees. A brief description of the system is in order to explain the particular properties of this two-sided market that are the focus of our model. For concreteness, we focus on the BPS, whose basic design features are shared among most other cryptocurrencies. Appendix A provides a more detailed description of the BPS which is targeted for economists.

Users post transactions over time; miners organize them into blocks, each block with the same, limited capacity; the block of a single randomly selected miner is added to the blockchain; this block selection amounts to processing of the transactions in that block; the timing of miner selection is a Poisson process with a fixed rate which is independent of the aggregate computing resources used by the miners.[4] That, and the fixed capacity of the blocks imply that the BPS has a fixed expected transaction-processing capacity.

The system's limited capacity coupled with the randomness of transaction arrival and processing times imply that, at times, transactions will be processed with delays of random lengths. To make the presentation cleaner, we assume, that on average, the system has sufficient capacity to process all transactions. In addition, the analysis assumes that the mining resources are sufficient to guarantee the system's reliability and security. When so, increases in the mining resources do not affect the system's transaction-processing capacity.

All miners perform the same tasks. Participation in the miner selection tourna-

[4]This is a simplification, see Appendix A for a precise description.

ment is the most resource-consuming among these. A miner's chance of being selected is proportional to his share of the total computational resources. The selected miner is said to have mined a block, and is rewarded with a fixed, system-generated reward plus the fees associated with the transactions in that block. Each user chooses the fee associated with his transaction. Each miner is free to enter and exit the system at no cost. Each participating miner chooses which transactions to include in his block.

We set up a model of fees, priority levels, and mining intensity that captures the main features of the BPS. Its analysis highlights differences between the BPS and a traditional payment system operated by a profit-maximizing firm. The analysis delivers explicit formulas of the fees and delays, thereby enabling suggestions for design improvements. Figure 1 suggests an agreement between the fee formula and the data.

Beyond the quantitative results, the analysis offers a series of qualitative insights as follows.

The BPS processes all transactions, albeit with delay; all users receive strict positive surplus. In contrast, in our setting a profit-maximizing firm excludes low willingness to pay (WTP) transactions but processes the rest without delay. In the BPS, the fee level does not increase if user WTP increases (e.g., if users lose their alternative options) whereas the firm charges more if users' WTP increases.

User payments under the BPS are determined by a congestion market and are payments for service speed. A profit-seeking miner excludes the transactions which offer the lowest fees when the assembled block is full. Therefore, users to whom delays are costly will offer relatively high fees to gain priority and be served faster.

In equilibrium, users with higher delay costs receive higher processing priority and therefore shorter delays. The fee a user pays is equal to the expected delay externality he imposes on others who offer lower fees. Thus, fees are equal to those obtained by allocating priority through a Vickrey–Clarke–Groves (VCG) mechanism, although the BPS employs no auctioneer. User WTP does not affect fees, assuming WTP is sufficiently high. This implies users are protected from price increase should alternative payment service providers leave the market.

An increase (respectively, decrease) in the arrival rate of new transactions results in increased (resp., decreased) congestion, which in turn causes fees to be higher (resp., lower). No delays imply no fees. The analysis offers an explicit relation between block size (which reflects congestion) and the USD-denominated fee. Figure 1 provides a theoretical and an empirical summary of this relation. Notably, the dependence of

fees on congestion is highly non-linear: fees are negligible when blocks are below 50% of their maximal size, positive when blocks are at 80% of their maximal size, and substantially higher when blocks are close to their maximal size.

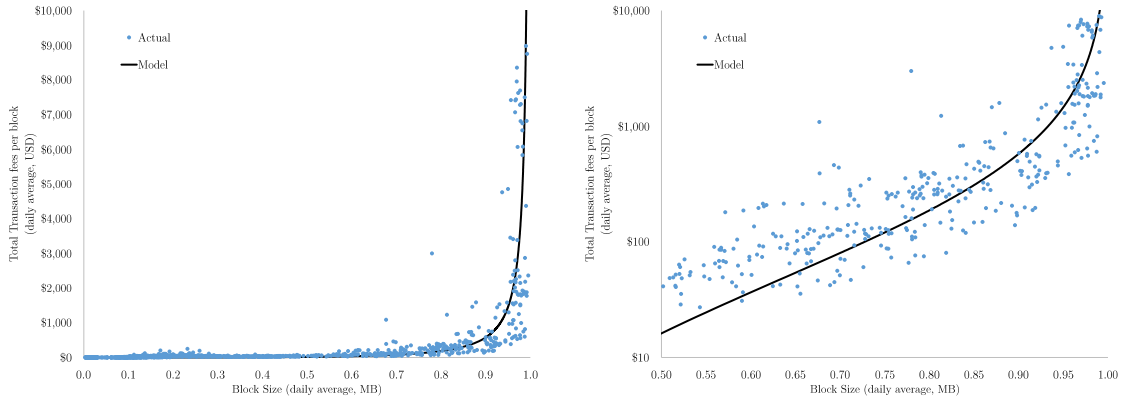

Figure 1: Actual and model predicted transaction fees per block (in USD) and block size for the Bitcoin Payment System (daily averages, April 1, 2011–June 30, 2017). The chart on the left shows fees on a linear scale from the entire range of dates; the chart on the right shows fees on a logarithmic scale over periods when block size was above 0.5MB. See Section 6.2 for details.

We show that even a miner who controls a substantial fraction of the mining resources cannot profitably affect the fees paid by users. While a large miner can affect a user's choice of fees, an increase in user fees will attract entry by new miners, leading to increased competition and lower profits for a miner who attempts to affect user fees. In contrast to standard platform competition, new miners face no barriers to entry as they enter and compete within the same platform. Free entry of miners is essential to this result.

Newly minted coins and transaction fees fund the miners who acquire mining resources in USD-denominated markets. Exchange rate and fee-level fluctuations affect miners' aggregate income, which in turn affects aggregate mining power in the BPS. There is no mechanism that drives the level of infrastructure resources acquired and deployed to an efficient level, however defined.

The analysis points to an efficiency contrast between the BPS and a profit-maximizing firm. Namely, the latter's service is associated with dead-weight loss, whereas the BPS can operate with excess capacity, serving all users and awarding each with strictly positive surplus. If miners are homogeneous, all surplus accrues to the users.

However, the costs of operating the BPS are likely to be higher than those of a

traditional firm: its decentralized architecture requires duplication of computations and expenditure of efforts in the miner selection tournament; the aggregate mining level can be too high; costly delays are necessary to induce users to pay transaction fees. Thus, welfare under the BPS can be higher or lower than that under a traditional system, depending on the value of eliminating monopoly dead-weight loss.

Hundreds of variants of Bitcoin have emerged, with many aiming to improve on the original Nakamoto (2008) design. Our analysis provides the following messages to designers. First, it suggests that congestion is not merely an engineering necessity, but also a device to motivate users to pay transaction fees. Second, the analysis suggests a simple modification that avoids the variation in revenue from transaction fees. In the BPS, capacity is fixed and congestion varies with demand; consequently, the revenue and infrastructure levels vary over time.

We suggest an improved design: a protocol rule that automatically adjusts the system's capacity according to the volume of transactions, thereby steadying congestion, aggregate fees, and mining level. This design has two advantages over alternatives such as a fixed transaction fee: (i) it allows the system to raise revenue without excluding transactions, as users can choose to pay no fees but incur delays; (ii) it allows the protocol to obtain the USD market value of delay reduction without the need to learn the exchange rate. Alternatives such as fixed transaction fees (or newly minted coins) need to be set within the protocol and be denominated in the system's coin, implying revenue fluctuation with the coin's exchange rate.

The analysis also allows us to optimize parameter choices. We offer an analytic expression for the delay costs required to raise a certain revenue level. Analysis and examples suggest that large blocks are less efficient in that they require longer delays to sustain a given level of revenue.

Related Literature

Famously, a white paper by Nakamoto (2008) coined the term Bitcoin and described the BPS. Its opening paragraph criticizes the costs of the existing financial system and its usefulness to small transactions, "Completely non-reversible transactions are not really possible, since financial institutions cannot avoid mediating disputes. The cost of mediation increases transaction costs, limiting the minimum practical transaction size and cutting off the possibility for small casual transactions." Section 6 ("Incentive") predicts that transaction fees will eventually fund the system, "The incentive

can also be funded with transaction fees... Once a predetermined number of coins have entered circulation, the incentive can transition entirely to transaction fees... " The Section's title notwithstanding, Nakamoto (2008) is silent on the incentive to pay transaction fees, their relation to other parameters, and their implications; understanding these is the present paper's task.

Kroll et al. (2013) offer an analysis of the incentives faced by participants in the system, and especially the incentives faced by miners. They conclude a brief discussion of transaction fees by stating, "We therefore do not expect transaction fees to play a significant long-term role in the economics of the Bitcoin system, under the current rules. We believe that a rules change would be necessary before transactions fees can play any major role in the Bitcoin economy." The present paper shows otherwise, i.e., that transaction fees have dual and crucial roles in the Bitcoin system: (i) They are supplanting newly minted coins as the funding source of the mining community; (ii) They are the arbiters of priority in the congestion of messages to be processed by the miners, i.e., they determine priority in the message queue.

Following the initial version of this paper, the design of transaction fee mechanisms has received attention from both academics and practitioners (for example, Buterin (2018)). Easley et al. (2017) is a contemporaneous piece which proposes and empirically examines an equilibrium model of exogenously specified transactions fees and block size assumed restricted to a single transaction. Their model predicts that miners' profits are zero and that fees are positively correlated with transaction waiting times. The data appear consistent with these predictions. Lavi et al. (2017), Yao (2018) and Basu et al. (2019) suggest alternative mechanisms for transaction fees.

Prat & Walter (2018) study the dynamics of miner entry as it is influenced by changes in exchange rates and technological changes and predictions thereof. Felten (2013) argues that in equilibrium miners break even. Cong, He & Li (2018) argue that large mining pools confer risk-sharing advantages on their members, which are mitigated by the larger fees which larger pools charge their members. Arnosti & Weinberg (2018) develop a model where miners are heterogeneous in their cost structure, and quantifies how such asymmetries lead to the formation of oligopolies and concentration of mining power.

Eyal & Sirer (2014), Sapirshtein et al. (2016) analyze the equilibrium between miners and show that proper design of the blockchain protocol produces a reliable system in equilibrium if all miners are sufficiently small. Babaioff et al. (2012) analyze the incentives to propagate information in the BPS. Narayanan et al. (2016) offer an

elaborate description and analysis of the system. Croman et al. (2016) provide cost estimates for the BPS and analyze the potential for transaction-processing capacity. Eyal et al. (2016) suggest an alternative design aimed to construct a system with a higher capacity. Carlsten et al. (2016) analyze how incentives for miners change when miners are rewarded with transaction fees instead of newly created coins. Chiu & Koeppl (2017) evaluate the welfare implications of printing new coins.

The protocol proposed by Nakamoto (2008) posits that in case of a fork, miners will follow the longest branch. Biais et al. (2018) study the robustness of this rule. Budish (2018) studies the system's vulnerability to attacks and argues that the cost of securing Bitcoin is inefficiently high. Abadi & Brunnermeier (2018) posit three desired properties of distributed ledger technologies, (i) correctness, (ii) decentralization, and (iii) cost efficiency and argue that no ledger can satisfy all three properties simultaneously.

Yermack (2015) reviews the history of Bitcoin and its price history to "argue that bitcoin does not behave much like a currency according to the criteria widely used by economists. Instead bitcoin resembles a speculative investment similar to the Internet stocks of the late 1990s."

Gandal & Halaburda (2014) analyze competition between the different cryptocurrencies. Halaburda & Sarvary (2016) review the cryptocurrency market, its development, and future potential of blockchain technology. Gans & Halaburda (2015) analyze the economics of digital currencies, focusing on platform-sponsored credits. Catalini & Gans (2020) discuss possible opportunities that can arise from blockchain technology. Huberman et al. (2019) provides a broader comparison between services provided by the BPS and services provided by a firm.

Recent work considers the valuation of bitcoin relative to fiat currencies and other goods. That work usually assumes away the limited capacity of the BPS, although it induces delays and transaction fees. Ron & Shamir (2013) and Athey et al. (2016) provide analysis of the usage of bitcoin and its value as a currency. Schilling & Uhlig (2018) analyze the evolution of bitcoin prices relative to fiat currency and its implications for monetary policy. Makarov & Schoar (2018) report arbitrage opportunities across cryptocurrency exchanges, primarily across regions.

Cong, Li & Wang (2018) study a dynamic pricing and adoption model in which wider adoption renders the cryptocurrency more valuable. Pagnotta & Buraschi (2018) study bitcoin pricing under the assumption that, at all levels, higher aggregate mining effort delivers higher value to users. Sockin & Xiong (2018) propose a pricing

model for an ICO for a platform on which households can exchange certain goods or services if they own the platform's native coin.

Lui (1985), Glazer & Hassin (1986), Kittsteiner & Moldovanu (2005) and Hassin (1995) study a queuing system in which users with different waiting costs volunteer to pay transaction fees (termed bribes in Lui (1985) to gain priority in a queue to a single service station which serves customers one at a time. The main observation of Lui is that the server may increase its profits by increasing the speed of service. Hassin (1995) shows that the service rate that maximizes the server's profits is always slower than the socially optimal service rate. Hassin & Haviv (2003) provide a summary of the results, and Hassin (2016) provides an updated review. Kittsteiner & Moldovanu (2005) show that convexity or concavity of delay costs determines the queue-discipline.

The present analysis considers a queuing system in which transaction arrival and service arrival is stochastic, but the service is processed in batches of fixed maximal size. The prior work corresponds to a batch size of one. The interaction among arrival rates, service rates, and the maximal batch size, and their impact on the transaction fees and server's revenues are of major concern.

Organization of the Paper

Section 2 provides a model of traditional payment systems, the BPS, and users who may use either. For the sake of completeness, Section 3 provides the standard analysis of a traditional payment systems operated by a firm. Section 4 provides our main analysis and characterizes the equilibrium under the BPS. Section 5 leverages our analysis to provide design suggestions. Section 6 brings empirical evidence to bear on some of the model's predictions. Section 7 provides some final remarks. Appendix A provides a simplified explanation of the BPS and the underlying blockchain technology.

The online appendix contains all omitted proofs and additional discussion. Appendix B extends our analysis of the BPS to parameters where the participation constraint of some users binds. Appendix C extends our analysis to allow for endogenous determination of the user's WTP. Appendix D gives additional properties of transaction fees under the BPS. Additional figures are in Appendix E. Omitted proofs are in Appendix F.

2 Economic Model of Traditional Payment Systems and the BPS

This section sets up a model of a payment system to facilitate a comparison between a decentralized protocol like Bitcoin and a conventional payment system which is controlled by a profit-maximizing firm. Section 2.1 describes the users. Their preferences are the same across the two payment systems. Section 2.2 very briefly states the familiar problem of a firm providing payment services. Section 2.3 describes succinctly the features of the Bitcoin Payment System (BPS) relevant to its economic analysis and its comparison with a traditional system. Sections 4 and 5 offer equilibrium analyses of the firm and of the BPS, respectively.

2.1 Users

Each user has a single potential transaction; hence, references to users and their transactions are interchangeable. Users are heterogeneous in two distinct dimensions. First, users differ in their willingness to pay (WTP) for using the system. The value a user derives from sending a transaction in the system above the value available via an alternative is his WTP $R = v - v_{alt}$. Second, users have different delay costs per unit time c. The net reward of user (R, c) from sending a transaction that is processed after delay W and paying a transaction fee b is

$$u\left(W, b \mid R, c\right) = R - c \cdot W - b. \tag{1}$$

The variables R and b are denominated in USD;[5] the variable c is in USD per unit time. By the definition of R, a potential user will prefer using the system over the alternative (outside option) if $u\left(W, b \mid R, c\right) \geq 0$.

To make the cleanest distinction between the systems, we consider a setting where $R \in \{R_L, R_H\}$ $(R_L \leq R_H)$ and is not correlated with c.[6] One interpretation is that users with WTP R_H have no compelling alternative of making the transfer, and therefore their WTP R_H is almost the entire value of processing the transaction,

[5] In practice, transaction fees in the BPS are denominated in bitcoin. However, since users decide transaction fees as they submit transactions, we will consider them as USD denominated without loss of generality. This is in contrast to the block reward S discussed in Section 2.3, which is fixed by the protocol, and hence is impacted by the USD/bitcoin exchange rate.

[6] An alternative and analogous model entails $u = V\delta^W - b - v_{alt}$. Variation in R is variation in v_{alt}. Variation in c is variation in δ. All have the same V.

while users with WTP R_L can use an alternative method, and therefore their WTP is equal to the cost of the alternative method.

WTP reflects various features of the system. Currently, users of the BPS face costs and risks due to the volatility of the bitcoin to USD exchange rate. Likewise, users may have concerns about the long-run viability of the system, security, privacy, or ease of use (e.g., the lack of password recovery service). On the other hand, the BPS may facilitate transactions that are difficult to conduct through other means. We capture such considerations by the WTP R.

Potential users arrive over time according to a Poisson process. The arrival rate of users with value R_j is λ_j with $j = L, H$ and $\lambda = \lambda_L + \lambda_H$. Both of these populations of users have heterogeneous delay costs per unit time c that are distributed $c \sim F[0, \bar{c}]$, independently of the user's WTP R. The cumulative distribution function $F(\cdot)$ has a density $f(\cdot)$, and its tail probability is denoted $\bar{F}(c) \triangleq 1 - F(c)$.

For tractability, users know the steady-state behavior of the system, but do not observe other pending transactions at the time they submit their transaction. Users are risk neutral and maximize their expected net reward.

We focus our analysis on the case summarized below which gives the cleanest distinction between the BPS and a firm.

Assumption 1. *The following hold:*

- $\lambda_H R_H > (\lambda_L + \lambda_H) R_L$
- $R_H \geq R_L > \bar{R} > 0$ *where* $\bar{R}$ *is defined in Lemma 2.*
- *User delay costs* c *are distributed independently of WTP* R*.*

The assumption that $R > 0$ entails that transaction-processing by the BPS is valuable to its potential users after accounting for exchange-rate risk, the BPS's other limitations, and the possibility of using alternative systems. In particular, users consider the system to be a reliable means of sending transactions.

2.2 Payment System Run by a Firm

A firm-run conventional payment system can process transactions without delay at a marginal cost of c_f per transaction. The firm sets its price in response to the distribution of consumer demand. The firm faces no capacity constraints, can costlessly delay transactions, and can offer different prices for processing transactions with different

delays. In Section 3, we show that the firm does not pursue these policies because they do not increase its profit.

2.3 Decentralized Cryptocurrency

The BPS offers users a similar functionality to that offered by familiar payment systems, i.e., the ability to transfer balances from one user to another. In contrast to traditional payment systems, the BPS uses a decentralized network of computers (so called miners) to process transactions and maintain the ledger containing their history. The novel blockchain design ensures the system as a whole is reliable and trustworthy without the need to trust any individual miners.

A computer protocol governs the system and dictates the rules for how miners and users interact within the system. Thus, the BPS system is a two-sided market with rules that are fixed by a computer protocol. The description in Appendix A provides further details regarding the protocol's operations and functionality. In this Section, we provide the implications of the design for the structure of the two-sided market.

Users send their transactions as they would under any payment system but also select the transaction fee they will pay. Transactions need not be processed in their order of arrival. Processing may take time.

Miners provide their computational infrastructure to the BPS at will and can switch between being active and inactive. Collectively, the miners maintain a ledger of all transaction history. Transactions are periodically added to the ledger in batches, in the form of a block of transaction data. These additions are according to a Poisson process[7] with rate μ, irrespective of the number of miners. For each block, a randomly chosen active miner selects which pending transactions are processed in the block. That miner is said to have mined the block. The probability that a miner is chosen is equal to his share of the total computational power. A block can contain up to K transactions.[8] Pending transactions not included in a block wait to be processed in a future block. Miners observe all pending transactions and the fees associated with them. Each miner applies his own selection of up to K pending transactions. We say that transactions included in the miner's block are processed by that miner.

Miners incur a cost per unit time while they are active. A miner who mines a new block is rewarded with the transaction fees paid by the transactions included in

[7] A Poisson process is the limit of many independent binomial trials. See footnote 30.

[8] While in practice transactions may vary in size, for the sake of tractability we assume all transactions are of the same size.

that block as well as a fixed block reward of newly minted coins. We denote by S the expected number of coins the system awards per unit time.[9] We use e to denote the USD/bitcoin exchange rate, which is assumed fixed and exogenous. Of particular interest will be the case where $S = 0$, which describes the operation of the BPS in the long term.[10]

Each miner chooses the computation power it deploys. We denote the aggregate computational power by N. The total expected processing capacity of the system is an average μK transactions per unit time. The values μ, K are predetermined by the protocol and are unaffected by the number of miners, their total computational power N, or the transaction volume λ.

Realized processing capacity is random because block arrival time is random. The load parameter is $\rho = \lambda/\mu K$, which is the ratio of average demand to capacity. The parameter ρ is a measure of the system's congestion. To make the presentation cleaner, we assume, that on average, the system has sufficient capacity to process all transactions.

Assumption 2. *The system has sufficient capacity to eventually process all transactions, that is,* $\rho < 1$.

Miners who possess a small fraction of the total computational power N have a small probability of mining a block. We assume that each of these miners cannot influence the system or the choices of other miners and users. We refer to these as small miners. To capture that each small miner has a negligible effect on transaction-processing delays, the model distinguishes between large miners and small miners. Each large miner i can choose to deploy computational power $x_i \geq 0$. When taking actions, each of these large miners take into account the actions' impact on the system, including the way they influence other actors' choices. We assume there are finitely many large miners indexed by i, each with computational power bounded by $\bar{x}_i \in \mathbb{R} \cup \{\infty\}$ and a cost of computational power $c_i : [0, \bar{x}_i) \to \mathbb{R}$ that is smooth, increasing, strictly convex, and satisfies $c_i(0) = 0$, $\lim_{x \to \bar{x}_i} c'(x) = \infty$. There are infinitely many small miners, and each small miner who chooses to be active deploys an identical, infinitesimal amount of computational power at infinitesimal cost $c_m > 0$. If selected to mine a block, the miner's revenue from the block does not depend on the computational power he deploys.

[9] Note that all values are given per unit time.

[10] In the BPS, the block reward is halved every 4 years, until it is rounded down to 0.

A miner's block assembly policy $A \in \mathcal{A}$ captures his transaction selection. Formally, the collection of pending transactions is associated with a list of transaction fees $\boldsymbol{b} = (b_1, \ldots, b_n)$. The block assembly function A assigns for every $\boldsymbol{b}$ a vector $A(\boldsymbol{b})$ of zeros and ones of the same length; transaction j is included in the corresponding entry of $A(\boldsymbol{b})$ is one. Compliance with the protocol requires that the vector $A(\boldsymbol{b})$ has no more than K entries equal to one.

Next, we describe the interactions between users, between miners (large and small), and across these two groups. Users play a congestion queueing game, in which each user chooses b, the fee he offers, to maximize his expected utility (1). A user's delay W depends on the selection of transactions by a randomly chosen miner. That selection is sensitive to the fee offered by the transaction and its level relative to other transaction fees. Each miner is a profit maximizer who chooses whether to be active or not; those who choose to be active choose a block assembly policy. Large miners who choose to be active also choose their computational power. Miners may enter or exit in response to profit opportunities, leading to an increase or a decrease in the total computational power. An increase in the total computational power lowers the probability a given miner is chosen to mine a block, and therefore lowers the miner's payoffs. Large miners' block assembly policies can affect the transaction fees offered by users.

Behaviors in systems like the one we are studying could be time- and state-dependent. We abstract from both. We focus on equilibria such that the system is time invariant and has a steady-state distribution. We assume all participants know the system parameters and steady-state distribution. We imagine the equilibria being on a time horizon where the model parameters (arrival rates, exchange rate, etc.) are fixed. Over a longer time horizon these parameters may be changing, and hence the system may move from one equilibrium to another.

Formally, we study the three-step, extensive-form game which summarizes the interactions among the various actors. These steps are:

(i) Each large miner i chooses whether to be inactive or to be active with some computational power $x_i > 0$ and some block assembly policy $A_i \in \mathcal{A}$.

(ii) Small miners observe the actions taken by the large miners in the first step and choose whether to be inactive, or to be active with infinitesimal computational power and some block assembly policy A. For each $A \in \mathcal{A}$, let $\eta(A)$ be the aggregate computational power of the miners (small and large) who choose a block assembly policy A, i.e., η is the the distribution of block assembly policies. The aggregate of

$\eta(A)$ is N. The probability a block is assembled according to A is $\eta(A)/N$.

(iii) Users play the congestion queueing game implied by η. We restrict attention to deterministic stationary strategies. A user's expected waiting time depends on the fee he offers, the fraction γ of users who participate, and the distribution of transaction fees $G(\cdot)$. Each user type (R, c) chooses whether to opt out and receive a payoff of 0 or participate and send a transaction with fee $b(R, c) \geq 0$ to receive his expected steady-state payoff,

$$R - b(R, c) - c \cdot W(b(R, c) \mid G, \gamma, \eta)$$

where $W(b(R, c) \mid G, \gamma, \eta)$ is the expected waiting time for a transaction with fee $b(R, c)$ under the steady-state distribution of the system.[11]

The payoff of an active large miner i with computational power $x_i > 0$ and block assembly policy A_i is

$$\frac{x_i}{N}\Big(\text{Rev}(A_i \mid G, \gamma, \eta) + e \cdot S\Big) - c_i(x_i) \ ,$$

where $c_i(x_i)$ is large miner i's cost of computational power, and $\text{Rev}(A \mid G, \gamma, \eta) = \mathbb{E}_{\boldsymbol{b}}[\boldsymbol{b} \cdot A(\boldsymbol{b})]$ is the expected transaction fees per block assembled by A under the steady-state distribution of pending transactions $\boldsymbol{b}$.[12] The payoff of an active small miner with block assembly policy A is proportional to

$$\frac{1}{N}\Big(\text{Rev}(A \mid G, \gamma, \eta) + e \cdot S\Big) - c_m \ ,$$

where all small miners have the same cost of an infinitesimal unit of computational power $c_m > 0$. All inactive miners receive a payoff of 0.

Assumption 3. *Given any feasible profile of choices by large miners, in any subgame perfect equilibrium of the induced subgame for small miners and users there are some*

[11] The decisions of all participants specify a continuous time Markov process and its steady-state distribution as follows. The states are lists of transaction fees of pending transactions $\boldsymbol{b} = (b_1, \ldots, b_n) \in \cup_n \mathbb{R}^n \cup \{\phi\}$. There are two kinds of transitions. At Poisson rate $\gamma\lambda$ a user arrives and posts a transaction with transaction fee independently drawn from G, and the state is updated by appending the new transaction. At Poisson rate μ a block is mined, and the block assembly policy $A(\cdot)$ is applied with probability $\eta(A)/N$. The system transitions to a new state by erasing all transactions selected by $A(\cdot)$. For completeness, if given G, γ, η the system does not have a steady-state distribution, we set $W(\cdot \mid G, \gamma, \eta) \equiv \infty$.

[12] The distribution of pending transactions observed by a miner who is selected to mine a block is identical to the steady-state distribution of pending transactions. For completeness, if given G, γ, η the system does not have a steady-state distribution we set $\text{Rev}(A \mid G, \gamma, \eta) \equiv 0$.

small miners that are active and some small miners that are inactive.

Assumption 3 requires the presence of sufficiently many players who can become active small miners. This is likely to be satisfied if it is possible to become a small miner by buying standard computational resources on the open market.[13] Because a miner's payoff decreases with aggregate computational power, this implies that in any equilibrium some potential small miners are inactive. The second part of Assumption 3 requires that some small miners are active given any choices by large miners. This will be satisfied if the total computational resources employed by large miners are limited and the computational resources used by small miners are sufficiently efficient (i.e., c_m sufficiently small).

To highlight the distinctive properties of the system, the analysis focuses on the parameter range where all potential transactions can be processed. The assumptions in Section 2.1 imply that there are sufficiently many miners for the system to operate reliably and securely. In Section 4, we analyze the BPS under these assumptions and verify when they indeed hold.

To avoid technical issues with equilibrium existence, we restrict the set of block assembly policies $\mathcal{A}$. We require that for any profile of large miners' block assembly policies chosen from $\mathcal{A}$, the induced subgame (played by small miners and users) has at least one subgame perfect Nash equilibrium in pure strategies. We restrict attention to deterministic strategies and implicitly assume that small miners can use a public coordination device to coordinate their entry decisions.

Miners procure the resources they need in fiat currency-denominated markets. Therefore, we consider all payments and costs denominated in USD rather than in bitcoin. In particular, the USD value of the block reward fluctuates with the exchange rate. No miner can affect this exchange rate.

3 Analysis of the Firm

The firm's problem is standard and stated here for completeness. We consider the profit-maximizing mechanism, allowing for probabilistic or dynamic mechanisms. By the revelation principle, it is sufficient to consider direct mechanisms in which the firm offers a menu to each user. Since the firm faces no capacity constraints, it

[13] A miner who controls a sufficiently large fraction of the mining resources may behave in a way that disadvantages small miners (e.g. selfish mining (Eyal & Sirer 2014)). Our results hold as long as the miner is unable to prevent small miner entry.

can optimize its menu separately for each user. Therefore, a menu of options that maximizes profit from a single randomly drawn user delivers the firm's optimal profit.

The following proposition shows that, unable to distinguish high and low WTP customers, the firm sets a transaction fee that precludes low WTP customers from using the system and processes all the transactions that pay this fee with no delay. The firm can and does change the price it charges if R_H changes.

Proposition 1. *When $\lambda_H R_H > (\lambda_H + \lambda_L) R_L$, the firm's optimal menu includes a single option: it charges the fee $b = R_H$ and processes all transactions that are willing to pay the fee with no delay. Thus, only high value customers are served. Consumer surplus is 0 and social surplus is $\lambda_H (R_H - c_f)$, all accruing to the firm.*

The intuition for the result is that the firm cannot use delays to screen between high and low WTP customers, and therefore avoids delays that decrease a user's willingness to pay.[14] When $\lambda_H R_H > (\lambda_H + \lambda_L) R_L$, the firm makes higher profits by selling only to high WTP users. The proof is in Appendix F.5.

A few observations facilitate the comparison with the BPS, which is presented in Section 4.3. First, the distribution of the user delay costs F is irrelevant to the equilibrium outcome when the firm is the service provider. Second, pricing out the low WTP customers entails a dead-weight loss of $\lambda_L (R_L - c_f)$. Third, the high WTP customers pay exactly their WTP. They will pay more, e.g., if these customers lose their best outside option.

The firm's profit is likely to draw the attention of potential entrants seeking to establish a competing payment service. Such competing payment services provide agents with alternative options, thereby reducing their WTP R (the value of using the BPS relative to the alternative options) and the price the firm can charge. However, the strong network effects and high setup costs that characterize the payments industry are likely to deter entry.[15] Even if there are multiple payment providers in the market, as long as each serves a separate segment, the service providers enjoy pricing power.[16]

[14] Recall that in our setting were there is no correlation between WTP and delay costs. Proposition 1 may not hold if such correlation exists.

[15] See, for example, Morningstar (2019), Evans & Schmalensee (2005), and references therein.

[16] Edelman & Wright (2015) argue that price coherence of payment cards (i.e., the restriction not to surcharge for payment by card) results in inefficiency that is not mitigated by competition. Another illustration of the ability to exercise market power is Apple's ability to collect 15 basis points from each transaction using the iPhone's NFC capabilities (https://www.cardfellow.com/blog/introduction-to-apple-pay/, retrieved Jan 2020). See also https://qz.com/1726203/apple-is-suffocating-mobile-payment-rivals/.

4 Analysis of the BPS

We analyze the equilibrium of the system under the assumptions stated earlier. Subsection 4.1 analyzes the behavior of miners in steps (i) and (ii). Subsection 4.2 analyzes the behavior of users in step (iii). Subsection 4.3 completes the analysis, giving expressions for the system's infrastructure level and welfare.

4.1 Miners, Small and Large

A small miner's choice of block assembly policy does not affect the distribution of block assembly policies η and cannot affect users' fee choices. It follows that small miners maximize their payoffs by selecting the block assembly policy A^* that maximizes $A^*(\boldsymbol{b}) \cdot \boldsymbol{b}$ for any $\boldsymbol{b}$. In words, $A^*(\cdot)$ selects the K pending transactions offering the highest fees. (If there are fewer than K pending transactions, $A^*(\cdot)$ selects all of them.)

A large miner's choice of block assembly policy affects the distribution of block assembly policies η, changing the induced subgame for small miners and users. It may seem as though large miners can attempt to increase their payoffs by choosing a block assembly policy different from A^* to favorably affect users' fee choices. However, Theorem 1 shows that such attempts will not increase the miners payoffs; entry by small miners renders the block assembly policy A^* optimal for any large miners.

Theorem 1 considers the choices of large miners' behavior in step (i), fixing possible responses of small miners and users. That is, for any profile of choices of large miners, we select an equilibrium play of small miners and users in the resulting subgame. Each selection generates an induced game between large miners. We use x_i^* to denote the unique solution to $c_i'(x_i^*) = c_m$ or $x_i^* = 0$ if no solution exists.

Theorem 1. *In any induced game between large miners, it is a dominant strategy for each large miner i to choose the block assembly policy A^* and the computational power x_i^*. Moreover, for any choice of computational power x_i, we have that A^*, x_i dominates A, x_i for any block assembly policy A.*

In the equilibrium in which all miners choose A^*, the total amount of computational power in the network is

$$N = \frac{\text{Rev} + e \cdot S}{c_m}, \tag{2}$$

where Rev *is the total transaction fees in USD paid by users per unit time and* e *is the USD/bitcoin exchange rate.*

Theorem 1 holds regardless of the number of large miners. In particular, free entry of small miners precludes large miners from profitably affecting transaction fees even if all large miners collude.

The proof relies on free riding by small miners. For example, large miner i may choose to process only transactions with a fee above b', leading to a subgame in which some users increase their transaction fees above b' to avoid being delayed when miner i is selected. If the increased fees outweigh the loss from not processing transactions with a fee lower than b', choosing such a block assembly policy can increase miner i's expected transaction fees per block. However, this creates a larger increase in the expected transaction fees per block of small miners because small miners benefit from the increased fees while still processing all transactions. Entry by small miners increases the aggregate computational power so that small miners break even. Because small miners collect more fees than a large miner attempting to affect fees, free entry implies the large miner either breaks even or is strictly worse off.

Proof. Consider an arbitrary profile of choices by large miners and a subgame perfect equilibrium of the induced subgame for small miners and users. By Assumption 3, there are small miners that are active. Consider such an active small miner. Since small miners are non-atomic, the small miner's choice of block assembly policy $A(\cdot)$ does not affect η or N. Therefore, it does not affect G, γ and the steady-state distribution of $\boldsymbol{b}$. Since for any fixed distribution of $\boldsymbol{b}$ we have that $\text{Rev}(A^* \mid G, \gamma, \eta) = \max_A \{\text{Rev}(A \mid G, \gamma, \eta)\}$, it is a best response for the active small miner to choose A^*. Furthermore, any block assembly policy that constitutes a best response must give the small miner the same payoff as A^*.

Also by Assumption 3, there are inactive small miners, and therefore small miners must be indifferent between being inactive or active with A^*,

$$\frac{1}{N}\Big(\text{Rev}(A^* \mid G, \gamma, \eta) + e \cdot S\Big) - c_m = 0\,,$$

yielding

$$\text{Rev}(A^* \mid G, \gamma, \eta) + e \cdot S = c_m N\,. \tag{3}$$

Now consider a large miner i who can affect the distribution η and thereby affect

G, γ. Let x_i, A_i denote the computational power and block assembly policy of miner i, respectively. Fix the choices of other large miners, fix $x_i \geq 0$, and let G^{A_i}, γ^{A_i}, η^{A_i}, and N^{A_i} be distributions and values induced by a subgame perfect equilibrium of the subgame induced by miner i's choice of A_i, holding all other choices by large miners fixed. We have that

$$\begin{aligned}
&\frac{x_i}{N^{A_i}}\Big(\text{Rev}\left(A_i \mid G^{A_i}, \gamma^{A_i}, \eta^{A_i}\right) + e \cdot S\Big) - c_i\left(x_i\right) \\
\leq &\frac{x_i}{N^{A_i}}\Big(\text{Rev}\left(A^* \mid G^{A_i}, \gamma^{A_i}, \eta^{A_i}\right) + e \cdot S\Big) - c_i\left(x_i\right) \\
= &\frac{x_i}{N^{A_i}} c_m N^{A_i} - c_i\left(x_i\right) \\
= &c_m x_i - c_i\left(x_i\right) .
\end{aligned} \tag{4}$$

The inequality follows because holding G, γ, η, N fixed, A^* delivers higher revenue than any A. The first equality follows from (3).

For $A_i = A^*$, using (3) we have that

$$\begin{aligned}
&\frac{x_i}{N^{A_i}}\Big(\text{Rev}\left(A_i \mid G^{A_i}, \gamma^{A_i}, \eta^{A_i}\right) + e \cdot S\Big) - c_i\left(x_i\right) \\
= &\frac{x_i}{N^{A^*}} c_m N^{A^*} - c_i\left(x_i\right) \\
= &c_m x_i - c_i\left(x_i\right) .
\end{aligned}$$

We thus showed that given any profile of choices of other large miners and any best responses of users and small miners, miner i attains the maximal payoff of

$$\sup_{x_i}\left\{c_m x_i - c_i\left(x_i\right)\right\} = c_m x_i^* - c_i\left(x_i^*\right)$$

by selecting the block assembly policy A^* and the computational power x_i^* that is either the unique solution to $c_i'\left(x_i^*\right) = c_m$ or $x_i^* = 0$ if no solution exists.

Consider a profile where all active small miners choose A^* and each large miner i chooses A^* and x_i^*. Denote by η^* the implied distribution of computational power, which is given by $\eta^*\left(A^*\right) = N$ and $\eta^*\left(A\right) = 0$ for all $A \neq A^*$. Complete the description of the strategy profile by having small miners and users play some subgame perfect equilibrium following any possible deviation by a large miners. The arguments above show this profile constitutes a subgame perfect equilibrium, as large miners, small

miners, and users all play a best response.

Since $\rho < 1$, all transactions are eventually processed and $\text{Rev} = \text{Rev}\left(A^* \mid G, \gamma, \eta^*\right)$ is equal to the total transaction fees (in USD) per unit time under a subgame perfect equilibrium of the induced subgame for users (which will be characterized in the next section). Rewriting (3) we have that

$$N = \frac{\text{Rev} + e \cdot S}{c_m}.$$

□

Large miners can make positive profits if their average cost per computational unit is below c_m.[17] For the case where large miners do not have a computational cost advantage, we obtain the following immediate corollary of Theorem 1.

Corollary 1. *If all large miners have the same cost c_m per computational unit, that is, $c_i(x) = c_m x$ for all large miners i, then all miners make zero profit.*

While choosing block assembly policy A^* is a weakly dominant strategy, we have not yet ruled out other equilibria in which large miners may choose other block assembly policies. To formally preclude other equilibria, we introduce a perturbation that ensures the distribution of pending transaction has full support. Let G_0 be a distribution with strictly positive density over $\mathbb{R}_+$ (e.g., the half-normal distribution). The ε-perturbed system is given by adding to the original game exogenous arrivals of transactions and blocks. Additional transactions arrive according to a Poisson process with rate ε, each with a fee independently drawn from G_0. Additional blocks arrive according to a Poisson process with rate ε, and these blocks process all pending transactions (regardless of their number). The ε-perturbed game is identical to the original game, except for payoffs being determined by the steady-state distribution of the ε-perturbed system.

We say that two block assembly policies A, A' are G_0-equivalent if $A(\boldsymbol{b}) = A'(\boldsymbol{b})$ with probability 1 for $\boldsymbol{b}$ that is generated by independently drawing a geometrically distributed number of transactions from G_0. Notice that if A, A' are G_0-equivalent, they are also payoff equivalent. For any $\varepsilon > 0$, the argument in the proof of Theorem 1 implies that any block mining policy A that is not G_0-equivalent to A^* is strictly

[17] For example, miners who position their servers near dams can have lower cost due to cheap electricity. If such opportunities are scarce and can support only a limited number of servers they will not be competed away.

dominated. Thus, the equilibrium described in Theorem 1 is the unique equilibrium (up to payoff irrelevant variations) that survives the perturbation.

Proposition 2. *For any $\varepsilon > 0$, in any subgame perfect equilibrium of the ε-perturbed game, all active miners choose the block assembly policy that is G_0-equivalent to A^*.*

Proof. Let $\text{Rev}\left(A \mid G, \gamma, \eta, G_0, \varepsilon\right)$ denote the expected transaction fees per block in the ε-perturbed game. If $A\left(\boldsymbol{b}\right) \neq A^*\left(\boldsymbol{b}\right)$ with positive probability for $\boldsymbol{b}$ (given the steady-state distribution of pending transaction $\boldsymbol{b}$), we have that

$$\text{Rev}\left(A \mid G, \gamma, \eta, G_0, \varepsilon\right) < \text{Rev}\left(A^* \mid G, \gamma, \eta, G_0, \varepsilon\right) .$$

Thus, we can replace the weak inequality in (4) with a strict inequality. Following the remainder of the proof of Theorem 1, it follows that x_i, A is strictly dominated by x_i^*, A^*. Finally, we have that if $A\left(\boldsymbol{b}\right) = A^*\left(\boldsymbol{b}\right)$ with probability 1 (given the steady-state distribution of pending transaction $\boldsymbol{b}$) it must be that A, A^* are G_0-equivalent, since for any k there is a positive probability that a clearing block will be immediately followed by k arrivals of transactions drawn from G_0. □

Entry by small miners is essential for Theorem 1. Suppose a single large miner can control all the mining infrastructure and preclude entry. While the blockchain protocol provides some security guarantees even when there is a single miner, a single miner will be able to set a minimal transaction fee because the single miner can ensure that any transaction that offers a lower fee will not be processed. The single miner can preclude entry of small miners if it maintains the reward per computational unit strictly below c_m and can make positive profits if his own cost is lower than c_m.

Our analysis presents a stylized view of miners, thereby abstracting from various real-world issues. Actual miners incur fixed costs to purchase mining equipment; available equipment is heterogeneous in price, quality, and vintage; innovative equipment manufacturers are also miners; electricity costs are location- and possibly miner-dependent. Future work will take up these nuances.

4.2 User Behavior and Equilibrium Transaction Fees

We now characterize user behavior in step (iii). The analysis in Section 4.1 shows that all miners, small and large, choose the block assembly policy A^*. The remainder of the paper maintains that miners follow this behavior and characterizes the induced

subgame for users. In this context, the term equilibrium means the subgame perfect equilibrium behavior of users in the subgame induced by all miners choosing A^*, i.e., each block processes the K pending transactions which offer the highest transaction fees. The number of miners does not affect μ, the rate at which blocks are generated, or K, the block size, and therefore the number of miners does not affect users' choice of transaction fees.

From a user's perspective, the higher the fee he offers, the more likely the transaction will be included in an earlier block. Consider an equilibrium where all potential users participate and post their transactions in the system, with $G(\cdot)$ denoting the cumulative distribution function of the chosen transaction fees. A user i with delay cost c_i and WTP R_i who decides to post a transaction chooses his transaction fee b to maximize his net reward

$$R_i - b - c_i \cdot W(b \mid G), \tag{5}$$

with $W(b \mid G) = W(b \mid G, 1, \eta^*)$ denoting the equilibrium expected delay given transaction fee b and the CDF G. For brevity, we omit the dependence on the distribution of block assembly policies, as we maintain that all miners adopt the block assembly policy A^*. The following lemma characterizes the equilibrium expected delay.

Lemma 1. *In any equilibrium in which all potential users participate, the expected delay for a user with delay cost c_i is*

$$\mu^{-1} W_K(\hat{\rho}(c_i)) \tag{6}$$

where $\hat{\rho}(c_i) = \lambda \bar{F}(c_i)/K\mu = \rho \cdot \bar{F}(c_i)$ is the effective load from transaction with higher delay cost, and the function $W_K(\cdot)$ gives the expected number of blocks that pass until the transaction is processed.

The function $W_K(\cdot)$ is specified in Appendix F.1. In particular, $W_K(0) = 1$ and $W_K'(\hat{\rho}) \geq 0$ for $\hat{\rho} \in [0, 1)$.

The intuition for Lemma 1 is as follows. Users face a queuing game where higher transaction fees imply higher processing priority. Standard arguments (see Hassin & Haviv (2003)) imply that users with higher delay cost will pay higher transaction fees and receive higher priority, and therefore the arrival rate of transactions with higher priority is $\lambda \cdot \bar{F}(c)$. Analysis of the stochastic system shows that the number of blocks that pass until a transaction is processed depends only on the block size K and the

effective load from higher priority transactions $\hat{\rho}(c_i) = \lambda \bar{F}(c_i)/K\mu$. Although $\rho < 1$ implies the system has sufficient capacity to process all transactions on average, the randomness of the arrival times implies the possibility of backlogs. The expression (6) captures the expected wait from such cases. Finally, the term μ^{-1} in (6) enables the statement of the result in terms of calendar time rather than the number of blocks. The particular function $W_K(\cdot)$ endogenously arises by the incentives set in the protocol. Appendix E provides a plot of $W_K(\cdot)$.

Users' individual optimization implies:

Proposition 3. *Assuming that all potential users participate, there is a unique equilibrium. In it, a user with waiting cost $c_i \in [0, \bar{c}]$ chooses to pay a transaction fee $b(c_i)$, given by*

$$b(c_i) = \rho \int_0^{c_i} f(c) \cdot c \cdot \mu^{-1} W_K' \left(\rho \bar{F}(c)\right) dc. \tag{7}$$

These transaction fees coincide with the payments that result from selling priority of service in a VCG auction.

The net reward for a user with delay cost c_i and WTP R_i is

$$u(R_i, c_i) = R_i - \mu^{-1} \int_0^{c_i} W_K \left(\rho \bar{F}(c)\right) dc. \tag{8}$$

The Bitcoin protocol indirectly entails a priority auction, although no auctioneer is present. Users with higher waiting costs pay higher transaction fees and wait less. Users' bids have the VCG property that each user bids an amount equal to the externality he imposes on others by delaying their transactions. Equation (8) implies that users with lower delay cost c_i bear lower total costs (total of paid fees and delay costs). This is due to information rents. The highest costs are born by users with $c_i = \bar{c}$ and are equal to $\bar{R} = \mu^{-1} \int_0^{\bar{c}} W_K \left(\rho \bar{F}(c)\right) dc$.

The equilibrium allocation of priority is efficient. However, the allocation of delay takes the particular form because of the blockchain design. A different design or increased values of μ, K can reduce waiting costs for all transactions. Note that transaction fees depend on ρ and therefore will change with changes in λ, μ, K.

Finally, we verify that all potential users prefer to participate under the assumption that WTP is sufficiently high given the load ρ.

Lemma 2. *Let* $\bar{R} = \mu^{-1} \int_0^{\bar{c}} W_K\left(\rho \bar{F}(c)\right) dc$. *If* $R_H \geq R_L > \bar{R}$ *there is a unique equilibrium where all potential users participate. In equilibrium, all users receive strictly positive net reward.*

Thus, equilibrium behavior of users does not depend on their WTP R, assuming that it is sufficiently high. All users participate regardless of their WTP, and the transaction fees paid are independent of WTP. Each user pays a fee equal to the externality he imposes on other users, and since all transactions are eventually processed, the externality involves only delays to other transactions.

Transaction fees under the firm and the BPS depend on different parameters. The firm sets prices based on user WTP, and transactions that do not pay the required fee are not processed. Under the BPS, prices are determined in equilibrium based on user delay costs. All transactions are processed regardless of the fees they offer. Some users offer higher fees to reduce delays. Transactions which offer lower or zero fees are processed with greater delays. The BPS transaction fees depend only on the parameters K, μ, ρ, and the distribution of delay costs F. The transaction fees are nominally denominated in the system's native currency, but their value in USD is independent of the exchange rate e.

We summarize these results in the following theorem.

Theorem 2. *Let* $\rho = \lambda/\mu K \in (0,1)$ *and assume that*

$$R_H \geq R_L > \bar{R} = \mu^{-1} \int_0^{\bar{c}} W_K\left(\rho \bar{F}(c)\right) dc. \tag{9}$$

There is a unique equilibrium where all potential users participate and receive strictly positive surplus. Equilibrium transaction fees paid by users are independent of user WTP R_H, R_L, *and of the exchange rate* e.

Despite having excess capacity (i.e., $\rho < 1$*), the system raises strictly positive revenue from transaction fees.*

As seen in Section 3, a profit-maximizing firm will raise prices until some users receive no net benefit. The possibility that all users are net beneficiaries of the system distinguishes its service from a similar service provided by a profit-maximizing firm.

Another distinguishing feature of the system is its commitment to congestion pricing, a commitment that is difficult to modify even when circumstances change. Thus, the users are protected from being held up should they get locked into the

BPS: if users lose their alternative payment methods then their WTP for the system goes up, but because transaction fees are independent of the WTP R (given that R_H, R_L are sufficiently high), users are protected from price increases. In contrast, users should be wary of getting locked into a conventional payment system, as a firm would raise prices should its users lose their alternative options (Grossman & Hart 1986).

We highlight this as the following corollary.

Corollary 2. *Assume that the conditions of Theorem 2 are satisfied. Then, an increase in WTP R does not change equilibrium transaction fees.*

Corollary 2 may appear as good news to users. However, the pricing level depends on the congestion in the system $\rho = \lambda/\mu K$ and may be inefficient.

4.3 Determination of Infrastructure Level and Welfare

Building on the two preceding subsections, this subsection shows the total revenue from transaction fees and the system's level of infrastructure. Moreover, it calculates the welfare level associated with the BPS and compares it to that delivered by a profit-maximizing firm. The following considers the equilibrium characterized by Theorems 1 and 2, and assumes the conditions of Theorem 2 are satisfied.

Aggregating (7) over all users delivers

Theorem 3. *Total revenue from transaction fees per unit time is*

$$\text{Rev}_K(\rho) = K\rho^2 \int_0^{\bar{c}} cf(c)\bar{F}(c)W_K'\left(\rho\bar{F}(c)\right)\, dc. \tag{10}$$

Equation (10) complements equation (2) to determine the network's computational power in equilibrium. Equation (10) shows that total revenue from transaction fees depends only on K, ρ, and the distribution of delay costs F. It implies that the revenue depends on μ and λ only through $\rho = \lambda/\mu K$. Thus, holding the type distribution function F fixed, a system with double the demand λ and double the block rate μ will raise the same amount of revenue as the original system but will have twice as many users, each of whom will pay half the transaction fee paid by the corresponding user in the original system.

Note that there is no guarantee that the equilibrium number of miners is adequate for the system's reliability and security. The protocol can dictate the amount of newly

minted coins S that are awarded to miners, but the exchange rate e may fluctuate during the life of the system. The revenue from transaction fees does not depend on the exchange rate, but varies with the congestion ρ which is a function of the predetermined parameters μ, K as well as the potential demand λ that may change over time. Moreover, a shortage of mining resources does not lead to higher fees or a more favorable exchange rate; if anything, it is likely to result in the opposite. On the other hand, an abundance of mining resources does not lead to lower fees or a less favorable exchange rate. The equilibrium analysis is applicable if user WTP for the system R_H, R_L are sufficiently high given the equilibrium number of miners N.

Next, we calculate welfare by accounting for the total benefits and costs of the system. Since all users are served, the system generates $\lambda_H R_H + \lambda_L R_L$ for users per unit time. The users pay transaction fees and incur delay costs. All miners receive a reward equal to c_m per mining unit. Marginal miners whose cost is c_m will therefore break even and spend all the revenue they receive on operating costs.

Theorem 4. *If all miners have a cost c_m per computational unit and no new coins are minted*[18] *then welfare is given by*

$$\lambda_H R_H + \lambda_L R_L - \text{DelayCost}_K(\rho) - c_m \cdot N \tag{11}$$

where the total delay costs incurred by users is

$$\text{DelayCost}_K(\rho) = K\rho \int_0^{\bar{c}} c f(c) W_K\left(\rho \bar{F}(c)\right) dc. \tag{12}$$

Miners break even and spend all the revenue they receive on operating costs.

The total benefit from processing transactions is $\lambda_H R_H + \lambda_L R_L$, as all transactions are processed. The cost $c_m \cdot N$ is the cost of server infrastructure, where competition between the miners ensures that infrastructure is provided at cost c_m and miners make no profit. The delay costs $\text{DelayCost}_K(\rho)$ are necessary in order to raise revenue from users, as users have an incentive to pay higher transaction fees only if transactions with low fees suffer delays.

If, in deviation from the theorem's assumption, some miners have an average cost

[18] That is, $S = 0$, as will be the case for the BPS in the long run. Currently the BPS funds most of its mining cost by minting new coins. The welfare calculations remain unchanged if the BPS can mint a finite amount of new coins and the opportunity cost of awarding the coins to miners is equal to its value. We defer determination of the welfare costs of minting new coins to future work.

lower than c_m, they make a profit. In such case, welfare will be higher by these miners' profit.

This allows us to compare the BPS and a conventional payment system that is run by a firm. Under our assumptions, the cost of operating the BPS is $c_m \cdot N$, while the cost of operating a firm-run payment system is $c_f \cdot \lambda_H$. It appears that it is more expensive to run the BPS because the decentralized protocol requires additional computational overhead. Moreover, if the BPS is successful and popular, the implied congestion can lead to an equilibrium value of N that is too high. The BPS also has the additional delay cost $\text{DelayCost}_K(\rho)$, while the firm processes transactions immediately. On the other hand, the BPS serves all potential demand, while under the firm there is a dead-weight loss because R_L users are not served, losing $\lambda_L \cdot R_L$ of potential generated value. Altogether, we get that if

$$\lambda_L R_L > c_m \cdot N - c_f \lambda_H + \text{DelayCost}_K(\rho) \tag{13}$$

welfare is higher under the BPS than under a firm. Note that the two sides of inequality (13) depend on different sets of parameters, and therefore the comparison can go either way. Essentially, the BPS allows society to pay for a more costly infrastructure on which competitive pricing is guaranteed, and that can be beneficial if dead-weight loss is substantial.

Beyond this calculations-based comparison, there are differences worth mentioning. For instance, a firm-run system operates under the legal system and can offer procedures to retrieve lost accounts and reverse erroneous or fraud-inspired payments. The BPS cannot offer such services, but is transparent and does not require trust in any individual component.

5 Protocol Design for Efficient Congestion Pricing

The following corollary of Section 4 motivates this section's main question, namely how to set the system's parameters K and μ in response to λ in order to achieve desired combinations of fee revenue and delays.

Corollary 3. *In equilibrium, if $\rho = 0$, both delay cost and revenue are zero. For any fixed K, both revenue (and with it, infrastructure provision by miners) and delay cost are strictly increasing in ρ.*

Figure 2 shows how revenue from transaction fees and delay cost vary with ρ under the parameters $K = 2,000$ and $c \sim U[0,1]$. The figure assumes that all agents participate, and therefore revenue tends to infinity as $\rho \to 1$. When agents choose whether to participate, revenue will be bounded, as agents may not participate as the system gets congested (see Appendix B). The figure looks similar for other distributions of delay costs (see Appendix E for a plot of other distributions).

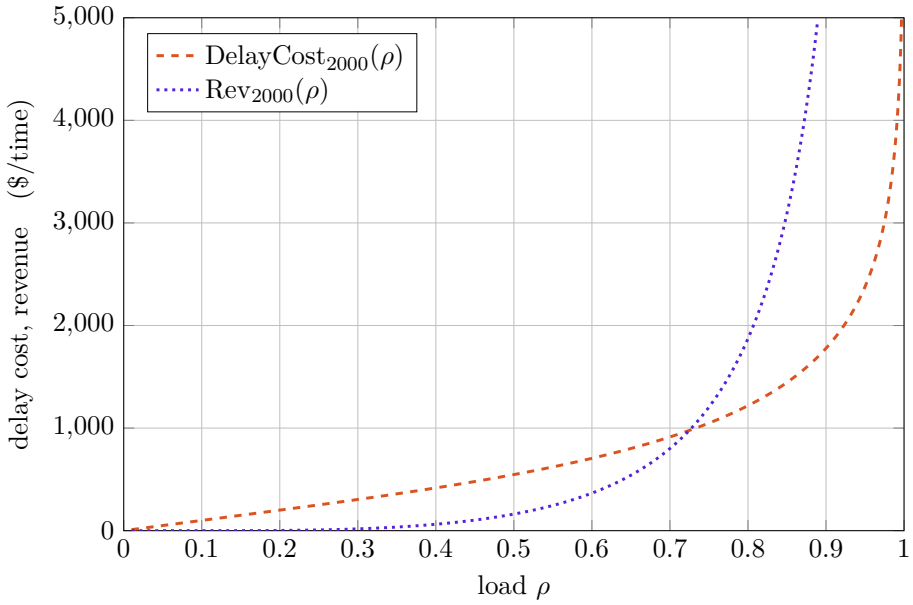

Figure 2: Revenue and delay cost for varying congestion level ρ. Delay costs are distributed according to $c \sim U[0,1]$ and the block size is $K = 2,000$.

The current Bitcoin protocol uses fixed capacity parameters K and μ, and therefore the congestion ρ varies with demand. This is undesirable, as the amount of revenue generated can be too high or too low relative to the desired levels of reliability and security. An alternative design should adjust K, μ to accommodate demand variations and thereby maintain desirable levels of congestion and revenue.

While our focus is on the economic aspects of the design, we note that designing such a decentralized protocol raises engineering challenges. First, the protocol must maintain agreement on K, μ among the independently operating miners. Thus, the parameter adjustment rule must be encoded in the protocol and use only information shared among all miners. If $\rho < 1$, a rule that uses the volume of recently processed transaction as a proxy for demand can dynamically adjust K, μ and maintain agreement on them.[19] Second, the consensus protocol may constrain K, μ. The Nakamoto

[19]Such a rule can be implemented by modifying the adjustment of the hash difficulty (as explained

consensus protocol requires that block inter-arrival times are sufficiently large relative to the network lag given the block size.[20] New designs may allow a larger range of parameters.[21] In the analysis below, we determine the ideal K, μ from an economic perspective. Addressing the engineering limitations is left for future work. Our suggestion can guide the choice of K, μ within the feasible range.

The choice of K, μ should achieve the target revenue from transaction fees and should minimize the delay costs imposed on users. Note that by an appropriate choice of K, μ in response to demand λ, we can achieve the desired ρ and desired revenue from transaction fees in USD, regardless of exchange rate fluctuations. Although transaction fees are denominated in bitcoin, their USD value reflects the USD value of shortening delays. The protocol obtains the USD market value of delay reduction without the need to learn the exchange rate.

Raising revenue from transaction fees requires positive ρ and therefore delay costs. To better understand the dependency on K, μ, and the implied trade-offs between revenue and delay costs, we provide the following simplified approximate expressions.

Lemma 3. *For any $\hat{\rho} \in [0, 1)$ we have that*[22]

$$\lim_{K \to \infty} W_K(\hat{\rho}) = W_\infty(\hat{\rho}) = 1 + \frac{1}{\rho} e^{-1/\rho} + o\left(\frac{1}{\rho} e^{-1/\rho}\right)$$

where the function $W_\infty \colon [0, 1) \to [1, \infty)$ is explicitly given in Appendix F.4. Moreover, $W_\infty(0) = 1, W'_\infty(0) = 0$ and $W'_\infty(\hat{\rho}) > 0$ for $\hat{\rho} \in (0, 1)$.

A given transaction with $\hat{\rho} \in [0, 1)$ will be processed within $W_K(\hat{\rho})$ blocks on average. We have that $1 \leq W_K(\hat{\rho}) < \infty$ because the inclusion of a transaction in

in Appendix A). Currently, the difficulty adjusts in accordance with the total computing power of the network to maintain average block mining frequency of 10 minutes. Our suggested alternative design can similarly adjust the difficulty to maintain that on average a fraction ρ of blocks is used.

[20] Croman et al. (2016) studies the limitations of the computer network operating the BPS. Pass et al. (2017) provides theoretical bounds for block rate in the Nakamoto consensus.

[21] Bitcoin's capacity limitations led to many suggestions of alternative protocols. For example, Sompolinsky & Zohar (2015) suggests the GHOST protocol in which blocks form a tree (instead of a chain); Gilad et al. (2017) and Bentov et al. (2016) suggest alternative proof of stake protocols. Many of these suggested protocols maintain the main features of our model (in particular, batch processing of transactions), and can incorporate similar congestion pricing mechanisms.

[22] Given arbitrary functions $f(\cdot)$ and $g(\cdot)$ and a positive function $h(\cdot)$, as $\rho \to 0$, we will say that $f(\rho) = g(\rho) + O(h(\rho))$ if $\limsup_{\rho \to 0} |f(\rho) - g(\rho)|/h(\rho) < \infty$, i.e., if the difference between f and g is asymptotically bounded above by *some* constant multiple of h. Similarly, we will say that $f(\rho) = g(\rho) + o(h(\rho))$ if $\limsup_{\rho \to 0} |f(\rho) - g(\rho)|/h(\rho) = 0$, i.e., if the difference between f and g is asymptotically dominated by *every* constant multiple of h.

a block depends both on how many pending transactions have accumulated at the time the block is generated as well as how the priority of the given transaction ranks among the accumulated transactions. The former is random due to the random time between blocks, and the latter is random due to the random arrival of transactions. When blocks are fairly large, there is still randomness due to their random arrival time, but the arrival of higher priority transactions does not create much additional randomness.[23] As a result, $W_K(\hat{\rho})$ is almost independent of K for large K. Calculations show that the approximation already appears good for $K = 20$; with Bitcoin's $K = 2000$ we can comfortably use this approximation. For additional intuition and the proof of Lemma 3, see Appendix F.4.

Using Lemma 3, we can give the following simplified expressions for revenue and delay costs.

Theorem 5. *For a fixed-load $\rho \in [0, 1)$, as the block size $K \to \infty$, we have that*[24]

$$\begin{aligned}\mathrm{Rev}_K(\rho) &= K \cdot \mathrm{Rev}_\infty(\rho) + o(K),\\ \mathrm{DelayCost}_K(\rho) &= K \cdot \mathrm{DelayCost}_\infty(\rho) + o(K),\end{aligned}$$

where

$$\begin{aligned}\mathrm{Rev}_\infty(\rho) &\triangleq \rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_\infty\left(\rho\bar{F}(c)\right)\, dc,\\ \mathrm{DelayCost}_\infty(\rho) &\triangleq \rho \int_0^{\bar{c}} cf(c) W_\infty\left(\rho\bar{F}(c)\right)\, dc.\end{aligned}$$

Theorem 5 offers simple approximations of the dependencies of revenue and delay

[23] To gain intuition, consider a user i with delay costs c_i that posts a transaction at time t_0 when there are no pending transactions. The following block arrives after some random time $t \cdot \mu^{-1}$, where $t \sim Exp(1)$. The probability that i's transaction is included in the following block is the probability that, between t_0 and $t_0 + t \cdot \mu^{-1}$, less than K higher priority transactions arrive. The number of higher priority transactions given t has distribution $X_t \sim \mathrm{Poisson}\left(\lambda\bar{F}(c_i) \cdot t\mu^{-1}\right) = \mathrm{Poisson}(t \cdot K\hat{\rho})$. The realized number is random because t is random and also because the number of arrivals given t, X_t, is random. However, the variance of X_t is of order K, and therefore, as $K \to \infty$, the number of arrivals given t measured in block equivalents, X_t/K, can be well approximated by its expectation $t\hat{\rho}$. Thus, the probability that the transaction will be included in the next block converges according to $\mathsf{P}(X_t < K) \to \mathsf{P}\left(t < \hat{\rho}^{-1}\right)$, which only depends on $\hat{\rho}$.

[24] Given arbitrary sequences $\{f_K\}$ and $\{g_K\}$, and a positive sequence $\{h_K\}$, as $K \to \infty$, we will say that $f_K = g_K + o(h_K)$ if $\limsup_{K\to\infty} |f_K - g_K|/h_K = 0$, i.e., if the difference between f and g is asymptotically dominated by *every* constant multiple of h. Similarly, we will say that $f_K = g_K + \Omega(h_K)$ if $\liminf_{K\to\infty} |f_K - g_K|/h_K > 0$, i.e., if the difference between f and g is asymptotically bounded below by *some* constant multiple of h.

costs on K. The expressions $\text{Rev}_\infty(\rho), \text{DelayCost}_\infty(\rho)$ are functions of only ρ and F. To a good approximation, the dependency of $\text{Rev}_K(\rho), \text{DelayCost}_K(\rho)$ on K is only through a scaling factor of both of these expressions. See Appendix E for plots showing the goodness of approximation.

Note that Theorem 5 critically relies on the randomness of block inter-arrival times. If $\rho < 1$ and blocks were to arrive at deterministic fixed time intervals (say, exactly every 10 minutes), then for large K every pending transaction would be processed in the next block. Hence users would not have incentive to pay any transaction fees. The random arrival of blocks allows the system with large blocks to generate revenue even when $\rho < 1$.

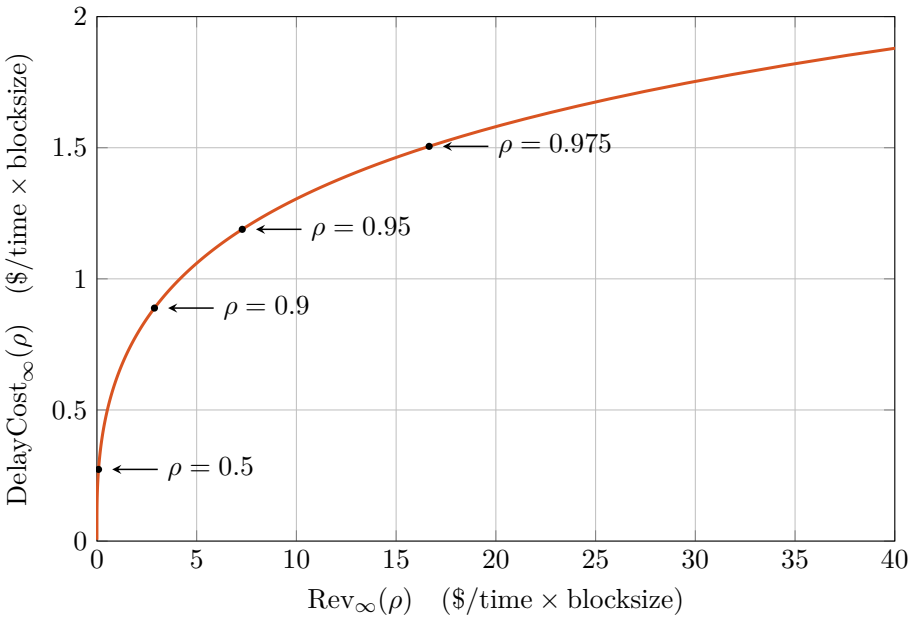

Figure 3: The parametric curve $(\text{Rev}_\infty(\rho), \text{DelayCost}_\infty(\rho))$ *for* $\rho \in [0, 1)$*, describing (up to a scaling by blocksize) the achievable combinations of revenue and delay cost for systems with large blocksize. The distribution of delay costs is taken to be* $c \sim U[0, 1]$*.*

Figure 3 plots how the pairs $(\text{Rev}_\infty(\rho), \text{DelayCost}_\infty(\rho))$ vary with ρ, assuming the distribution of delay costs is $c \sim U[0, 1]$. From Theorem 5, the pairs $(\text{Rev}_K(\rho), \text{DelayCost}_K(\rho))$, for any fixed K and varying ρ, are scaled versions of the depicted curve. Thus, the curve informs us of the delay costs that are necessary for raising a given amount of revenue for any K.

The figure shows that a significant amount of delay cost is necessary to raise even a small amount of revenue. We formally show this in Theorem 6.

Theorem 6. *For any F, as $\rho \to 0$, we have that*

$$\mathrm{Rev}_\infty(\rho) = O\left(e^{-1/\rho}\right),$$
$$\mathrm{DelayCost}_\infty(\rho) = \rho \cdot \mathbb{E}[c] + o(\rho).$$

In other words, for small values of the load ρ, the delay cost grows linearly but the revenue grows more slowly than any polynomial.

The intuition is as follows. For $\rho \approx 0$, all transactions are likely to be processed in the next block regardless of their priority because the block is unlikely to reach its maximal size. In contrast, total delay costs scale linearly, as every transaction needs to wait for at least one block and higher ρ implies more waiting. Therefore, as the load increases from $\rho \approx 0$, both revenue and delay costs increase but delay costs grow more than exponentially faster than revenue.

Together with Theorem 5, this implies that using a larger K to raise a desired level of revenue R^* will yield unfavorable results. We formally state this as the following theorem.

Theorem 7. *Consider a desired level of revenue $R^* > 0$ and a block size K. Define $\mathrm{DelayCost}^*_K(R^*)$ to be the delay cost required to achieve revenue R^* under the approximation for large K, i.e.,*

$$\mathrm{DelayCost}^*_K(R^*) \triangleq K\,\mathrm{DelayCost}_\infty\left(\mathrm{Rev}^{-1}_\infty(R^*/K)\right),$$

with $\mathrm{Rev}^{-1}_\infty(R^*) \triangleq \inf\left\{\rho > 0 \;:\; \mathrm{Rev}_\infty(\rho) \geq R^*\right\}$ being the minimal load required to achieve revenue R^*.

Then,

$$\mathrm{DelayCost}^*_K(R^*) = \Omega\left(\frac{K}{\log K}\right).$$

Figure 4 illustrates the possible attainable values for revenue and delay given different values of K and ρ, assuming delay costs are distributed uniformly in $[0, 1]$. Each curve shows the attainable values for revenue and delay for a fixed value of K and a range of possible ρ. The plot shows that a lower value of K allows raising any level of revenue at a lower delay cost to users.

Each curve's two main features are (i) monotonicity, i.e., longer delays are required to generate more revenue, and (ii) the curve is asymptotically vertical at the origin, i.e., to move from zero to some revenue, the delay cost has to be substantial. These

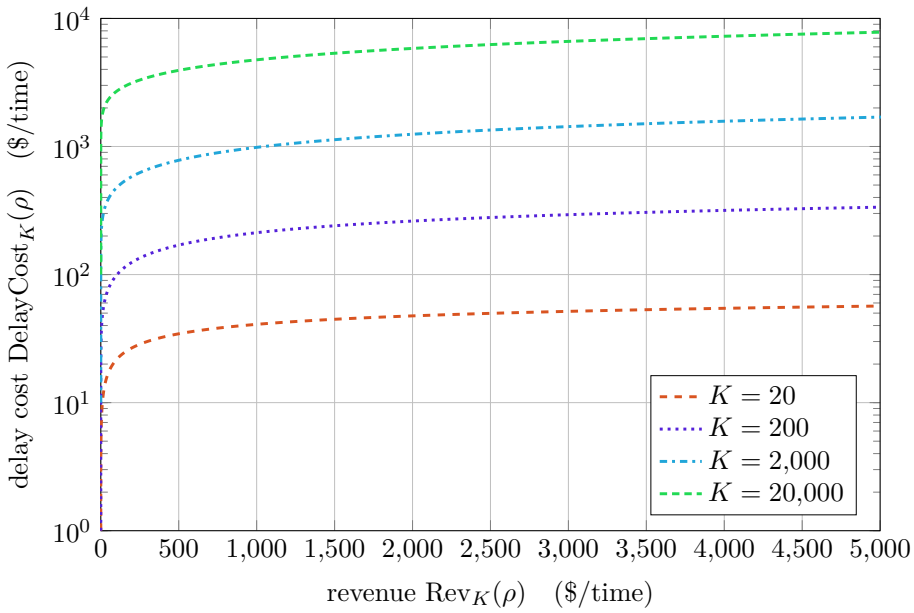

Figure 4: Possible pairs of revenue and delay cost as ρ varies, for different values of K, where delay costs are distributed according to $c \sim U[0, 1]$.

insights transcend the specific $U[0, 1]$ distribution of c underlying the figure. However, note that these calculations ignore technological constraints and assume that no users opt out of the system. All curves are approximately a scaled version of the curve in Figure 3 (note the logarithmic scale for the vertical axis), as implied by Theorem 5.

To summarize, this analysis suggests the following simple adaptations to the current protocol. First, a smaller block size K is preferable. Second, an adjustment of the block rate to $\mu = \lambda / (K\rho^*)$ in response to demand λ. This keeps congestion constant at ρ^*, yielding a stable, desired level of revenue.[25]

6 Data

6.1 Mining Profitability

We compare our results to empirical estimates given by Croman et al. (2016), who estimate that the total expenditure of miners during October 2015 was approximately 5,840 USD per block. Croman et al. (2016) attribute the vast majority of the cost to the costs of electricity and hardware used in the attempts to get selected to mine

[25] Clearly, there are communication and other limitations that limit the range of feasible μ and K. This paper ignores these engineering challenges.

the next block. During that period, the mining reward per block was 25 bitcoins plus negligible transaction fees, or approximately 6,000 - 7,500 USD (the bitcoin-USD exchange rate fluctuated during the month). This back of the envelope calculation suggests that miners who buy electricity at market prices approximately break even, which is consistent with our analysis. Websites that offer information to potential miners about mining profitability of various cryptocurrencies[26] give advice that is consistent with this observation. Furthermore, while some groups controlled a significant fraction of the computational power in the network, there is no evidence that even large miners tried to influence fee levels.

6.2 The Relation Between Congestion and Transaction Fees

Average block size in MB can be used as a measure of the actual congestion in the BPS. In practice, the BPS limited blocks to 1MB of data per block until August 21st 2017. This corresponds to approximately $K = 2,000$ transactions per block. In our model, the congestion parameter ρ is equal to the average number of transactions per block divided by K. Analogously, we interpret the average size of a block relative to the 1MB limit as a proxy for congestion ρ. Each point in Figure 1 corresponds to one day in the BPS, displaying daily average transaction fees per block and daily average block size.[27] The plot also includes a solid line generated by our model as follows. We set $K = 2,000$, and normalize time so that a time unit is 10 minutes and set $\mu = 1$. The distribution of users' delay cost is unknown and arbitrarily set to $F = U[0, \bar{c}]$ with $\bar{c} = 0.1$ USD/10 minutes. The resulting total revenue per unit time $\text{Rev}_{2000}(\cdot)$ is the expected total transaction fees per block, which is displayed by the solid black line in Figure 1.

Note that the solid line produced by our model matches the broad patterns in the data. Figure 1 shows that transaction fees are negligible when congestion is low. Transaction fees become substantial when congestion reaches 80%. Transaction fees increase rapidly as congestion approaches 1, even though the system has excess capacity.

[26] https://www.coinwarz.com/cryptocurrency/, retrieved 6/20/2017.

[27] Transaction fee and block size data is from http://blockchain.info, and the number of blocks per day is from https://data.bitcoinity.org. Each point is a daily average over the interval 4/1/2011–6/30/2017. The starting date 4/1/2011 was selected as this is roughly when the fees per block started exceeding 1 USD. The end date does not extend to present day because the BPS changed the method for calculating a block's size in August 2017.

7 Conclusion

Bitcoin presents a computer science breakthrough, showing the feasibility of a decentralized payment system that relies on a collection of unrelated parties without the need for a central intermediary. This paper shows that Bitcoin also provides an economic innovation that can address concerns of the harm of monopoly power of platforms. The BPS shows the feasibility of a decentralized platform in which users are protected from the harms of monopoly pricing, even if users have no alternative to the platform. The platform can fund itself by user fees that are determined in a market equilibrium. Competition and free entry among the service providers renders all participants to be price takers.

Critical ingredients of our analysis are costly effort on the part of miners combined with free entry and exit. Our results can be extended to other protocols, e.g., Proof of Stake, should they retain these ingredients. Issues left unaddressed in this paper include engineering challenges limiting the scale at which a decentralized system can operate; ensuring the security of the system against attacks; determination of the coin's exchange rate and its volatility.

A comprehensive comparison between the BPS and a traditional payment system operated by a profit-maximizing firm requires consideration of multiple attributes, many of which are outside the scope of this paper's analysis. As opposed to traditional systems, the BPS does not require trust in any entity. On the other hand, the BPS cannot provide some services: for instance, transactions cannot be reversed in case of error or fraud, and users who lose the credentials to their accounts cannot retrieve their balances.

We think of the BPS as a blueprint showing the feasibility of a decentralized design. The BPS demonstrates the power of competition and free entry of service providers within a platform. Future work is likely to improve upon these insights and apply them in other domains.

Since service provision requires resource expenditure, the operation of a decentralized platform necessitates a means to transfer value from users to service providers. The BPS allows such value transfers under the assumption that balances within its system (denominated in the system's native coin, bitcoin) are valuable. Determination of this value is left for future work.

Another feature that sets Bitcoin apart is that a protocol, rather than a managing organization, runs Bitcoin. Unlike a managing organization, a protocol lacks an easily

workable mechanism to change prices, offerings, and rules, implying the stability of these attributes. Such stability can be considered an asset or a liability of the system.

The blockchain protocol presents a novel economic design that would merit an economist's attention and scrutiny even if it had not been functional. Currently, the BPS handles daily transactions worth several billion dollars in aggregate. It can serve as a compelling proof of concept that should further encourage economists to study this marvelous structure and its future descendants.

References

Abadi, J. & Brunnermeier, M. (2018), Blockchain economics. NBER Working Paper No. 25407.

Arnosti, N. & Weinberg, S. M. (2018), Bitcoin: A natural oligopoly, *in* '10th Innovations in Theoretical Computer Science Conference (ITCS 2019)', Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Athey, S., Parashkevov, I., Sarukkai, V. & Xia, J. (2016), Bitcoin pricing, adoption, and usage: Theory and evidence. Stanford University Graduate School of Business Research Paper No. 16-42.

Babaioff, M., Dobzinski, S., Oren, S. & Zohar, A. (2012), On Bitcoin and red balloons, *in* 'Proceedings of the 13th ACM conference on electronic commerce', ACM, pp. 56–73.

Basu, S., Easley, D., O'Hara, M. & Sirer, E. (2019), 'Towards a functional fee market for cryptocurrencies', *arXiv preprint arXiv:1901.06830* .

Bentov, I., Pass, R. & Shi, E. (2016), 'Snow white: Provably secure proofs of stake.', *IACR Cryptol. ePrint Arch.* **2016**, 919.

Biais, B., Bisiere, C., Bouvard, M. & Casamatta, C. (2018), The blockchain folk theorem. Swiss Finance Institute Research Paper No. 17-75.

Budish, E. (2018), The economic limits of Bitcoin and the blockchain. NBER Working Paper No. 24717.

Buterin, V. (2018), Blockchain resource pricing.

Carlsten, M., Kalodner, H., Weinberg, S. M. & Narayanan, A. (2016), On the instability of bitcoin without the block reward, *in* 'Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security', ACM, pp. 154–167.

Catalini, C. & Gans, J. S. (2020), 'Some simple economics of the blockchain', *Communications of the ACM* **63**(7), 80–90.

Chiu, J. & Koeppl, T. (2017), The economics of cryptocurrencies - Bitcoin and beyond. Working paper.

Cong, L. W., He, Z. & Li, J. (2018), Decentralized mining in centralized pools. George Mason University School of Business Research Paper No. 18-9.

Cong, L. W., Li, Y. & Wang, N. (2018), Tokenomics: Dynamic adoption and valuation. Columbia Business School Research Paper No. 18-46.

Croman, K., Decker, C., Eyal, I., Gencer, A. E., Juels, A., Kosba, A., Miller, A., Saxena, P., Shi, E. & Gün, E. (2016), On scaling decentralized blockchains, *in* 'Proc. 3rd Workshop on Bitcoin and Blockchain Research'.

Easley, D., O'hara, M. & Basu, S. (2017), From mining to markets: The evolution of bitcoin transaction fees. Working paper.

Edelman, B. & Wright, J. (2015), 'Price coherence and excessive intermediation', *The Quarterly Journal of Economics* **130**(3), 1283–1328.

Evans, D. S. & Schmalensee, R. (2005), *Paying with plastic: the digital revolution in buying and borrowing*, Mit Press.

Eyal, I., Gencer, A. E., Sirer, E. G. & Van Renesse, R. (2016), Bitcoin-ng: A scalable blockchain protocol, *in* '13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)', pp. 45–59.

Eyal, I. & Sirer, E. G. (2014), Majority is not enough: Bitcoin mining is vulnerable, *in* 'International Conference on Financial Cryptography and Data Security', Springer, pp. 436–454.

Felten, E. (2013), 'Basic economics of Bitcoin mining'.
URL: *https://freedom-to-tinker.com/2013/02/05/basic-economics-of-bitcoin-mining/*

Gandal, N. & Halaburda, H. (2014), Competition in the cryptocurrency market. CEPR Discussion Paper No. DP10157.

Gans, J. S. & Halaburda, H. (2015), Some economics of private digital currency, *in* 'Economic Analysis of the Digital Economy', University of Chicago Press, pp. 257–276.

Gilad, Y., Hemo, R., Micali, S., Vlachos, G. & Zeldovich, N. (2017), Algorand: Scaling byzantine agreements for cryptocurrencies, *in* 'Proceedings of the 26th Symposium on Operating Systems Principles', pp. 51–68.

Glazer, A. & Hassin, R. (1986), 'Stable priority purchasing in queues', *Operations Research Letters* **4**(6), 285–288.

Grossman, S. J. & Hart, O. D. (1986), 'The costs and benefits of ownership: A theory of vertical and lateral integration', *Journal of Political Economy* **94**(4), 691–719.

Halaburda, H. & Sarvary, M. (2016), 'Beyond bitcoin', *The Economics of Digital Currencies* .

Hassin, R. (1995), 'Decentralized regulation of a queue', *Management Science* **41**(1), 163–173.

Hassin, R. (2016), *Rational queueing*, CRC press.

Hassin, R. & Haviv, M. (2003), *To queue or not to queue: Equilibrium behavior in queueing systems*, Vol. 59, Springer Science & Business Media.

Hayashi, F. & Maniff, J. L. (2019), 'Public authority involvement in payment card markets: Various countries–august 2019 update', *Federal Reserve Bank of Kansas City* .

Herkenhoff, K. F. & Raveendranathan, G. (2020), Who bears the welfare costs of monopoly? the case of the credit card industry, Technical report, National Bureau of Economic Research.

Huberman, G., Leshno, J. D. & Moallemi, C. (2019), An economist's perspective on the bitcoin payment system, *in* 'AEA Papers and Proceedings', Vol. 109, pp. 93–96.

Kittsteiner, T. & Moldovanu, B. (2005), 'Priority auctions and queue disciplines that depend on processing time', *Management Science* **51**(2), 236–248.

Kleinrock, L. (1975), *Queueing Systems. Volume 1: Theory*, Wiley-Interscience.

Kroll, J. A., Davey, I. C. & Felten, E. W. (2013), The economics of bitcoin mining, or bitcoin in the presence of adversaries, *in* 'The Twelfth Workshop on the Economics of Information Security (WEIS 2013)'.

Lavi, R., Sattath, O. & Zohar, A. (2017), 'Redesigning bitcoin's fee market', *arXiv preprint arXiv:1709.08881* .

Lui, F. T. (1985), 'An equilibrium queuing model of bribery', *Journal of Political Economy* **93**(4), 760–781.

Makarov, I. & Schoar, A. (2018), Trading and arbitrage in cryptocurrency markets. Working paper.

McKinsey & Company (2019), 'Global payments report 2019: Amid sustained growth, accelerating challenges demand bold actions'.
URL: *https://www.mckinsey.com/ /media/mckinsey/industries/financial services/our insights/tracking the sources of robust payments growth mckinsey global payments map/global-payments-report-2019-amid-sustained-growth-vf.ashx*

Morningstar (2019), 'Visa inc class a analysis'. Economic Moat by Brett Horn, Updated Dec 17, 2019.

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

Narayanan, A., Bonneau, J., Felten, E., Miller, A. & Goldfeder, S. (2016), *Bitcoin and cryptocurrency technologies*, Princeton University Press.

Olver, F. J. W., Lozier, D. W., Boisvert, R. F. & Clark, C. W., eds (2010), *NIST Handbook of Mathematical Functions*, Cambridge University Press.

Pagnotta, E. & Buraschi, A. (2018), An equilibrium valuation of bitcoin and decentralized network assets. Working paper.

Pass, R., Seeman, L. & Shelat, A. (2017), Analysis of the blockchain protocol in asynchronous networks, *in* 'Annual International Conference on the Theory and Applications of Cryptographic Techniques', Springer, pp. 643–673.

Prat, J. & Walter, B. (2018), An equilibrium model of the market for bitcoin mining. CESifo Working Paper Series No. 6865.

Ron, D. & Shamir, A. (2013), Quantitative analysis of the full bitcoin transaction graph, *in* 'International Conference on Financial Cryptography and Data Security', Springer, pp. 6–24.

Rosenbaum, A., Baughman, G., Manuszak, M. D., Stewart, K., Hayashi, F. & Stavins, J. (2017), 'Faster payments: market structure and policy considerations', *Federal Reserve Bank of Kansas City Working Paper No. RWP* pp. 17–14.

Sapirshtein, A., Sompolinsky, Y. & Zohar, A. (2016), Optimal selfish mining strategies in bitcoin, *in* 'International Conference on Financial Cryptography and Data Security', Springer, pp. 515–532.

Schilling, L. & Uhlig, H. (2018), Some simple bitcoin economics. NBER Working Paper No. 24438.

Sockin, M. & Xiong, W. (2018), A model of cryptocurrencies. Working paper.

Sompolinsky, Y. & Zohar, A. (2015), Secure high-rate transaction processing in bitcoin, *in* 'International Conference on Financial Cryptography and Data Security', Springer, pp. 507–527.

Tanenbaum, A. S. & Van Steen, M. (2007), *Distributed systems: principles and paradigms*, Prentice-Hall.

United States Cong. House Committee on Financial Services Task Force on Financial Technology (2020), 'Is cash still king? reviewing the rise of mobile payments', 116th Cong. 2nd sess. Testimony of Aaron Klein, Fellow, Economic Studies, Brookings Institution.

Wright, J. (2012), 'Why payment card fees are biased against retailers', *The RAND Journal of Economics* **43**(4), 761–780.

Yao, A. C.-C. (2018), 'An incentive analysis of some bitcoin fee design', *arXiv preprint arXiv:1811.02351* .

Yermack, D. (2015), Is bitcoin a real currency? an economic appraisal, *in* 'Handbook of Digital Currency', Elsevier, pp. 31–43.

Zohar, A. (2015), 'Bitcoin: under the hood', *Communications of the ACM* **58**(9), 104–113.

A A Brief Description of the Bitcoin Payment System

This appendix provides a simplified explanation of the permissionless blockchain protocol that underlies the Bitcoin Payment System and is the basis of many other cryptocurrencies. The description focuses on the economic elements.[28] In order to describe what the Bitcoin system does, it is useful to first explain what is needed for a payment system, such as PayPal or FedWire, or the maintenance of electronic balances in a modern bank.

An electronic payment system functions as a record (or a ledger) of accounts. Each account is associated with a user and his balance. It allows users to check their balances, and it allows a user to debit his balance and credit the debited amount to another account. Only an account owner can debit the account. Balances do not change without a legal transfer, i.e., a transfer that conforms to the system's stated rules.

One simple implementation is just a spreadsheet (or another bookkeeping device) that only a trusted authority can modify. Allowing multiple computers to maintain and update the ledger requires a more elaborate structure. This distributed ledger structure requires synchronization across the servers but is, in principle, more robust than a single server system. Maintaining consensus in a distributed computer system has been known to be straightforward as long as the computers are trusted (see Tanenbaum & Van Steen (2007)).

The Bitcoin system is designed for an environment which lacks a trusted authority. Therefore, its ledger must be maintained and updated by a collection of computer servers, called miners, none of which are trusted. They are assumed to be selfish, i.e., to respond to incentives in a profit-maximizing way. Moreover, they offer or withdraw their services according to profit opportunities they perceive.

Although legal transactions are processed by untrusted miners, the system as a whole is secure, i.e., it processes all legal transactions and no other transactions. The collection of miners jointly holds a single ledger, meaning that there must be consensus among miners about current balances. Moreover, consensus must be maintained as balances change.

[28] In particular, this description omits discussion of potential attacks on the system. For further details and an explanation of the cryptographic elements of the system please refer to Narayanan et al. (2016).

Bitcoin's ledger is a public database called blockchain, which can be verified by third parties through cryptography. The system arranges for the miners to be compensated for their services in such a way that when each of them maximizes his profit and believes that other miners similarly maximize their profits, the system has the properties sketched above.

Initially, all balances are at zero. Over time, the protocol mints new coins which it adds to the balances of successful miners. The system holds the record of all balance changes. The manifestation of a transaction is a message which a sending account transmits to all the miners. It states the sending account, receiving account, amount transferred, transaction fee, and cryptographic signature by the sending account. A transaction is processed by adding the appropriate message to the end of the ledger. The cryptographic signature allows any third party to verify that the transaction was indeed authorized by the holder of the sending account. Since the ledger is public, any third party can verify that the sender indeed held a balance sufficient for the transfer.

The public ledger is saved in the distributed blockchain format, in which the transaction data is partitioned into a sequence of blocks. These blocks are periodic updates to the ledger. Notably, the ledger does not update instantly following the appearance of a new transaction. Rather, it updates on average every ten minutes with a block summarizing a subset of the recent pending transactions which hadn't been included in a previous block. Remaining unprocessed transactions wait to be processed in future blocks. As of July 2017, the maximal block size is 1MB.[29]

New transactions are processed when they are included in a block that is added to the ledger, which happens as follows. Each miner holds a copy of the current ledger i.e., all previous blocks. All transaction requests are broadcast to all miners. The set of pending transactions that reaches each miner may vary slightly across miners due to network imperfections, rendering non-trivial the choice of a universally agreed upon record of transactions. To ensure that Bitcoin maintains a unique record of transactions, a single miner is selected to add a block of transactions to the ledger. Since there is no trusted authority to make the selection, a tournament is used to randomly select a winning miner. To participate in the tournament, miners exert effort[30] (known as proof of work) that is useful only for generating a verifiable random

[29] As of July 2017, the protocol limits each block to 1MB of data to ensure each block can be transmitted promptly throughout the network. This limits each block to no more than approximately 2,000 transactions, as the average transaction uses 0.5KB of data (Zohar 2015).

[30] The tournament selects a random winner without the need of a trusted authority through use of

selection of a miner without the need of a trusted randomization device.

Periodically (currently approximately every 10 minutes), the tournament randomly selects one miner as the winner, assigning his block as the next in the chain, thereby making that block a mined block. The mined block is transmitted to all the other miners, who verify the legality of that block and vet all transactions included in the block. Miners add a newly mined legal block to their copy of the ledger and proceed to add new blocks on top of it. Miners ignore mined blocks that are not legal.

The tournament-winning miner is paid a reward when he mines a new block but can withdraw his reward only after newer blocks augment the chain on top of his block. Other miners will build on top of his block only if they consider it legal. Hence, the incentive is to assemble and create legal blocks. Consensus forms on a ledger that includes the new block. The process continues in the same manner for the following ten minutes (on average) and so on.[31]

The miner that created a block is paid from two sources. One consists of newly minted coins, the exact number of which is protocol-determined and is decreasing with time. (Crediting successful miners with newly minted coins moves the system early on from having zero balances to having positive ones.) The second consists of the fees offered by the transactions in the mined block. This second source is the focus of the paper.

This system will have the following desired properties. All miners are synchronized

a hash function. The hash function is a deterministic one-way function that produces a hash value, interpreted as a pseudo-random real number between 0 and 1. A block is said to be a winning block if it is a legal block and its hash value is below a target value. A legal block contains, in addition to transaction data, an unrestricted "nonce" field for which the miner can input any numerical value. The cryptographic properties of the hash function imply that finding such a block requires a brute-force search, iterating over numerical values for the nonce and computing the hash value for each of them. Roughly speaking, each attempt for a value of the nonce generates an independent random draw of a hash value, distributed uniformly between 0 and 1.

To participate in the tournament, miners assemble their blocks and use their computational power to iterate over values of the nonce. Each attempt for a nonce value has an independent probability of generating a winning block, with probability equal to the target value. Because the target value is very small, a miner's chance to win the tournament within a time period is proportional to the number of nonce values attempted within the period. A miner with a winning block is said to "mine the block", and the winning block can be verified by any third party by recomputing the hash.

The target value adjusts over time so that a block is mined every 10 minutes (on average). For example, if the overall computational power of miners doubles, then the target value is halved and twice as many attempts (on average) are required to find a winning block.

[31] There is a small probability that two or even more blocks are vying to be accepted as the newest block. This situation is called a fork. Bitcoin's convention calls for newer blocks to be built on top of the longest chain. This convention resolves forks. Eyal & Sirer (2014) analyze strategic issues between miners.

to hold the same ledger of processed transactions. No single miner controls the system, because every 10 minutes the ability to process transactions is given to a randomly chosen miner. Balances change only with a legal transaction because any transaction that is added is vetted by other miners to be valid, and transactions cannot be deleted from the ledger.

Online Appendix for "Monopoly without a Monopolist: An Economic Analysis of the Bitcoin Payment System"

B Endogenous Entry

The analysis in Section 4.2 assumed that the reward R_L, R_H is sufficiently high for all users receive positive net reward. Lemma 2 shows that all users receive positive net reward if

$$\int_0^{\bar{c}} \mu^{-1} W_K\left(\rho \bar{F}\left(c\right)\right) dc \leq R_L.$$

This section extends the analysis to values of R for which the inequality is not satisfied. For simplicity, assume that $R_H = R_L = R \geq 0$ and let $c^* \in [0, \bar{c}]$ be the unique solution to

$$\int_0^{c^*} \mu^{-1} W_K\left(\rho\left(\bar{F}\left(c\right) - \bar{F}\left(c^*\right)\right)\right) dc = R.$$

It is straightforward to verify that, in equilibrium, users with delay cost $c_i \notin [0, c^*]$ opt out of the system, and that a user with delay cost $c_i \in [0, c^*]$ chooses a transaction fee

$$b\left(c_i\right) = \rho \int_0^{c_i} f\left(c\right) \cdot c \cdot \mu^{-1} W_K'\left(\rho\left(\bar{F}\left(c\right) - \bar{F}\left(c^*\right)\right)\right) dc.$$

The system's revenue and total delay cost are given by

$$\text{Rev}_K(\rho|R) = K\rho^2 \int_0^{c^*} cf(c)\left(\bar{F}(c) - \bar{F}\left(c^*\right)\right) W_K'\left(\rho\left(\bar{F}(c) - \bar{F}\left(c^*\right)\right)\right)\, dc,$$

$$\text{DelayCost}_K(\rho|R) = K\rho \int_0^{c^*} cf(c) W_K\left(\rho\left(\bar{F}(c) - \bar{F}\left(c^*\right)\right)\right)\, dc.$$

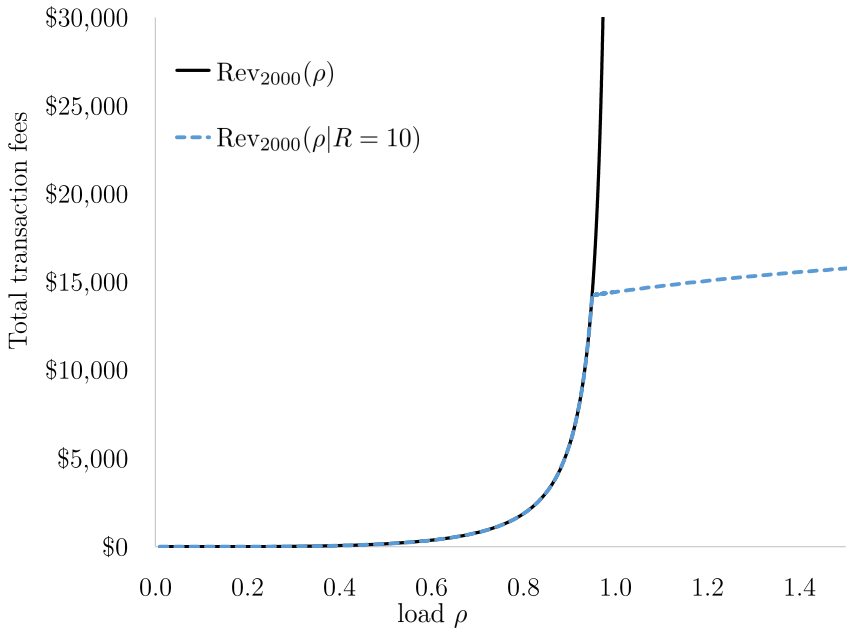

Figure 5: Total revenue per block as a function of ρ when $c \sim U[0,1]$. The curve $\text{Rev}_{2000}(\rho)$ shows total revenue from transaction fees when WTP is sufficiently high so that the participation constraint does not bind for any user, and it is only defined for $0 \leq \rho < 1$. The curve $\text{Rev}_{2000}(\rho|R = 10)$ shows total revenue from transaction fees when all users have WTP equal to 10 USD, and it is defined for any $\rho \geq 0$.

The infrastructure available to the system is given by the number of miners

$$N = \frac{\text{Rev}_K(\rho|R)}{c_m}.$$

Note that these expressions coincide with their counterparts in Section 4.2 when $c^* = \bar{c}$. Figure 5 provides an illustration of these results.

C Endogenous Willingness To Pay

The model allows us to solve for miner and user behavior given exogenously specified user WTP. The analysis assumed (Assumption 1) that users consider the system to be a reliable means of sending transactions and, in particular, that the system has sufficient mining resources for its operation and security. This section builds up on Appendix B to extend the analysis and allow for endogenous determination of the

user's WTP R given the system's aggregate computational power N.[32] Analogous extensions can extend the model to allow for an endogenous exchange rate e.

For tractability, we assume all agents have the same WTP $R = \psi(N)$, which is a function of the system's aggregate computational power N. Users endogenously choose whether to participate as a function of their perceived WTP $\psi(R)$. In particular, ψ can capture that users believe the system is unreliable with computational power N' by $\psi(N') < 0$. Negative WTP implies that users choose to not participate.

We change the game described in Section (2) to allow for endogenous WTP by requiring that agents have correct beliefs on N and that their WTP is $R = \psi(N)$. That is, equilibrium R, N must satisfy

$$\begin{aligned} R &= \psi(N) \\ N &= \frac{\text{Rev}(R) + e \cdot S}{c_m}. \end{aligned}$$

Appendix B derives $\text{Rev}(R)$ for any possible R. If $R \leq 0$, then none of the users participate and $\text{Rev}(R) = 0$. If $R \geq 0$ we have that

$$\text{Rev}(R) = \text{Rev}_K(\rho|R) = K\rho^2 \int_0^{c^*} cf(c)\left(\bar{F}(c) - \bar{F}(c^*)\right) W_K'\left(\rho\left(\bar{F}(c) - \bar{F}(c^*)\right)\right) dc,$$

where c^* is the unique solution to

$$\int_0^{c^*} \mu^{-1} W_K\left(\rho\left(\bar{F}(c) - \bar{F}(c^*)\right)\right) dc = R,$$

if $R \leq \bar{R} = \mu^{-1}\int_0^{\bar{c}} W_K\left(\rho\bar{F}(c)\right) dc$, and $c^* = \bar{c}$ if $R \geq \bar{R}$.

Let $\text{Rev}\left(\bar{R}\right) = \max_R \text{Rev}(R)$ be the maximal total revenue from transaction fees, which is achieved when all users participate. Let $\bar{N} = \left(\text{Rev}\left(\bar{R}\right) + e \cdot S\right)/c_m$ denote the corresponding aggregate computational power. The following corollaries are immediate.

Corollary 4. *If $\psi\left(e \cdot S/c_m\right) < 0$, that is, the system is not reliable if there is zero revenue from transaction fees, then there is an equilibrium in which none of the users participate.*

[32] For some considerations (e.g., security of the system), the users WTP should depend on the total payment to miners in USD, rather than the system's total computational power N. The derivation below allows for either interpretation because, in equilibrium, the total payment to miners is $c_m N$, which is a constant multiple of the system's total computational power N.

Corollary 5. *If $\psi\left(\bar{N}\right) \geq \bar{R}$, the equilibrium analyzed in Section 4 is also an equilibrium under endogenous WTP.*

It is natural to consider users that deem the system to be unreliable when the computational power is below some minimal required N_0, that is, $\psi(x) < 0$ for any $x \leq N_0$. Currently, the majority of miner compensation comes from newly minted coins $e \cdot S$. This amount provides sufficient computational power for the reliability of the system, that is, ${}^{e \cdot S}/_{c_m} > N_0$. If newly minted coins by themselves are insufficient (because, e.g., the protocol mints less coin), then the system is susceptible to failure when congestion is low and revenue from transaction fees is insufficient.

Corollary 6. *Suppose that $\psi(x) < 0$ for any $x \leq N_0$, and that ${}^{e \cdot S}/_{c_m} < N_0$. Then there exists ρ_0 such that for any $\rho < \rho_0$ there is a unique equilibrium in which none of the users participate. The proof follows from Corollary (3), which shows that the maximal total revenue from transaction fees $\mathrm{Rev}\left(\bar{R}\right)$ is increasing in ρ and is equal to zero when $\rho = 0$.*

The following example provides simplified expressions under additional assumptions.

Example. Suppose that $\mu = 1, K = 1$, and $c \sim U[0,1]$. For these parameters, we have that $\bar{R} = 1/(1-\rho)$, and the equation that defines c^* simplifies to

$$R = \int_0^{c^*} \mu^{-1} W_K\left(\rho\left(\bar{F}(c) - \bar{F}(c^*)\right)\right) dc = \frac{c^*}{1 - c^*\rho}.$$

Therefore, we have that for $0 \leq R \leq \bar{R}$

$$c^* = \frac{R}{1 + \rho R},$$

and the implied revenue from transaction fees is

$$\begin{aligned}
\mathrm{Rev}_K(\rho|R) &= K\rho^2 \int_0^{c^*} c f(c) \left(\bar{F}(c) - \bar{F}(c^*)\right) W_K'\left(\rho\left(\bar{F}(c) - \bar{F}(c^*)\right)\right) dc \\
&= \frac{c^*(2 - c^*\rho)}{1 - c^*\rho} + \frac{2\log(1 - c^*\rho)}{\rho} \\
&= \frac{R(2 + \rho R)}{1 + \rho R} - \frac{2\log(1 + \rho R)}{\rho}.
\end{aligned}$$

Plugging these expressions into the endogenous WTP conditions, we get that WTP R can arise in equilibrium only if: (i) $R = \psi\left(\bar{N}\right) \geq \bar{R}$, or (ii) $R = \psi\left(0\right) \leq 0$, or (iii) $0 \leq R \leq \bar{R}$ and

$$R = \psi\left(\frac{\frac{R(2+\rho R)}{1+\rho R} - \frac{2\log(1+\rho R)}{\rho} + e \cdot S}{c_m}\right).$$

D Attributes of Transaction Fees

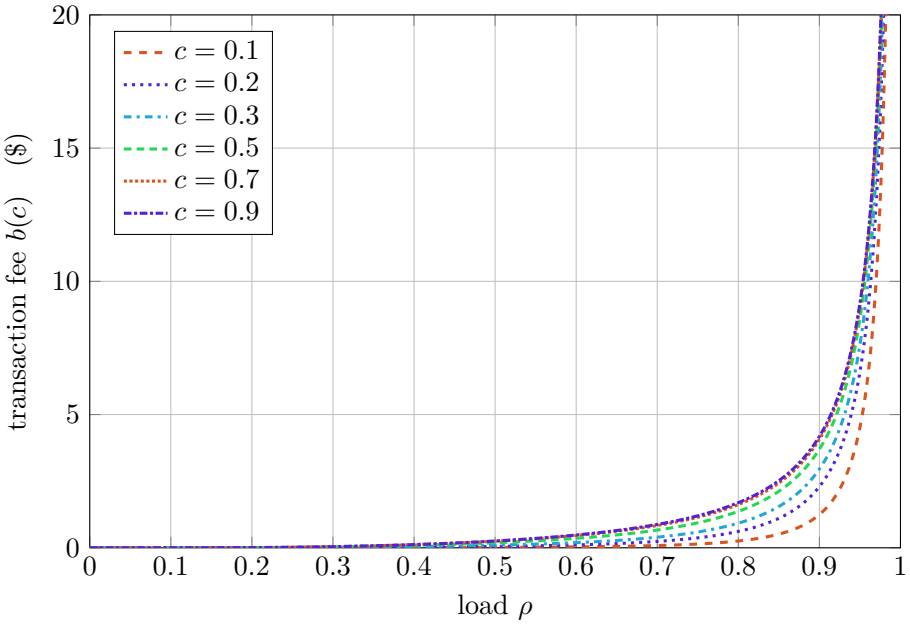

Figure 6: The dependence of equilibrium transaction fees on congestion ρ for fixed user's delay cost c. Block size is taken to be $K = 2,000$, block arrival rate $\mu = 1$, and delay costs are distributed according to $c \sim U[0,1]$.

Figure 6 and 7 illustrate how transaction fees depend on the user's delay cost c and the overall congestion ρ. Both figures display equilibrium fees when c is distributed uniformly over $[0,1]$, the block size is $K = 2,000$, and $\mu = 1$. Figure 6 shows how the transaction fees chosen by users in equilibrium vary with the overall system congestion ρ. Transaction fees are very small when the system is not congested but can be arbitrarily high as ρ approaches 1.

Figure 7 shows that the transaction fees increase with the user's delay cost but do not vary much among users with high delay cost. An intuitive explanation is that such users that offer high fees the probability that a transaction is processed in the next block is high and does not vary much with further fee increases. Because all

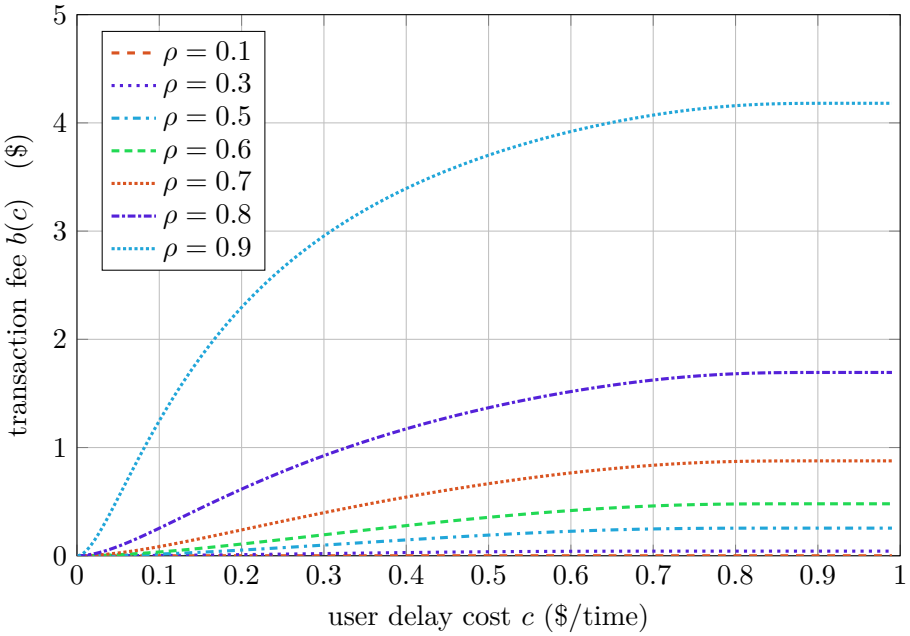

Figure 7: The dependence of equilibrium transaction fees on the user's delay cost c for fixed congestion ρ. Block size is taken to be $K = 2,000$, block arrival rate $\mu = 1$, and delay costs are distributed according to $c \sim U[0,1]$.

users within the same block are treated equally, there is little competition for priority among users with high delay costs.

To form a complementary interpretation, observe that the expected wait for a user with cost c_i is $W_K(\hat{\rho})$ with $\hat{\rho} \triangleq \rho \bar{F}(c_i) < \bar{F}(c_i)$. When $\hat{\rho}$ is small, the expected wait $W_K(\hat{\rho})$ is not very sensitive to variations in $\hat{\rho}$, and therefore users with a high c_i are only slightly harmed when someone gains priority over them. However, $W_K(\hat{\rho})$ can be very sensitive to changes in $\hat{\rho}$ when $\hat{\rho}$ is close to 1, and thus the externality on users with low delay cost can be substantial. All users with sufficiently high delay cost, for example $c_i > 0.7$, impose the same externality to other users with delay costs $c_j \in [0, 0.7]$ plus a relatively small externality to other users with delay costs $c_j \in (0.7, c_i)$.

E Additional Figures

This appendix provides additional plots showing the goodness of approximation in Theorem 5, illustrating the delay function $W_K(\rho)$, and showing that different waiting cost distribution yield similar results. Table 1 presents a regression analysis to

complement Figure 1.

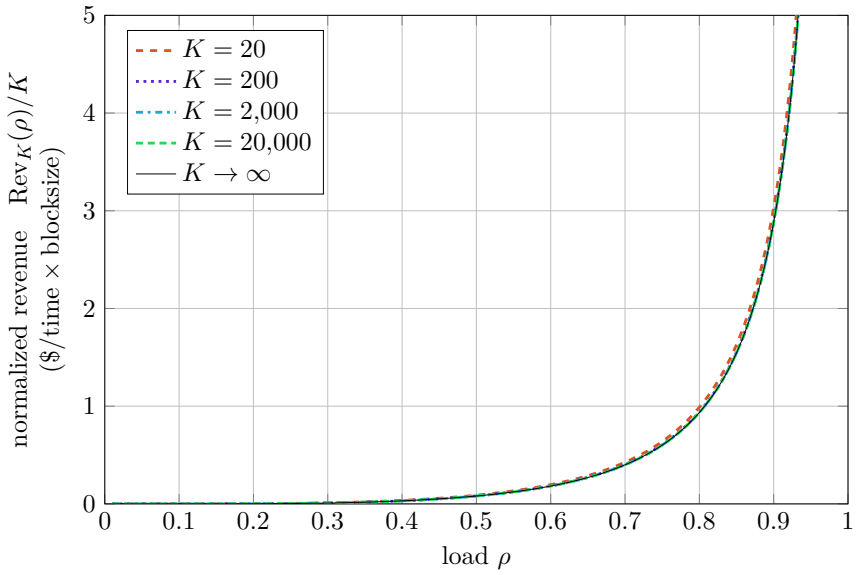

Figure 8: Normalized revenue $\mathrm{Rev}_K(\rho)/K$ *when* $c \sim U[0,1]$ *and* $K \in \{20, 200, 2000, 20000\}$*, compared to the limiting values obtained from the approximation using* $W_\infty(\cdot)$*. The plot may appear to have only one line because all lines overlap.*

OLS Regression Results

Dep. Variable:	FeeTotUSD	R-squared:	0.802
Model:	OLS	Adj. R-squared:	0.801
Method:	Least Squares	F-statistic:	1840.
Date:	Thu, 10 Sep 2020	Prob (F-statistic):	0.00
Time:	18:19:30	Log-Likelihood:	-28760.
No. Observations:	2283	AIC:	5.753e+04
Df Residuals:	2277	BIC:	5.757e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3214.2896	2759.495	1.165	0.244	-2197.098	8625.677
predictedRev	11.4300	1.194	9.575	0.000	9.089	13.771
BlkSizeMeanByte	-0.2900	0.009	-32.948	0.000	-0.307	-0.273
PriceUSD	209.1827	6.613	31.631	0.000	196.214	222.151
HashRate	0.0881	0.004	21.462	0.000	0.080	0.096
ROI30d	5.2354	20.068	0.261	0.794	-34.118	44.589

Omnibus:	1500.842	Durbin-Watson:	0.147
Prob(Omnibus):	0.000	Jarque-Bera (JB):	52550.476
Skew:	2.585	Prob(JB):	0.00
Kurtosis:	25.928	Cond. No.	2.45e+06

Table 1: Regression of total daily transaction fees in USD from April 1, 2011 to June 30, 2017 on predicted transaction fees (see Section 6.2), daily average block size, the bitcoin to USD exchange rate, Hashrate, and the 30 day change in the bitcoin to USD exchange rate. Data source: https://coinmetrics.io/community-network-data/.

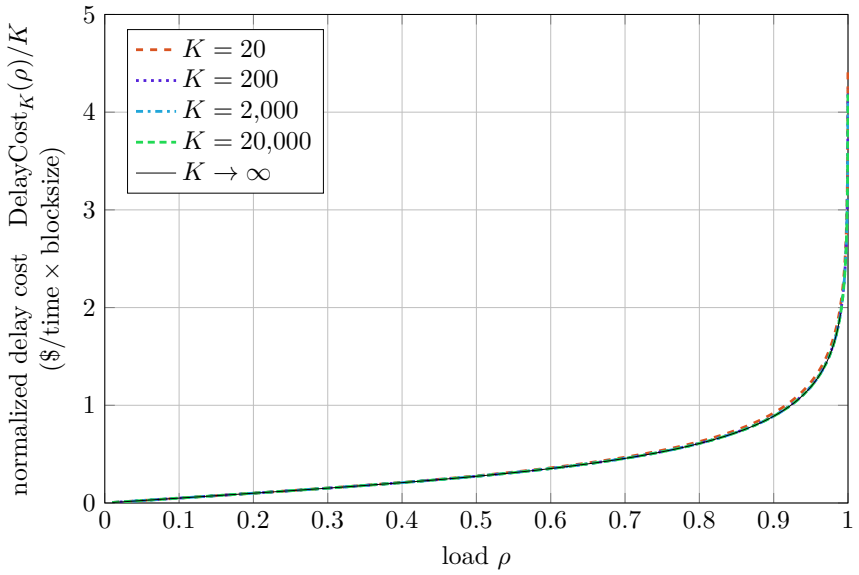

Figure 9: Normalized revenue $\mathrm{Rev}_K(\rho)/K$ when $c \sim U[0,1]$ and $K \in \{20, 200, 2000, 20000\}$, compared to the limiting values obtained from the approximation using $W_\infty(\cdot)$. The plot may appear to have only one line because all lines overlap.

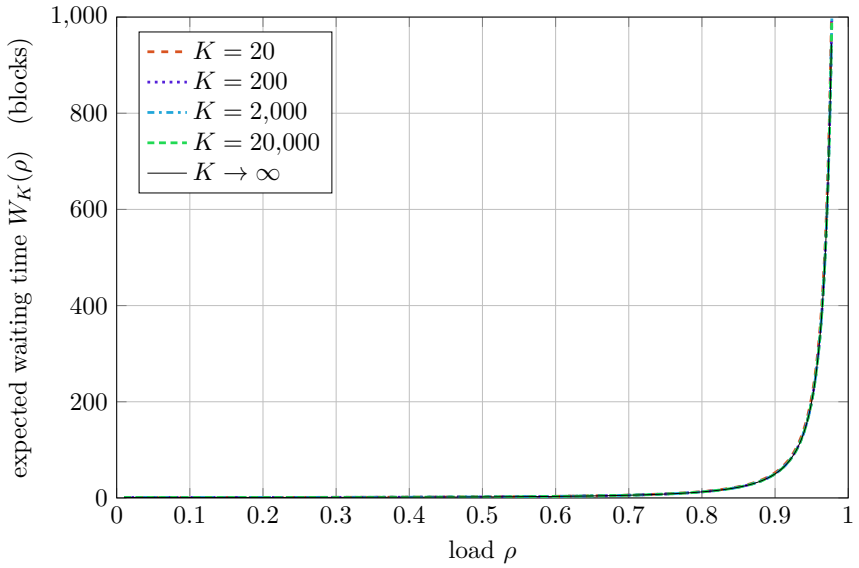

Figure 10: The expected delay in blocks $W_K(\rho)$ of the lowest priority transaction given $\rho = \lambda/\mu K$ and $K \in \{20, 200, 2000, 20000\}$.

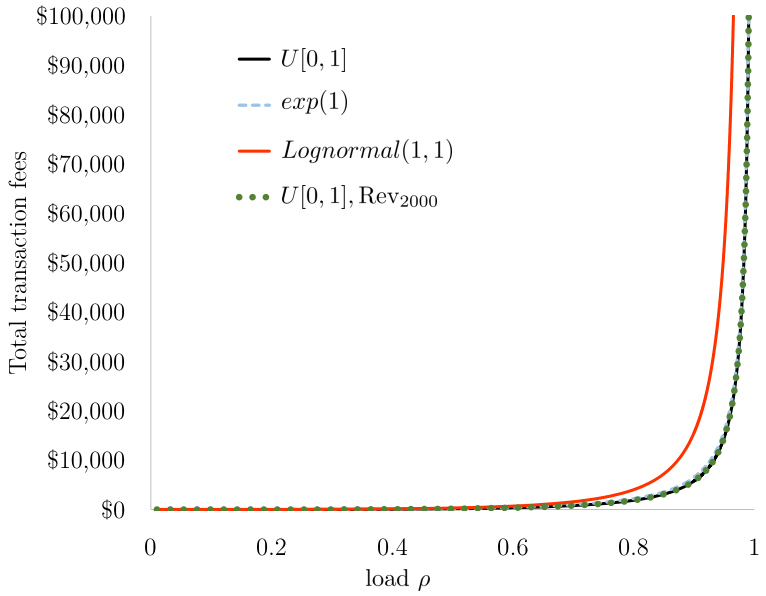

Figure 11: Revenue for $K = 2000$ and waiting costs c distributed (i) uniformly on $[0, 1]$, (ii) as an exponential with mean 1, (iii) as a Log-normal with mean and variance equal to 1. All were calculated using the asymptotic approximation. The plot also shows $\text{Rev}_{2000}(\rho)$ for the uniform distribution in a dotted line that overlaps the asymptotic approximation.

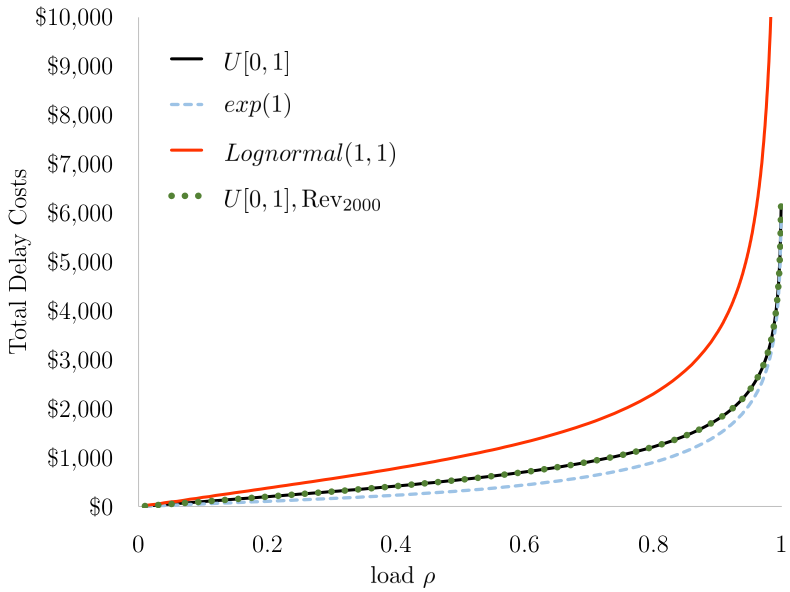

Figure 12: Delay costs for $K = 2000$ and waiting costs c distributed (i) uniformly on $[0, 1]$, (ii) as an exponential with mean 1 (iii) as a Log-normal with mean and variance equal to 1. All were calculated using the asymptotic approximation. The plot also shows $\mathrm{Rev}_{2000}(\rho)$ for the uniform distribution in a dotted line that overlaps the asymptotic approximation.

F Proofs

F.1 Queueing Analysis

In this section, we will establish the main queueing result, which is the waiting time expression of Lemma 1. We begin with a standard result from the analysis of bulk service systems (e.g., Section 4.6, Kleinrock 1975):

Lemma A1. *Consider a queue system consisting of a single queue, with arrivals according to a Poisson process of rate $\lambda \geq 0$ and bulk service in batches of size up to $K \geq 1$ with service times exponentially distributed with parameter $\mu > 0$. Suppose that the load $\rho \triangleq \lambda/(\mu K) \geq 0$ satisfies $\rho < 1$. Then, the queueing system is stable, and the steady-state queue length Q has the geometric distribution*

$$\mathsf{P}(Q = \ell) = (1 - z_0)z_0^{\ell}, \quad \ell = 0, 1, \ldots.$$

Here, the parameter of the geometric distribution $z_0 \triangleq z_0(\rho, K)$ is given as unique

solution of the polynomial equation

$$z^{K+1} - (K\rho + 1)z + K\rho = 0,$$

in the interval $[0, 1)$.

Lemma A1 and Little's Law are used to prove the following, which implies Lemma 1:

Lemma A2. *Consider a transaction, and let* $\hat{\lambda}$ *be the arrival rate of higher priority transactions (i.e., transaction that offer greater fees). The expected time until the transaction is processed is a function of the block size* K*, the block arrival rate* μ*, and the load parameter* $\hat{\rho} \triangleq \hat{\lambda}/\mu K \in [0, 1)$*, and is equal to*

$$\mu^{-1} W_K(\hat{\rho}) = \frac{1}{\mu} \frac{1}{(1 - z_0)\big(1 + K\hat{\rho} - (K+1)z_0^K\big)}.$$

Here, $z_0 \triangleq z_0(\hat{\rho}, K) \in [0, 1)$ *is the polynomial root defined in Lemma A1.*

The quantity $W_K(\hat{\rho}) \geq 1$ *is the expected waiting time measured in blocks. It satisfies*

$$W_K'(\hat{\rho}) > 0, \quad \forall\ \hat{\rho} \in (0, 1).$$

Finally, we have that

$$W_K(0) = 1; \qquad \lim_{\hat{\rho}\to 1} W_K(\hat{\rho}) = \infty; \qquad W_K'(0) = 0, \ \textit{if } K > 1; \qquad \lim_{\hat{\rho}\to 1} W_K'(\hat{\rho}) = \infty.$$

Proof. While this result can be established directly using a generating function argument, we will instead use a more intuitive approach based on Little's Law.

To start, consider a queueing system with arrival according to a Poisson process of rate $\hat{\lambda}$, exponential service time with parameter μ, and batch size K. Define $\bar{W}_K(\rho)$ to be the average waiting time of a user in this system measured in multiples of the mean service time μ^{-1}. Here, we highlight the dependence on the load $\hat{\rho} = \hat{\lambda}/\mu K$. Lemma A1 implies that the mean queue length is given by

$$\mathsf{E}[Q_K] = \frac{z_0(\hat{\rho}, K)}{1 - z_0(\hat{\rho}, K)}.$$

Applying Little's Law,

$$\frac{z_0(\hat{\rho}, K)}{1 - z_0(\hat{\rho}, K)} = \hat{\lambda} \frac{\bar{W}_K(\hat{\rho})}{\mu}. \tag{14}$$

Now, Little's Law (14) holds no matter what the service discipline. In particular, we can specialize to the case where users are given preemptive priority service, where each user is given a priority type drawn uniformly over the interval $[0, \hat{\rho}]$, and where service for users of lower numerical priority type preempts service for higher numerical priority type. Define $W_K(\rho)$ to be the expected waiting time (in multiples of the mean service time) for users with priority type $\rho \in [0, \hat{\rho}]$. Then,

$$\bar{W}_K(\hat{\rho}) = \frac{1}{\hat{\rho}} \int_0^{\hat{\rho}} W_K(\rho)\, d\rho.$$

Substituting into (14), we have that

$$\frac{z_0(\hat{\rho}, K)}{1 - z_0(\hat{\rho}, K)} = K \int_0^{\hat{\rho}} W_K(\rho)\, d\rho.$$

Differentiating with respect to $\hat{\rho}$ and simplifying, we have that

$$W_K(\hat{\rho}) = \frac{\partial_{\hat{\rho}} z_0(\hat{\rho}, K)}{K\left(1 - z_0(\hat{\rho}, K)\right)^2}. \tag{15}$$

In order to simplify this expression, we will use the implicit function theorem. Denote by $Q_K(z, \hat{\rho})$ the degree K polynomial in z defined by

$$z^{K+1} - (K\hat{\rho} + 1)z + K\hat{\rho} = \big(z_0(\hat{\rho}, K) - z\big) Q_K(z, \hat{\rho}), \quad \forall\ (z, \hat{\rho}) \in \mathbb{R} \times [0, 1). \tag{16}$$

This polynomial exists and is unique since $z_0 \triangleq z_0(\hat{\rho}, K)$ is a root of the degree $K+1$ polynomial on the left side. We apply the implicit function theorem and differentiate (16) with respect to $(z, \hat{\rho}) \in \mathbb{R} \times [0, 1)$ to obtain

$$(K+1)z^K - (K\hat{\rho} + 1) = -Q_K(z, \hat{\rho}) + \big(z_0(\hat{\rho}, K) - z\big)\partial_z Q_K(z, \hat{\rho}), \tag{17}$$

$$-Kz + K = \partial_{\hat{\rho}} z_0(\hat{\rho}, K) Q_K(z, \hat{\rho}) + \big(z_0(\hat{\rho}, K) - z\big)\partial_{\hat{\rho}} Q_K(z, \hat{\rho}). \tag{18}$$

Substituting $z = z_0(\hat{\rho}, K)$ into (17), we have that

$$Q_K(z_0, \hat{\rho}) = 1 + K\hat{\rho} - (K+1)z_0^K. \tag{19}$$

The same substitution into (18) yields that

$$\partial_{\hat{\rho}} z_0(\hat{\rho}, K) = K \frac{1 - z_0}{Q_K(z_0, \hat{\rho})} = K \frac{1 - z_0}{1 + K\hat{\rho} - (K+1) z_0^K}. \tag{20}$$

Substituting (19)–(20) into (15) yields the desired result that

$$W_K(\hat{\rho}) \triangleq \frac{1}{(1 - z_0)\big(1 + K\hat{\rho} - (K+1) z_0^K\big)}. \tag{21}$$

We will now show that $W_K'(\hat{\rho}) > 0$. Differentiating (21),

$$W_K'(\hat{\rho}) = \frac{\big(Q_K(z_0, \hat{\rho}) + K(K+1)(1 - z_0) z_0^{K-1}\big)\, \partial_{\hat{\rho}} z_0(\hat{\rho}, K) - K(1 - z_0)}{\big((1 - z_0) Q_K(z_0, \hat{\rho})\big)^2}$$

Substituting $z = z_0(\hat{\rho}, K)$ into (17), we have that

$$\partial_{\hat{\rho}} z_0(\hat{\rho}, K) = \frac{K(1 - z_0)}{Q_K(z_0, \hat{\rho})} = K(1 - z_0)^2 W_K(\hat{\rho}).$$

Then,

$$\begin{aligned} W_K'(\hat{\rho}) &= K \frac{\big(Q_K(z_0, \hat{\rho}) + K(K+1)(1 - z_0) z_0^{K-1}\big) - Q_K(z_0, \hat{\rho})}{(1 - z_0) Q_K(z_0, \hat{\rho})^3} \\ &= \frac{K^2 (K+1) z_0^{K-1}}{Q_K(z_0, \hat{\rho})^3} \\ &= K^2 (K+1) z_0^{K-1} (1 - z_0)^3 W_K(\hat{\rho})^3. \end{aligned} \tag{22}$$

Since the waiting time must be at least one block, $W_K(\hat{\rho}) \geq 1$. Since $z_0 < 1$ and, if $\hat{\rho} \in (0, 1)$, $z_0 \neq 0$ also, we have that $W_K'(\hat{\rho}) > 0$. Furthermore, since $z_0(0, K) = 0$, it is clear that

$$W_K(0) = 1, \qquad W_K'(0) = \begin{cases} 2 & \text{if } K = 1, \\ 0 & \text{if } K > 1. \end{cases}$$

Finally, we consider the asymptotic limits of $W_K(\cdot)$ and $W_K'(\cdot)$ as $\hat{\rho} \to 1$. Factoring the defining polynomial for $z_0 \in [0, 1)$, we have that

$$0 = z_0^{K+1} - (K\hat{\rho} + 1) z_0 + K\hat{\rho} = (1 - z_0) \left(K\hat{\rho} - \sum_{\ell=1}^{K} z_0^{\ell} \right).$$

Therefore, z_0 satisfies

$$\hat{\rho} = \frac{1}{K}\sum_{\ell=1}^{K} z_0^{\ell} \leq \frac{1}{K}\sum_{\ell=1}^{K} z_0 = z_0 < 1,$$

where the inequalities follow since $z_0 \in [0,1)$. Taking a limit as $\hat{\rho} \to 1$, clearly $z_0 \to 1$ and $Q_K(z_0, \hat{\rho}) \to 0$. Therefore, from (21), $W_K(\hat{\rho}) \to \infty$, and also from (22),

$$\lim_{\hat{\rho}\to 1} W_K'(\hat{\rho}) = \lim_{\hat{\rho}\to 1} \frac{K^2(K+1)z_0^{K-1}}{Q_K(z_0,\hat{\rho})^3} = \infty.$$

□

F.2 Equilibrium

Proof of Proposition 3: We consider agents equilibrium decisions conditional on being forced to participate. Let G denote the the cumulative distribution function of transaction fees in some equilibrium, and let $b(c_i)$ be a transaction fee chosen by agents with delay cost c_i. Consider a user i with delay cost c_i. The user chooses his transaction fee b to maximize his net reward

$$R_i - b - c_i \cdot W(b \mid G),$$

with $W(b \mid G)$ denoting the expected delay given transaction fee b and the CDF G. By Lemma 1, the expected delay is decreasing with b, and standard arguments (see Lui (1985), Hassin & Haviv (2003)) imply that $b(c_i)$ is increasing in c_i and $b(0) = 0$. Monotonicity of $b(\cdot)$ implies that $G(b(c)) = F(c)$. Therefore, we have that

$$\hat{\rho}(c_i) = \frac{\lambda \cdot (1 - G(b(c_i)))}{\mu K} = \rho \cdot \bar{F}(c_i),$$

and

$$\begin{aligned} W(b \mid G) &= \mu^{-1} W_K\left(\rho \cdot \bar{G}(b)\right) \\ &= \mu^{-1} W_K\left(\rho \cdot \bar{F}(c_i)\right). \end{aligned}$$

Each agent is bidding optimally if and only if

$$b(c_i) \in \arg\min_b \{c \cdot W(b \mid G) + b\}.$$

The first order condition implies

$$W'(b_i \mid G) = -\frac{1}{c_i}.$$

Plugging in $G'(b_i) = f(c_i)/b'(c_i)$, we have that

$$\mu^{-1} W_K'\left(\rho \cdot \bar{G}(b)\right) \cdot \left(-\rho f(c_i)/b'(c_i)\right) = -\frac{1}{c_i},$$

or

$$b'(c_i) = c_i \rho f(c_i) \mu^{-1} W_K'\left(\rho \bar{F}(c_i)\right).$$

Integration, together with the fact that $b(0) = 0$ yields

$$b(c_i) = \rho \int_0^{c_i} f(c) \cdot c \cdot \mu^{-1} W'\left(\rho \bar{F}(c)\right) dc.$$

Transaction fees coincide with the payments that result from selling priority in a VCG auction because of revenue equivalence. To directly see that $b(c_i)$ is the externality imposed by c_i, write the expected wait in terms of arrival rate of higher priority transactions as $\mu^{-1}\tilde{W}_K\left(\hat{\lambda}\right) \triangleq \mu^{-1} W_K\left(\hat{\lambda}/\mu K\right)$. The transaction sent by c_i affects the waiting time of transactions with lower priority that are sent by users with $0 \leq c < c_i$; higher priority transactions are not affected. Integration over all affected types implies that the externality imposed by a marginal increase in the volume of transaction from users with c_i is

$$\int_0^{c_i} \lambda f(c) \cdot c \cdot \mu^{-1} \tilde{W}_K'\left(\lambda \bar{F}(c)\right) dc = b(c_i).$$

Finally,

$$\begin{aligned} b(c_i) &= \rho \int_0^{c_i} c f(c) \mu^{-1} W_K' \left(\rho \bar{F}(c)\right) dc \\ &= -\int_0^{c_i} c \left(\mu^{-1} W_K \left(\rho \bar{F}(c)\right)\right)' dc \\ &= \int_0^{c_i} \mu^{-1} W_K \left(\rho \bar{F}(c)\right) dc - \left[c \mu^{-1} W_K \left(\rho \bar{F}(c)\right)\right]\Big|_0^{c_i} \\ &= \int_0^{c_i} \mu^{-1} W_K \left(\rho \bar{F}(c)\right) dc - c_i \mu^{-1} W_K \left(\rho \bar{F}(c_i)\right) \\ &= \int_0^{c_i} \mu^{-1} W_K \left(\rho \bar{F}(c)\right) dc - c_i W(b \mid G). \end{aligned}$$

Therefore,

$$\begin{aligned} u(R_i, c_i) &= R_i - c_i \cdot W(b(c_i) \mid G) - b(c_i) \\ &= R_i - \int_0^{c_i} \mu^{-1} W_K \left(\rho \bar{F}(c)\right) dc. \end{aligned}$$

□

Proof of Lemma 2: First, assume that all users participate. From Proposition 3, the equilibrium net surplus of an agent (R_i, c_i) conditional on all agents participating is

$$u(R_i, c_i) = R_i - \mu^{-1} \int_0^{c_i} W_K \left(\rho \bar{F}(c)\right) dc.$$

Because $u(R_i, c_i)$ is decreasing in R_i, c_i we have that for all (R_i, c_i)

$$\begin{aligned} u(R_i, c_i) &\geq u(R_L, \bar{c}) \\ &= R_L - \mu^{-1} \int_0^{\bar{c}} W_K \left(\rho \bar{F}(c)\right) dc \\ &= R_L - \bar{R} > 0. \end{aligned}$$

Additionally, we have that W_K is an increasing function, which implies that the utility $u(R_L, \bar{c})$ increases if less agents participate. Therefore, it is a strict best response for all agents to participate regardless of the participation decisions of other users. In other words, all agents participate in equilibrium and receive net surplus $u(R_i, c_i) \geq u(R_L, \bar{c}) > 0$. □

Proof of Theorem 2: From Lemma 2, we have that all agents participate and receive strictly positive surplus. From the expressions derived in Proposition 3, we have that transaction fees $b(c_i)$ are independent of the user's WTP and the exchange rate (a change in the exchange rate may change the nominal value written into the transaction, as users observe the exchange rate. Users trade off fees in USD against delay cost in USD equivalents).

Finally, if $\rho > 0$ we have that $b(c_i) > 0$ and the system raises strictly positive revenue. □

Proof of Corollary 2: Note that if the conditions of Theorem 2 are satisfied, they will also be satisfied if we increase WTP R of some or all the users. Therefore, both before and after the increase, the equilibrium transaction fees are given by $b(c_i)$ which is independent of WTP R. □

F.3 Delay and Revenue

In this section, we establish results relating to the total revenue generated by users and the total delay cost experienced by users in equilibrium. Theorems 3 and 4, which provide an expressions for the total revenue and delay cost, are implied by the following result:

Theorem A3. *The total revenue per unit time raised from users is*

$$\mathrm{Rev}_K(\rho) = K\rho^2 \int_0^{\bar{c}} cf(c)\bar{F}(c)W_K'\left(\rho\bar{F}(c)\right)\, dc \tag{23}$$

$$= K\rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_K\left(\rho\bar{F}(c)\right)\, dc. \tag{24}$$

The total delay cost per unit time incurred by users is

$$\mathrm{DelayCost}_K(\rho) = K\rho \int_0^{\bar{c}} cf(c)W_K\left(\rho\bar{F}(c)\right)\, dc. \tag{25}$$

The total overall cost per unit time borne by users is

$$\mathrm{TotalCost}_K(\rho) \triangleq \mathrm{Rev}_K(\rho) + \mathrm{DelayCost}_K(\rho) = K\rho \int_0^{\bar{c}} \bar{F}(c)W_K\left(\rho\bar{F}(c)\right)\, dc. \tag{26}$$

Proof. Transactions arrive per unit time at rate λ, and the expected revenue per

transaction is

$$\int_0^{\bar{c}} f(c)b(c)\,dc.$$

Therefore, the total expected revenue per unit time is

$$\begin{aligned}
\mathrm{Rev}_K(\rho) &= \lambda \int_0^{\bar{c}} f(c)b(c)\,dc \\
&= K\rho^2 \int_0^{\bar{c}} \int_0^{c} f(c)sf(s)W_K'\left(\rho\bar{F}(s)\right)\,ds\,dc \\
&= K\rho^2 \int_0^{\bar{c}} \int_s^{\bar{c}} f(c)sf(s)W_K'\left(\rho\bar{F}(s)\right)\,dc\,ds \\
&= K\rho^2 \int_0^{\bar{c}} sf(s)\bar{F}(s)W_K'\left(\rho\bar{F}(s)\right)\,ds.
\end{aligned}$$

This establishes (23). For (24), we integrate by parts with

$$\begin{aligned}
u = K\rho s\bar{F}(s),\ du = K\rho\left(\bar{F}(s) - sf(s)\right)\,ds, \\
dv = \rho f(s)W_K'\left(\rho\bar{F}(s)\right)\,ds,\ v = -W_K\left(\rho\bar{F}(s)\right),
\end{aligned}$$

to obtain

$$\begin{aligned}
\mathrm{Rev}_K(\rho) &= uv\Big|_0^{\bar{c}} - \int_0^{\bar{c}} v\,du \\
&= K\rho \int_0^{\bar{c}} \left(\bar{F}(s) - sf(s)\right) W_K\left(\rho\bar{F}(s)\right)\,ds,
\end{aligned}$$

as desired.

For the delay cost, note that the expected delay cost per transaction is

$$\int_0^{\bar{c}} f(c)\cdot c\mu^{-1}W_K\left(\rho\bar{F}(c)\right)\,dc.$$

Since transactions arrive at rate λ, the total expected revenue per unit time is then

$$\begin{aligned}
\mathrm{DelayCost}_K(\rho) &= \lambda \int_0^{\bar{c}} cf(c)\mu^{-1}W_K\left(\rho\bar{F}(c)\right)\,dc \\
&= K\rho \int_0^{\bar{c}} cf(c)W_K\left(\rho\bar{F}(c)\right)\,dc,
\end{aligned}$$

as desired. The expression for total cost per unit time (26) follows by combining (24)

and (25). □

Corollary 3, which establishes that total revenue and delay costs are increasing as functions of the load parameter ρ, is implied by the following result:

Corollary A4. *In equilibrium, if $\rho = 0$, both revenue and delay cost are zero. For all $\rho \in (0, 1)$,*

$$\mathrm{Rev}'_K(\rho) = K\rho \int_0^{\bar{c}} \bar{F}(c)^2 W'_K\left(\rho \bar{F}(c)\right)\, dc > 0,$$

$$\mathrm{DelayCost}'_K(\rho) = \frac{\mathrm{TotalCost}_K(\rho)}{\rho} > 0.$$

In other words, both revenue (and with it, infrastructure provision by miners) and delay cost are strictly increasing in ρ.

Proof. Differentiating (24) and applying (23),

$$\begin{aligned}
\mathrm{Rev}'_K(\rho) &= K \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_K\left(\rho \bar{F}(c)\right)\, dc \\
&\quad + K\rho \int_0^{\bar{c}} \left(\bar{F}(c)^2 - cf(c)\bar{F}(c)\right) W'_K\left(\rho \bar{F}(c)\right)\, dc \\
&= \frac{\mathrm{Rev}_K(\rho)}{\rho} + K\rho \int_0^{\bar{c}} \bar{F}(c)^2 W'_K\left(\rho \bar{F}(c)\right)\, dc - \frac{\mathrm{Rev}_K(\rho)}{\rho} \\
&= K\rho \int_0^{\bar{c}} \bar{F}(c)^2 W'_K\left(\rho \bar{F}(c)\right)\, dc.
\end{aligned}$$

Similarly, differentiating (25) and applying (23) and (26),

$$\begin{aligned}
\mathrm{DelayCost}'_K(\rho) &= K \int_0^{\bar{c}} cf(c) W_K\left(\rho \bar{F}(c)\right)\, dc + K\rho \int_0^{\bar{c}} cf(c)\bar{F}(c) W'_K\left(\rho \bar{F}(c)\right)\, dc \\
&= \frac{\mathrm{DelayCost}_K(\rho)}{\rho} + \frac{\mathrm{Rev}_K(\rho)}{\rho} = \frac{\mathrm{TotalCost}_K(\rho)}{\rho}.
\end{aligned}$$

□

F.4 Large Block Asymptotics

In this section, we establish asymptotic results in a "large block size" asymptotic regime. This is a regime where we consider a sequence of systems where the load parameter $\rho \triangleq \lambda/(\mu K) \in [0, 1)$ is held constant, while the block size $K \to \infty$.

The first result we establish in this regime is Lemma 3. The core of this Lemma is the observation that, in the large block regime, the expected waiting time measured in blocks, $W_K(\rho)$, is independent of K. The main intuition for this result is as follows. Fix the value of ρ. Consider a sequence of systems, indexed by the block size K, each with load ρ, as $K \to \infty$. When K is large, the arrival rate of new transactions must be very large relative to the service rate as which blocks are generated. Without loss of generality, suppose that the arrival rate of the Kth system is $\lambda_K = \rho K$ and the service rate of every system is $\mu = 1$, so the the load of each system is $\lambda_K/(\mu K) = \rho$ as desired. Now, over an interval of time of length t, the number of arrivals is given by a Poisson$(\lambda_K t)$ = Poisson$(\rho K t)$ distribution. Measured in units of the block size, this scaled number of arrivals process has the distribution

$$\frac{1}{K}\text{Poisson}(\rho K t) \to \rho t,$$

as $K \to \infty$, where the convergence is because the random variable on the left side has variance tending to zero, and hence is well-approximated by its mean. In other words, in this asymptotic regime, the number of new transactions is approximately deterministic and of order K, while services are at random times and also of order K. Therefore, it is natural to expect that the number of queued transactions, scaled by the block size K, converges in distribution as $K \to \infty$.

The following lemma makes this intuition precise:

Lemma A5. *Consider a sequence of bulk service queueing systems (as in Lemma A1) indexed by block size $K \geq 1$ with a fixed load parameter $\rho \in (0, 1)$, as $K \to \infty$. Define the random variable Q_K to be the steady-state distribution of the system when the block size is K.*

Then, Q_K is geometrically distributed with parameter $z_0(\rho, K)$ (cf. Lemma A1), where $z_0(\rho, K)$ asymptotically satisfies

$$z_0(\rho, K) = 1 - \alpha(\rho)/K + o(1/K), \tag{27}$$

as $K \to \infty$. Here, where $\alpha(\rho) > 0$ is the unique strictly positive root of the transcendental algebraic equation

$$e^{-\alpha} + \rho\alpha - 1 = 0.$$

Moreover, define $\tilde{Q}_K \triangleq Q_K/K$ to be the random variable corresponding to the steady-state queue length when the block size is K, measured in units of the block size

K. Then, as $K \to \infty$, $\tilde{Q}_K$ converges in distribution to an exponential distribution with parameter $\alpha(\rho)$.

Proof. Fix $\rho \in (0,1)$.

First, we will show that $\alpha(\rho)$ is well-defined. Define the transcendental function

$$T(\alpha) \triangleq e^{-\alpha} + \rho\alpha - 1.$$

Clearly $T(0) = 0$, $T'(0) < 0$, and $\lim_{\alpha\to\infty} T(\alpha) = \infty$. By the intermediate value theorem, there is at least one strictly positive root. Further, since $T''(\alpha) > 0$ for all $\alpha \geq 0$, the root must be unique. Thus,

$$T(\alpha) < 0, \quad \forall\ 0 < \alpha < \alpha(\rho); \qquad T(\alpha) > 0, \quad \forall\ \alpha > \alpha(\rho). \tag{28}$$

Next, we wish to prove (27). From Lemma A1, recall the polynomial defining z_0,

$$P_K(z) \triangleq z^{K+1} - (K\rho + 1)z + K\rho.$$

Note that

$$P_K(0) = K\rho > 0, \quad P_K(1) = 0, \quad P_K'(1) = K(1-\rho) > 0,$$

so $P_K(z)$ must be positive for z sufficiently close to zero, and must be negative for z sufficiently close to (but less than) 1. Since z_0 is the unique root of $P_K(\cdot)$ in the interval $[0,1)$, we have that

$$P_K(z) > 0, \quad \forall\ 0 \leq z < z_0(\rho, K); \qquad P_K(z) < 0, \quad \forall\ z_0(\rho, K) < z < 1. \tag{29}$$

Now, fix an arbitrary $\epsilon > 0$. Define

$$\underline{\nu}_K \triangleq 1 - \frac{\alpha(\rho) + \epsilon}{K}, \qquad \overline{\nu}_K \triangleq 1 - \frac{\alpha(\rho) - \epsilon}{K}.$$

Then,

$$\begin{aligned}\lim_{K\to\infty} P_K(\underline{\nu}_K) &= \lim_{K\to\infty} \underline{\nu}_K^{K+1} - (K\rho+1)\underline{\nu}_K + K\rho \\ &= \lim_{K\to\infty} \underline{\nu}_K \left(1 - \frac{\alpha(\rho)+\epsilon}{K}\right)^K + (K\rho+1)\frac{\alpha(\rho)+\epsilon}{K} - 1 \\ &= e^{-\left(\alpha(\rho)+\epsilon\right)} + \rho\big(\alpha(\rho)+\epsilon\big) - 1 \\ &= T\big(\alpha(\rho)+\epsilon\big) \\ &> 0,\end{aligned}$$

where (28) is used for the final inequality. Thus, for all K sufficiently large, $P_K(\underline{\nu}_K) > 0$. By (29), this implies that, for all K sufficiently large, $z_0(\rho, K) > \underline{\nu}_K$. Combining this with an analogous argument applied to $\overline{\nu}_K$, we have that, for all K sufficiently large,

$$1 - \frac{\alpha(\rho)+\epsilon}{K} < z_0(\rho, K) < 1 - \frac{\alpha(\rho)-\epsilon}{K},$$

or equivalently,

$$\left| z_0(\rho, K) - \left(1 - \frac{\alpha(\rho)}{K}\right)\right| < \frac{\epsilon}{K}.$$

Since ϵ is arbitrary, we have established (27).

To prove the convergence of $\tilde{Q}_K$ to the appropriate exponential distribution, notice that, for $t \geq 0$,

$$\mathsf{P}(\tilde{Q}_K \geq t) = \mathsf{P}(Q_K \geq tK) = \mathsf{P}(Q_K \geq \lceil tK \rceil) = z_0(\rho, K)^{\lceil tK \rceil} = z_0(\rho, K)^{K(\lceil tK \rceil / K)}. \tag{30}$$

Then,

$$\begin{aligned}\lim_{K\to\infty} \log \mathsf{P}(\tilde{Q}_K \geq t) &= \lim_{K\to\infty} (\lceil tK \rceil / K) \cdot K \log z_0(\rho, K) \\ &= t \cdot \lim_{K\to\infty} K \log z_0(\rho, K) \\ &= -t\alpha(\rho),\end{aligned} \tag{31}$$

where we have applied (27) and the fact that $\log(1-x) = -x + O(x^2)$ as $x \to 0$. □

The following lemma builds on the prior result to establish the first part of Lemma 3, which is that the expected waiting time (measured in blocks) converges and is independent of K:

Lemma A6. *Consider a fixed load parameter $\hat{\rho} \in (0,1)$. As block size K increases, the expected waiting time measured in blocks converges according to*

$$\lim_{K\to\infty} W_K(\hat{\rho}) = W_\infty(\hat{\rho}).$$

Here, $W_\infty(\hat{\rho})$ is the asymptotic expected delay (measured in blocks), defined for $\hat{\rho} \in (0,1)$ by

$$W_\infty(\hat{\rho}) \triangleq \frac{1}{1-\big(1+\alpha(\hat{\rho})\big)e^{-\alpha(\hat{\rho})}}, \tag{32}$$

where $\alpha(\hat{\rho}) > 0$ is defined in Lemma A5. For $\hat{\rho} = 0$, define $W_\infty(\hat{\rho}) \triangleq 1$ to coincide with the limiting value.

Moreover, the asymptotic expected delay satisfies

$$W'_\infty(0) = 0; \qquad W'_\infty(\hat{\rho}) > 0, \quad \forall\ \hat{\rho} \in (0,1).$$

Proof. The result is trivial for $\hat{\rho} = 0$.

Fix $\hat{\rho} > 0$. Equation (27) implies that there exists a sequence $\{\epsilon_K\}$ with limit $\epsilon_K \to 0$, such that

$$z_0(\hat{\rho}, K) = 1 - \frac{\alpha(\hat{\rho}) + \epsilon_K}{K}.$$

Then,

$$\begin{aligned}\lim_{K\to\infty} W_K\,(\hat{\rho})^{-1} &= \lim_{K\to\infty} (1-z_0)\big(1+K\hat{\rho}-(K+1)z_0^K\big)\\ &= \alpha(\hat{\rho})\hat{\rho} - \lim_{K\to\infty}\frac{K+1}{K}\big(\alpha(\hat{\rho})+\epsilon_K\big)z_0^K.\end{aligned}$$

But, as in (30)–(31), $z_0^K \to e^{-\alpha(\hat{\rho})}$. Also, from the transcendental algebraic equation defining $\alpha(\hat{\rho})$, we have that

$$\hat{\rho} = \frac{1-e^{-\alpha(\hat{\rho})}}{\alpha(\hat{\rho})}.$$

Therefore,

$$\lim_{K\to\infty} W_K\,(\hat{\rho})^{-1} = \alpha(\hat{\rho})\hat{\rho} - \alpha(\hat{\rho})e^{-\alpha(\hat{\rho})} = 1-\big(1+\alpha(\hat{\rho})\big)\,e^{-\alpha(\hat{\rho})},$$

as desired.

It remains to establish that $W'_\infty(\hat{\rho}) > 0$. Applying the implicit function theorem

to differentiate the equation $T\big(\alpha(\hat{\rho})\big) = 0$ with respect to $\hat{\rho}$, we have that

$$-e^{-\alpha(\hat{\rho})}\alpha'(\hat{\rho}) + \alpha(\hat{\rho}) + \hat{\rho}\alpha'(\hat{\rho}) = 0.$$

Simplifying, we obtain that

$$\alpha'(\hat{\rho}) = \frac{\alpha(\hat{\rho})}{e^{-\alpha(\hat{\rho})} - \hat{\rho}} = -\alpha(\hat{\rho})^2 W_\infty(\hat{\rho}).$$

Then, differentiating (32), we have that

$$W'_\infty(\hat{\rho}) = -\frac{e^{-\alpha(\hat{\rho})}\alpha(\hat{\rho})\alpha'(\hat{\rho})}{\left(1 - (1+\alpha(\hat{\rho}))\, e^{-\alpha(\hat{\rho})}\right)^2} = e^{-\alpha(\hat{\rho})}\alpha(\hat{\rho})^3 W_\infty(\hat{\rho})^3 > 0,$$

where the inequality holds for $\hat{\rho} \in (0,1)$. Observing that $\alpha(\hat{\rho}) \to \infty$ as $\hat{\rho} \to 0$, it follows that $W'_\infty(0) = 0$. □

Finally, we establish the second part of Lemma 3, which described the behavior of the large block asymptotic waiting time in the low load regime, as follows:

Lemma A7. *As $\rho \to 0$, we have that*

$$W_\infty(\rho) = 1 + \frac{1}{\rho}e^{-1/\rho} + o\left(\frac{1}{\rho}e^{-1/\rho}\right),$$

Proof. First, we will derive an asymptotic expression for $\alpha(\rho)$ when $\rho \to 0$. Suppose $\rho > 0$, if $\alpha > 0$ is the solution of

$$e^{-\alpha} + \rho\alpha - 1 = 0,$$

then $\beta \triangleq \alpha - 1/\rho > -1/\rho$ must solve

$$-\frac{1}{\rho}e^{-1/\rho} = \beta e^{\beta}.$$

The two real solutions to this transcendental equation can be expressed as

$$\beta = \mathcal{W}_i\left(-\frac{1}{\rho}e^{-1/\rho}\right), \quad \forall\, i = -1, 0,$$

where $\mathcal{W}_0(\cdot)$ and $\mathcal{W}_{-1}(\cdot)$ are the two branches of the Lambert W-function (for the

definition and properties of this function, see, e.g., Olver et al. 2010). Since $\beta > -1/\rho$, we can restrict to the $i = 0$ case (the so-called 'principal branch'), to obtain

$$\alpha(\rho) = \frac{1}{\rho} + \mathcal{W}_0\left(-\frac{1}{\rho}e^{-1/\rho}\right).$$

As $x \to 0$, from the Taylor expansion it is easy to see that $\mathcal{W}_0(x) = x + O(x^2)$. Then, as $\rho \to 0$,

$$\alpha(\rho) = \frac{1}{\rho} + O\left(\frac{1}{\rho}e^{-1/\rho}\right).$$

Now, we can analyze the asymptotic waiting time. As $\rho \to 0$, $\alpha(\rho) \to \infty$, so that

$$\big(1 + \alpha(\rho)\big)e^{-\alpha(\rho)} \to 0.$$

Since $1/(1-x) = 1 + x + O(x^2)$ as $x \to 0$, we have that

$$\begin{aligned} W_\infty(\rho) &= 1 + \big(1 + \alpha(\rho)\big)e^{-\alpha(\rho)} + o\left(\big(1 + \alpha(\rho)\big)e^{-\alpha(\rho)}\right) \\ &= 1 + \alpha(\rho)e^{-\alpha(\rho)} + o\left(\alpha(\rho)e^{-\alpha(\rho)}\right) \\ &= 1 + \frac{1}{\rho}e^{-1/\rho} + o\left(\frac{1}{\rho}e^{-1/\rho}\right). \end{aligned}$$

□

The following Theorem implies Theorems 5–6:

Theorem A8. *For a fixed load $\rho \in [0, 1)$, as the block size $K \to \infty$, we have that*

$$\begin{aligned} \mathrm{Rev}_K(\rho) &= K \cdot \mathrm{Rev}_\infty(\rho) + o(K), \\ \mathrm{DelayCost}_K(\rho) &= K \cdot \mathrm{DelayCost}_\infty(\rho) + o(K), \\ \mathrm{TotalCost}_K(\rho) &= K \cdot \mathrm{TotalCost}_\infty(\rho) + o(K), \end{aligned}$$

where

$$\begin{aligned} \mathrm{Rev}_\infty(\rho) &\triangleq \rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_\infty\left(\rho\bar{F}(c)\right)\, dc, \\ \mathrm{DelayCost}_\infty(\rho) &\triangleq \rho \int_0^{\bar{c}} cf(c) W_\infty\left(\rho\bar{F}(c)\right)\, dc. \\ \mathrm{TotalCost}_\infty(\rho) \triangleq \mathrm{Rev}_\infty(\rho) + \mathrm{DelayCost}_\infty(\rho) &= \rho \int_0^{\bar{c}} \bar{F}(c) W_\infty\left(\rho\bar{F}(c)\right)\, dc. \end{aligned}$$

Furthermore, for all $\rho \in (0,1)$,

$$\mathrm{Rev}'_\infty(\rho) = \rho \int_0^{\bar{c}} \bar{F}(c)^2 W'_\infty\left(\rho \bar{F}(c)\right)\, dc > 0,$$
$$\mathrm{DelayCost}'_\infty(\rho) = \frac{\mathrm{TotalCost}_\infty(\rho)}{\rho} > 0.$$

In other words, both the asymptotic revenue (and with it infrastructure provision by miners) and the asymptotic delay cost are strictly increasing in ρ.

Finally, as $\rho \to 0$,

$$\mathrm{Rev}_\infty(\rho) = O\left(e^{-1/\rho}\right),$$
$$\mathrm{DelayCost}_\infty(\rho) = \rho \cdot \mathbb{E}\left[c\right] + o\left(\rho\right).$$

In other words, for small values of the load ρ*, the asymptotic delay cost grows linearly in* ρ*, but the revenue grows slower than any polynomial in* ρ.

Proof. Note that, from (24),

$$\frac{\mathrm{Rev}_K(\rho)}{K} = \rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_K\left(\rho \bar{F}(c)\right)\, dc. \tag{33}$$

Since $W_K(\cdot)$ is strictly increasing,

$$\left|\left(\bar{F}(c) - cf(c)\right) W_K\left(\rho \bar{F}(c)\right)\right| \leq \left(\bar{F}(c) + cf(c)\right) W_K\left(\rho\right).$$

Now, pick any $\bar{\rho} \in (\rho, 1)$. Then $W_K(\rho) \to W_\infty(\rho) < W_\infty(\bar{\rho})$ by Lemma A6, so for K sufficiently large,

$$\left|\left(\bar{F}(c) - cf(c)\right) W_K\left(\rho \bar{F}(c)\right)\right| \leq \left(\bar{F}(c) + cf(c)\right) W_\infty\left(\bar{\rho}\right),$$

which is integrable over $c \in [0, \bar{c}]$. Then, we can apply the dominated convergence theorem to (33) to obtain

$$\lim_{K\to\infty} \frac{\mathrm{Rev}_K(\rho)}{K} = \rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_\infty\left(\rho \bar{F}(c)\right)\, dc \triangleq \mathrm{Rev}_\infty(\rho),$$

as desired.

The asymptotic $K \to \infty$ limits for delay cost and total cost can be established

using similar dominated convergence theorem arguments. Further, the derivative expressions can be derived directly by differentiation.

Finally, we wish to describe the asymptotic revenue $\mathrm{Rev}_\infty(\rho)$ and the asymptotic delay cost $\mathrm{DelayCost}_\infty(\rho)$ as $\rho \to 0$. For the asymptotic revenue,

$$\begin{aligned}\mathrm{Rev}_\infty(\rho) &= \rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_\infty\left(\rho \bar{F}(c)\right)\, dc \\ &= \rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) \left(W_\infty\left(\rho \bar{F}(c)\right) - 1\right)\, dc\end{aligned}$$

where we have used the fact that

$$\int_0^{\bar{c}} \bar{F}(c)\, dc = \int_0^{\bar{c}} cf(c)\, dc = \mathbb{E}[c].$$

Then, applying Lemma A7

$$\begin{aligned}\mathrm{Rev}_\infty(\rho) &\le \rho \int_0^{\bar{c}} \left|\bar{F}(c) - cf(c)\right| \cdot \left|W_\infty\left(\rho \bar{F}(c)\right) - 1\right|\, dc \\ &\le \rho \int_0^{\bar{c}} \left(\bar{F}(c) + cf(c)\right) \cdot \left|W_\infty\left(\rho\right)\right) - 1|\, dc \\ &\le 2\rho \mathbb{E}(c) \left|W_\infty\left(\rho\right)\right) - 1| \\ &\le 2\mathbb{E}(c) e^{-1/\rho} + o\left(e^{-1/\rho}\right).\end{aligned}$$

For the asymptotic delay cost, applying the dominated convergence theorem,

$$\lim_{\rho \to 0} \frac{\mathrm{DelayCost}_\infty(\rho)}{\rho} = \int_0^{\bar{c}} cf(c) W_\infty(0)\, dc = \mathbb{E}[c].$$

□

The following theorem implies Theorem 7:

Theorem A9. *Consider a target level of revenue $R^* > 0$ and a block size K. Define $\mathrm{DelayCost}_K^*(R^*)$ to be the delay cost required to achieve revenue R^*, under the asymptotic large K regime. That is, define*

$$\mathrm{DelayCost}_K^*(R^*) \triangleq K\, \mathrm{DelayCost}_\infty\left(\mathrm{Rev}_\infty^{-1}(R^*/K)\right),$$

where

$$\mathrm{Rev}_\infty^{-1}(r) \triangleq \inf\left\{\rho > 0 \ : \ \mathrm{Rev}_\infty(\rho) \geq r\right\},$$

for $r > 0$.

Then, as $K \to \infty$,

$$\mathrm{DelayCost}_K^*(R^*) = \Omega\left(\frac{K}{\log K}\right).$$

Proof. Define $\rho_K \triangleq \mathrm{Rev}_\infty^{-1}(R^*/K)$, so that $\mathrm{Rev}_\infty(\rho_K) = R^*/K$ for all K. Then,

$$\begin{aligned}
\mathrm{DelayCost}_K^*(R^*) &= K\,\mathrm{DelayCost}_\infty\left(\rho_K\right) \\
&= K\rho_K \int_0^{\bar{c}} c f(c) W_\infty\left(\rho_K \bar{F}(c)\right)\, dc \\
&\geq K\rho_K \mathbb{E}[c],
\end{aligned}$$

using the fact that $W_\infty(\cdot) \geq 1$. Hence, it suffices to prove that

$$\rho_K = \Omega\left(\frac{1}{\log K}\right) \tag{34}$$

as $K \to \infty$.

Now, if ρ_K is bounded away from zero as $K \to \infty$, (34) clearly holds. Assume otherwise that $\rho_K \to 0$ as $K \to \infty$. Theorem A8 implies that there exists a constant C such that, for K sufficiently large,

$$\frac{R^*}{K} = \mathrm{Rev}_\infty(\rho_K) \leq Ce^{-1/\rho_K}.$$

Equivalently,

$$\rho_K \geq \frac{1}{\log CK/R^*},$$

for K sufficiently large, which establishes (34). □

F.5 Profit-Maximizing Firm

Proof of Proposition 1. Notice that the firm can make a profit of $\lambda_H\left(R_H - c_f\right)$ by processing only transactions of R_H agents without delay at a fee R_H. Since this extracts all the possible surplus from R_H agents, this is optimal for the firm out of all pricing schemes that do not process transactions from R_L agents.

We follow to formulate the problem and show the firm cannot do better by processing some transactions from R_L agents. By the revelation principle, the firm's problem can be written as a choice of an incentive compatible direct mechanism where the firm offers a menu $\{x(\cdot,\cdot), W(\cdot,\cdot), b(\cdot,\cdot)\}$. Agents report their type $(R_i, c_i) \in \{R_H, R_L\} \times \mathbb{R}_+$. If $x(R_i, c_i) = 0$, the agent's transaction is not processed and the agent does not pay or wait. If $x(R_i, c_i) = 1$, the agent's transaction is processed after delay $W(R_i, c_i)$ and the agent is charged a transaction fee $b(R_i, c_i)$. If $x(R_i, c_i) \in (0, 1)$, the transaction is processed with probability $x(R_i, c_i)$, expected delay $W(R_i, c_i)$ and expected transaction fee $b(R_i, c_i)$.

The utility of a risk neutral agent of type (R_i, c_i) who reports (R, c) is

$$u(R, c|R_i, c_i) = x(R, c) R_i - c_i \cdot W(R, c) - b(R, c),$$

and we write $u(R_i, c_i) = u(R_i, c_i|R_i, c_i)$.

The firm's problem is stated by the following optimization problem:

$$\begin{aligned} \max_{x,W,b} \sum_{\tau \in \{H,L\}} \lambda_\tau \int_0^{\bar{c}} (b(R_\tau, c) - c_f x(R_\tau, c)) \, dF(c) & \\ u(R_i, c_i) \geq u(R, c|R_i, c_i) & \quad \forall R_i, c_i, R, c \text{ (IC-R,c)} \\ \text{s.t.:} \quad u(R_i, c_i) \geq 0 & \quad \forall R_i, c_i \text{ (PC-R,c)} \\ x(R, c) \in [0, 1], \; W(R, c) \geq 0, \; b(R, c) \geq 0. & \end{aligned} \tag{35}$$

The optimal value of (35) is bounded by the value of the firm's problem when the agent's waiting cost c_i is observed by the firm, which is given by

$$\begin{aligned} \max_{x,W,b} \sum_{\tau \in \{H,L\}} \lambda_\tau \int_0^{\bar{c}} (b(R_\tau, c) - c_f x(R_\tau, c)) \, dF(c) & \\ u(R_i, c_i) \geq u(R, c_i|R_i, c_i) & \quad \forall R_i, c_i, R \text{ (IC-R)} \\ \text{s.t.:} \quad u(R_i, c_i) \geq 0 & \quad \forall R_i, c_i \text{ (PC-R,c)} \\ x(R, c) \in [0, 1], \; W(R, c) \geq 0, \; b(R, c) \geq 0. & \end{aligned} \tag{36}$$

Because problem (36) is separable across different c_i, the optimal value of (36) is the total value of the optimal solutions for each fixed c_i. We rewrite the problem for

a fixed c_i and omit the dependency on c_i to obtain the problem (37)

$$\begin{aligned}
&\max_{x,W,b} \sum_{\tau\in\{H,L\}} \lambda_\tau \left(b\left(R_\tau\right) - c_f x\left(R_\tau\right)\right) \\
&\text{s.t.:} \quad \begin{aligned} u\left(R_i\right) &\geq u\left(R|R_i\right) && R_i, R \in \{R_H, R_L\} \ \text{(IC-R)} \\ u\left(R_i\right) &\geq 0 && R_i \in \{R_H, R_L\} \ \text{(PC-R,c)} \end{aligned} \\
&\qquad x\left(R\right) \in [0,1], \ W\left(R\right) \geq 0, \ b\left(R\right) \geq 0 .
\end{aligned} \tag{37}$$

Dropping the IC-R_L and PC-R_H constraints and plugging in expressions we obtain the relaxed problem (38)

$$\begin{aligned}
&\max_{x,W,b} \sum_{\tau\in\{H,L\}} \lambda_\tau \left(b\left(R_\tau\right) - c_f x\left(R_\tau\right)\right) \\
&\text{s.t.:} \quad \begin{aligned} x\left(R_H\right) R_H - c\cdot W\left(R_H\right) - b\left(R_H\right) &\geq x\left(R_L\right) R_H - c\cdot W\left(R_L\right) - b\left(R_L\right) && \text{(IC-}R_H\text{)} \\ x\left(R_L\right) R_L - c\cdot W\left(R_L\right) - b\left(R_L\right) &\geq 0 && \text{(PC-}R_L\text{)} \end{aligned} \\
&\qquad x\left(R\right) \in [0,1], \ W\left(R\right) \geq 0, \ b\left(R\right) \geq 0 .
\end{aligned} \tag{38}$$

If PC-R_L does not bind in (38), we can increase $b\left(R_L\right), b\left(R_H\right)$ by the same amount and increase the objective. Therefore, it must be that PC-R_L binds in (38) and we have

$$b\left(R_L\right) = x\left(R_L\right) R_L - c\cdot W\left(R_L\right).$$

This allows us to replace IC-R_H with

$$x\left(R_H\right) R_H - c\cdot W\left(R_H\right) - b\left(R_H\right) \geq x\left(R_L\right)\left(R_H - R_L\right),$$

and rewrite problem (38) as

$$\begin{aligned}
&\max_{x,W,b} \ \lambda_H \left(b\left(R_H\right) - c_f\cdot x\left(R_H\right)\right) + \lambda_L \left(x\left(R_L\right) R_L - c\cdot W\left(R_L\right) - c_f\cdot x\left(R_L\right)\right) \\
&\text{s.t.:} \quad \begin{aligned} x\left(R_H\right) R_H - c\cdot W\left(R_H\right) - b\left(R_H\right) &\geq x\left(R_L\right)\left(R_H - R_L\right) && \text{(IC-}R_H\text{)} \end{aligned} \\
&\qquad x\left(R\right) \in [0,1], \ W\left(R\right) \geq 0, \ b\left(R\right) \geq 0 .
\end{aligned} \tag{39}$$

Considering problem (39), we see that $W\left(R_L\right)$ only appears in the objective, and lowering it weakly increases the objective. $W\left(R_H\right)$ only appears in the constraint, and lowering it relaxes the constraint. If the IC-R_H does not bind, we can increase $b\left(R_H\right)$ and increase the objective. Therefore, in any optimal solution we have that

$W(R_H) = W(R_L) = 0$. This reduces (39) to a standard two-type price discrimination problem.

Because the IC-R_H must bind, we have

$$b(R_H) = x(R_H) R_H - x(R_L)(R_H - R_L).$$

Plugging this into the objective and rearranging we obtain

$$\begin{aligned}
&\lambda_H (b(R_H) - c_f \cdot x(R_H)) + \lambda_L (x(R_L) R_L - c \cdot W(R_L) - c_f \cdot x(R_L)) \\
=&\lambda_H (x(R_H) R_H - x(R_L)(R_H - R_L) - c_f \cdot x(R_H)) + \lambda_L (x(R_L) R_L - c_f \cdot x(R_L)) \\
=&x(R_H)(\lambda_H R_H - \lambda_H c_f) + x(R_L)((\lambda_L + \lambda_H)(R_L - c_f) - \lambda_H (R_H - c_f)).
\end{aligned}$$

We assumed $\lambda_H R_H > (\lambda_H + \lambda_L) R_L$, which implies that $(\lambda_L + \lambda_H)(R_L - c_f) < \lambda_H (R_H - c_f)$. Therefore, the unique optimal solution of (39) is obtained by

$$x(R_H) = 1, b(R_H) = R_H$$

and

$$x(R_L) = b(R_L) = W(R_H) = W(R_L) = 0.$$

It is straightforward to verify that this solution satisfies all the constraints of (37), and we have therefore obtained the unique optimal solution to (37). By integrating over all c we also obtain the solution to (36), which is therefore also the unique optimal solution to (35). That is, it is optimal for the firm to process only transactions of R_H agents without delay at a fee R_H. □