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

Gur Huberman Jacob D. Leshno Ciamac Moallemi
Columbia Business School

August 30, 2017

Abstract

Owned by nobody and controlled by an almost immutable protocol the Bitcoin payment system is a platform with two main constituencies: users and profit seeking miners who maintain the system's infrastructure. The paper seeks to understand the economics of the system: How does the system raise revenue to pay for its infrastructure? How are usage fees determined? How much infrastructure is deployed? What are the implications of changing parameters in the protocol?

A simplified economic model that captures the system's properties answers these questions. Transaction fees and infrastructure level are determined in an equilibrium of a congestion queueing game derived from the system's limited throughput. The system eliminates dead-weight loss from monopoly, but introduces other inefficiencies and requires congestion to raise revenue and fund infrastructure. We explore the future potential of such systems and provide design suggestions.

*We are grateful to Campbell Harvey, Refael Hassin, Seth Stephens-Davidowitz and Aviv Zohar for helpful conversations and to seminar participants at the Central Bank of Finland, Columbia, EIEF, MSR-NYC, NYCE and Stanford for helpful comments.

1 Introduction

A crypto-currency is a digital currency stored on an open and decentralized electronic payment system. Following Nakamoto (2008), crypto-currencies have caught the attention of industry, academia and the public at large, with Bitcoin being the most prominent. There are hundreds of such crypto-currencies, many running on large and reliable decentralized networks of anonymous computers. This wave has been enabled by an innovative computer science design named "blockchain". The blockchain supports the creation of a decentralized electronic payment system that can be trusted, although none of the system's servers is individually trusted. The novel blockchain design relies on a combination of cryptography and game theory-based incentives. It has received much public interest on its own right.

The blockchain design enables Bitcoin and other crypto-currencies to function similarly to conventional payment systems such as Fed Wire, Swift, Visa, and PayPal. These payment systems are natural monopolies in that they enjoy economies of scale and network effects. Each of them is operated by an organization that determines its rules and modifies them as circumstances change. These rules include how and how much participants pay for using the system. The governing organization ensures the system is trusted and is responsible for maintaining the required infrastructure for the system. Payment systems are often regulated (or outright owned by government agencies) in order to mitigate the welfare loss associated with their monopolistic positions.

The innovation in Bitcoin's blockchain design is in the absence of a governing organization. Rather, a protocol sets the system's rules, by which all constituents abide. Absent is a central entity that maintains the infrastructure. Rather, Bitcoin's infrastructure consists of computer servers (called "miners") which enter and exit the system at will, responding to perceived profit opportunities.[1] Participants follow the protocol because it is in their best interest to do so, assuming the other participants follow the protocol. Thus, the protocol-derived rules are practically fixed and binding on all parties.

The blockchain design carries an economic innovation. Unlike other payment systems, Bitcoin is a two-sided platform with rules that are pre-specified by a computer protocol. No participant has power to set or modify fees or rules of conduct or otherwise control the system. Each participant in the market place, users and miners alike, is a price taker. Users are provided protection from monopoly pricing: even if the system becomes a monopoly, there is no monopolist who charges monopolist fees. However, for

[1] Mining is permissionless in Bitcoin, i.e., any computer can serve as a miner.

the system to function properly it must raise sufficient revenue from the users to fund the required infrastructure. We aim to understand how fees, system revenue and amount of infrastructure are determined in equilibrium, and how these are affected by the rules set by the Bitcoin protocol.

In order to analyze the system, we first construct a simplified model that captures the economic environment generated by the Bitcoin protocol. We provide a simple description of the blockchain protocol and the Bitcoin system in Section 3, illustrating how the Bitcoin system is enabled by a combination of cryptographic tools that enable verification of the ledger, together with game theoretic structure that sets incentives to reach consensus on a legal ledger. Abstracting away from much of the technical detail, we translate this description to a simplified economic model.

The Bitcoin system's two main constituencies are users, identified with the transactions they send, and servers, called miners. The miners collectively maintain a ledger of all transaction in a format called the blockchain, where transactions are arranged in blocks. Each transaction is a cryptographically verified message. Every 10 minutes (on average) the Bitcoin system randomly selects one miner to add a block of transactions to the ledger, processing all the transactions within that block. The selected miner is said to have "mined the block". Equilibrium between many small miners ensures that all miners maintain consensus on the same ledger, and only legal transactions are processed. The protocol limits each block to no more than approximately 2,000 transactions.[2] Therefore the system's throughput is bounded, and does not depend on the number of miners.

To provide proper incentives, the system compensates miners for their effort by rewarding miners when they are selected to mine a block. The reward consists of newly minted coins and the transaction fees paid by the transactions processed in the block. The protocol specifies how many newly minted coins are awarded in each block. This number is cut in half approximately every four years. In contrast, transaction fees are not fixed by the protocol, and users choose the transaction fees they pay.

The simplified economic model captures these features of the system.

Two sets of seemingly disparate questions are the starting point of the analysis: (i) in the long run, who will pay the miners and why? (ii) if the system becomes popular how will it manage its limited throughput? Hpw will service priority be assgined? The absence of a Bitcoin-controlling organization renders both questions non trivial. The model asserts a single answer to both questions: The system's congestion due to its limited throughput

[2]The protocol limits each block to 1MB of data, and the average transaction requires 0.5KB of data (Zohar 2015).

leads users to pay transactions fees to gain processing priority. These very fees fund the miners. This answer raises follow up questions regarding social efficiency, stability, robustness and improved parameter choice.

The model allows us to analyze the long run behavior of the system, when miners are compensated solely from transaction fees. Throughout our analysis we assume that the system is operating reliably. We derive equilibrium behavior of users and miners to obtain expressions for equilibrium transaction fees, revenue and infrastructure. The structure of the system enables a separate analysis of the miners and the users, as follows.

Many small miners can enter or exit the system. Active miners compete to get selected and collect transaction fees. Transaction fees are chosen by users, and small miners cannot affect the choices of users. Therefore, every miner maximizes his reward by processing the block of transactions with the highest fees. Miners decide whether to enter the system by comparing their cost of operating to their expected revenue given the block reward and their chances of being selected. It follows that each miner's expected profit is zero, and the amount of revenue determines the number of miners through miners' entry decision. This is remarkable since the system as a whole is a monopolist, yet it provides its service at cost.

The analysis of miners implies that user decisions are independent of the number of miners, as long as there is a sufficient number of miners for the system to operate reliably. The system's throughput is fixed by the protocol. Higher fee transactions are prioritized by any miner. Therefore, users can choose to pay a higher transaction fee to avoid costly delays.

To understand how users select their transaction fees, we analyze the implied congestion queueing game. The system is stochastic due to random arrival of transactions and the random mining of new blocks. The stochastic nature of the system implies that some transaction will be delayed even if the system has sufficient capacity to eventually process all transactions. A transaction's delay increases with the overall congestion in the system. Gaining priority through higher transaction fees reduce delay. Analysis of the stochastic system allows us to calculate each user's trade-off between transaction fees and delay costs, and derive each user's equilibrium transaction fee. Each user's equilibrium transaction fee equals the externality his transaction imposes. Thus, equilibrium transaction fees coincide with the payments that result from selling priority of service in a VCG auction.

The derived expressions for total delay costs borne by users, total transaction fees paid to miners, and the equilibrium number of miners in the system have a few implications.

Users will pay fees only if the system is sufficiently congested and delays are costly to them. While the stochastic nature of the system can generate delays even if all transactions are processed, a system where capacity far exceeds demand will not generate sufficient delays to generate revenue. Moreover, unless users pay substantial delay costs, the system cannot raise even small amounts of revenue. While raising revenue through congestion pricing is a revenue generating mechanism that protects users from monopoly pricing, transaction fees vary with the congestion level, without regard for their role in determining the number of miners. Thus, the amount of revenue raised by the system is unlikely to lead to an efficient amount of infrastructure provision by miners.

Absence of sufficient congestion can be disastrous for the system. Without sufficient congestion, users pay almost no transaction fees, generating almost no revenue to fund miners. As miners exit, the system becomes less reliable, leading users to leave the system thereby reducing congestion further. Because the system's throughput does not depend on the number of miners, the miners' entry and exit choices do not help balance the system. Without an alternative way to maintain the miners, the system will collapse.

The Bitcoin system offers an alternative to regulating a monopoly or controlling prices via market competition. Bitcoin can be a monopoly in the sense that all potential users are using the Bitcoin system, without being priced at the monopoly price. Instead, transaction fees are determined in equilibrium. Miners compete within the system, and provide infrastructure at cost. However, maintaining the Bitcoin system requires spending of real resources. These resources include much of the mining effort as miners duplicate each other's effort and participate in a tournament to determine the miner whose block is next on the blockchain. Moreover, costly delays in transaction processing are a necessity. These inefficiencies can be lower or higher than the dead-weight loss of monopoly.

Our analysis suggests two simple design modifications to the protocol. First, Instead the fixed throughput of the current system, the system should adjust the frequency at which new blocks are added. Making this frequency a function of recent congestion is possible without raising any incentive problems, as past congestion can be inferred from the blockchain ledger which is observable to the protocol. This can allow the system to maintain the desired congestion level as the volume of users varies, keeping revenue and therefore the number of miners at the desired level.

Second, our analysis shows that raising a target revenue level requires imposing less delay costs on the users if the maximal block size is lower. Thus, it would be beneficial to redesign the system with the smallest block sizes possible (given engineering constraints), and maintaining throughput through frequent small blocks.

Finally, we pose the question of characterizing the set of feasible revenue generating mechanisms for distributed blockchain systems and identifying the optimal one. Revenue generating rules must provide proper incentives for miners to process transaction, and transaction fees must verifiable by third parties. We discuss some possible mechanisms that can be implemented by the protocol.

Organization of the paper

Section 2 provides a review of related literature. Section 3 provides a simplified explanation of the Bitcoin system and the underlying blockchain technology. Section 4 introduces our economic model and includes most of the technical analysis. The implications of the analysis for the current design of the Bitcoin system are discussed in Section 5. Section 6 discusses alternative design suggestions. Section 7 provides some final remarks. Omitted proofs are in the appendix.

2 Related Literature

2.1 Engineering of Bitcoin

Famously, a white paper by Nakamoto (2008) coined the term and described the Bitcoin system. 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 Bitcoin system. Narayanan et al. (2016) offer an elaborate description and analysis of the system. Croman et al. (2016) provide cost estimates for the Bitcoin system and analyze the potential for transaction throughput. Eyal et al. (2016) suggest an alternative design aimed to construct a system with a higher transaction throughput. Carlsten et al. (2016) analyze how incentives for miners changes when miners are rewarded with transaction fees instead of newly created coins. Chiu & Koeppl (2017) evaluate the welfare implications of printing new coins. Adopting a mostly empirical orientation, Easley et al. (2017) is a contemporaneous piece. Some of their results echo those reported here.

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.

2.2 Bitcoin usage as a currency and the crypto-currency market

Ron & Shamir (2013), Athey et al. (2016) provide analysis of the usage of Bitcoin and its value as a currency. Yermack (2013) reviews the history of Bitcoin and the statistical properties of 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 crypto-currencies. Halaburda & Sarvary (2016) review the crypto-currency 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 (2016) discuss possible opportunities that can arise from blockchain technology.

2.3 Related work in queuing theory

Lui (1985), Glazer & Hassin (1986), 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 single service station which serves customers one at a time. The main observation of Lui (1985) 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) provides an summary of the results.

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

Separately, Kasahara & Kawahara (2016) analyze delays in a priority queueing system with batch service inspired by Bitcoin, but do not consider user incentives or equilibrium considerations.

2.4 Work on competition, monopoly and its regulation

The social welfare implications of monopolistic vs. competitive provision of a good or service are of central concern to economic analysis, often leading to a debate regarding the extent to which regulation is desired and the best means through which to accomplish it.

Posner (1975) offers a clean position, "This paper presents a model [...] of the social cost of monopoly and monopoly-inducing regulation [...] [I]t assumes that competition to obtain a monopoly results in a transformation of monopoly profits into social costs. A major conclusion is that public regulation is probably a larger source of social cost than private monopoly."

A Posner-inspired interpretation of mining is that when a block is completed – i.e., the hard puzzle has been solved by one of the miners – the solving miner is a monopolistic winner who takes all the revenues associated with the completion of that block. The social cost of one miner's winning is the amount spent by the community of miners to try to solve the hard puzzle. Noteworthy is that the monopolist is *not* a price-setter, in contrast with standard monopoly models, including Posner's.

3 A Brief Description of the Bitcoin System

This section provides a simplified explanation of the permissionless blockchain protocol that underlies the Bitcoin system and is the basis of many other crypto-currencies. The description focuses on the economic elements.[3] In order to describe what the Bitcoin system does, it is useful to first explain what is needed for a payment system such as 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 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.

[3]For further details and an explanation of the cryptographic elements of the system please refer to Narayanan et al. (2016).

One simple implementation is just a spread-sheet (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 is 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 transaction. 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.

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 a 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. The maximal block size is 1MB. (The need to address network latency motivates this structure.)

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 reach 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[4] (known as proof of work) that is useful only for generating a verifiable random 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 when he mines a new block, but 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 to assemble and create legal blocks. Consensus forms on a ledger that includes the new block. The process continues

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

in the same manner for the following ten minutes (on average) and so on.[5]

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 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.

4 Economic Model

The description in Section 3 establishes the following attributes of Bitcoin when the system functions reliably. New blocks are added to the ledger at Poisson[6] rate μ (independently of the number of miners). Each block is mined (created) by a randomly chosen miner, and vetted by all other miners. A block can contain up to K transactions. A transaction is deemed processed once it is included in a legal block. Pending transactions not included in a block wait to be processed in a future block. A miner who mines a new block is rewarded with the transaction fees of transactions included in that block. In addition, the model assumes that no new coins are minted. These features are the ingredients of the model we study.

Identifying users with their transactions, we assume for simplicity that each user sends a single transaction. Transactions arrive according to a Poisson process of rate λ. Each transaction specifies a transaction fee b, which the user chooses. The system does not process transactions immediately, and delays are costly to users. Delay costs per unit time vary across users, are denoted by c, and are distributed $c \sim F[0, \bar{c}]$. 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

[5]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.

[6]A Poisson process is the limit of many independent binomial trials. See footnote 4.

observe other pending transactions at the time they submit their transaction. Users are risk neutral and select the fee b to maximizes their expected net reward $R-b-c\cdot W$, where R is the reward for having the transaction go through and $W = W(b)$ is the expected delay. The payoff for users who opt out of the system is normalized to 0.

The community of potential miners is large. Each of its members can join the system and become an active miner. Active miners employ their computational power in an attempt to get selected to mine a block, i.e. get selected assemble a block of transactions that is sent to all other miners and added to the blockchain ledger. Active miners also observe all pending transactions, keep a copy of the blockchain ledger, and append legally mined blocks to the ledger as they receive them. All miners maintain consensus on the same ledger, and the blockchain does not fork.[7] For simplicity, all active miners have the same computing power, and therefore have equal chances of being selected to mine the next block. Additionally, all miners observe all pending transactions and incur the same cost of c_m per unit time while they remain active.[8] Active miners can exit and become inactive without penalty. We denote the number of active miners by N.

Our main interest is the analysis of the system when it provides reliable service, which requires the following additional assumptions. The number of miners N is large enough for the system to be reliable and secure, and for each miner to be small. A large number of servers N guarantees consistent quality of service even when some servers occasionally fail or exit the system. Each miner needs to be small to ensure that miners cannot block or erase transactions from the ledger.[9] The system is also secure when its aggregate computational power is large because then it would be prohibitively expensive to marshal the computational resources to overtake the system.

The measure of the congestion is given by the load parameter $\rho = \lambda/\mu K$. To guarantee that all transactions are eventually processed, we assume that $\rho < 1$. The reward R is assumed to be large enough for all users to have a positive net reward, and thus to choose to participate in the system.

[7]Forks may happen because of communication latency or because of strategic behavior of sufficiently large miners (see footnote 5). We abstract away from both of these issues.

[8]The analysis can be naturally extended to miners who differ in their computational power or costs, as long as each miner remains small. See the discussion in Section 4.1.

[9]A miner that has strictly more than 50% of the computing power in the Bitcoin system can erase processed transactions or block transactions. Eyal & Sirer (2014) argue that the Bitcoin system is guaranteed to be reliable and secure only when every miner is small.

4.1 Miner behavior

Active miners use their computational power in an attempt to mine the next block and receive the reward. By assumption, each miner is small and cannot affect the behavior of users. Miners observe the current pool of pending transactions and the fees they offer, and maximize their profit by assembling a block that would deliver the highest possible reward.[10] Therefore, each miner maximizes his profit by assembling a legal[11] block that includes the K transactions offering the highest fees. (If there are fewer than K pending transactions the block includes all of them.)

All miners observe the same pool of pending transactions, and therefore assemble identical blocks if selected. Thus, all miners expect to receive the same reward if selected to mine the next block. Since each miner has equal chance of mining the next block, each miner in expectation receives a payment of Rev/N per unit time, where Rev is the total transaction fees from processed transactions per unit time.[12] The users determine transaction fees and therefore Rev which we calculate in Section 4.2. That analysis shows that Rev is independent of N.

Free entry and exit of miners imply that active miners leave the system if $\text{Rev}/N < c_m$ and non active miners enter if $\text{Rev}/N > c_m$. Therefore, in equilibrium the number of miners satisfies $\text{Rev}/N = c_m$.

Proposition 1. *Miners' expected profit is zero. All the revenue generated from transaction fees is paid to miners. The total infrastructure employed by the system (i.e., the number of miners) is*

$$N = \frac{\text{Rev}}{c_m}.$$

Proposition 1 has several implications for the system. Free entry implies that miners will provide their service to the system at cost. It also implies that the amount of infrastructure in the system is fully determined by Rev, the total amount of transaction fees paid by the users.

[10] Miners can alter the block they attempt to mine as new transaction arrive. According to Croman et al. (2016) the computational costs associated with vetting transactions and arranging the block are orders of magnitude smaller than the costs associated with the computational efforts spent to be selected to mine a block. We thus assume that the computational efforts spent on a assembling a block are negligible.

[11] If the mined block is not legal (for example, has more than K transactions) the other miners would ignore it. An illegal blocks is not added to the consensus blockchain, and the miner who creates it is not rewarded.

[12] Under our assumptions all arriving transactions are eventually processed, and therefore the total fees per unit time of incoming transactions is equal to the total fees per unit time of processed transactions.

The analysis abstracts away from several aspects of the competition between miners. The total infrastructure employed by the system will be determined by Rev through a zero profit condition for the marginal active miner. If miners differ in their cost of mining, it is possible for miners with a cost advantage to make positive profits. If potential miners incur fixed costs to become active miners (for example, they purchase dedicated computer hardware), then the entry decision of miners will depend on their beliefs regarding future rewards and probability of winning them. In a stationary equilibrium active miners will make positive profits to allow them to recover the fixed costs of entry. Again, the main result that the total infrastructure employed by the system is determined by Rev will still hold, and further examination of such issues is left for future research.

We compare the results of Proposition 1 to empirical estimates given by Croman et al. (2016) who estimate that the total expenditure of miners during October 2015 was approximately USD5,840 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 the next block. During that period the mining reward per block was 25 bitcoins plus negligible transaction fees, or approximately USD6,000 - 7,500 (the BTC-USD exchange rate fluctuated during the month). This back of the envelope calculation suggests that miners approximately break even. The information provided by websites that offer information to potential miners about mining profitability of various crypto-currencies[13] is consistent with this observation.

The model assumes that each miner is small, and therefore cannot induce users to change their behavior. In contrast, a large miner or collection of miners can induce users to change their behavior by ignoring specific transactions. For example, a single miner that controls all the infrastructure in the system can impose a minimal transaction fee $\underline{b} > 0$ by ignoring any transaction with a lower fee, leading some users to either increase their transaction fee or leave the system. However, free entry of small independent miners implies that such behavior is not profitable even for a large miner. To see that, observe that entry by small independent miners implies zero profit for a miner who assembles a block with the K highest fee transactions. The profit of a miner who constrains himself to assemble blocks with different transactions is strictly lower. Therefore any miner, small or large, cannot make positive profits in equilibrium and finds it optimal to assemble blocks with the K highest fee transactions.

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

4.2 User behavior and equilibrium transaction fees

Bitcoin's protocol calls for new blocks to arrive according to a Poisson process of rate μ. The analysis in Section 4.1 shows that miners' optimization implies that each block processes the K pending transactions which offer the highest transaction fees. Therefore users play a queuing game where capacity is determined by the parameters μ, K and higher transaction fees imply higher processing priority. In particular, users perceive a system where the number of miners N is irrelevant (under the assumption that N is large enough for the system to be reliable and secure).

The expected time until a transaction with transaction fee b is processed is equal to the expected time until the arrival of a block in which there are fewer than K pending transactions which offer transaction fees greater than b. Analysis of the stochastic queuing model gives the following characterization of expected delay.

Lemma 2. *The expected time until a transaction is processed is a function of the block size K, the block arrival rate μ, and the load parameter $\hat{\rho} \triangleq \hat{\lambda}/\mu K \in [0,1)$, where $\hat{\lambda}$ is the arrival rate of higher priority transactions (i.e., transaction that offer greater fees), and is equal to*

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

Here, $z_0 \triangleq z_0(\hat{\rho}, K)$ is defined to be the unique solution of the polynomial equation

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

in the interval $[0,1)$.

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

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

The expected waiting time measured in blocks $W_K(\hat{\rho})$ characterized by Lemma 2 has several features. Delay increases with the load parameter $\hat{\rho}$. Its minimal value is $W_K(0) = 1$, which is the expected delay for a transaction that is processed in the next block. For low values of $\hat{\rho}$ the delay $W_K(\hat{\rho})$ is low and insensitive to $\hat{\rho}$. For values of $\hat{\rho}$

close to 1 the delay $W_K(\hat{\rho})$ is high, and $W_K(\hat{\rho})$ varies dramatically with small changes in $\hat{\rho}$.

Let $G(\cdot)$ denote the cumulative distribution function of transaction fees in some equilibrium. Consider a user i with delay cost c_i. The user chooses his transaction fee b to maximize his net reward

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

with $W(b \mid G)$ denoting the expected delay given transaction fee b and the CDF G. By Lemma 2 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 $G(b(c)) = F(c)$. Thus, for c_i we have that

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

and

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

Users' individual optimality implies the following proposition.

Proposition 3. *In the unique equilibrium of the queuing game, a user with waiting cost $c_i \in [0, \bar{c}]$ chooses a transaction fee $b(c_i)$, given by*

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

where $\rho \triangleq \lambda/\mu K \in [0, 1)$ is the load.

Payments in the Bitcoin system are determined by the equilibrium of the implied queuing game. Users with higher waiting costs pay higher transaction fees and wait less. A user with delay cost c_i pays his externality, which is the additional delay cost imposed on lower priority transactions.[14] We therefore have the following immediate corollary.

[14] To 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

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

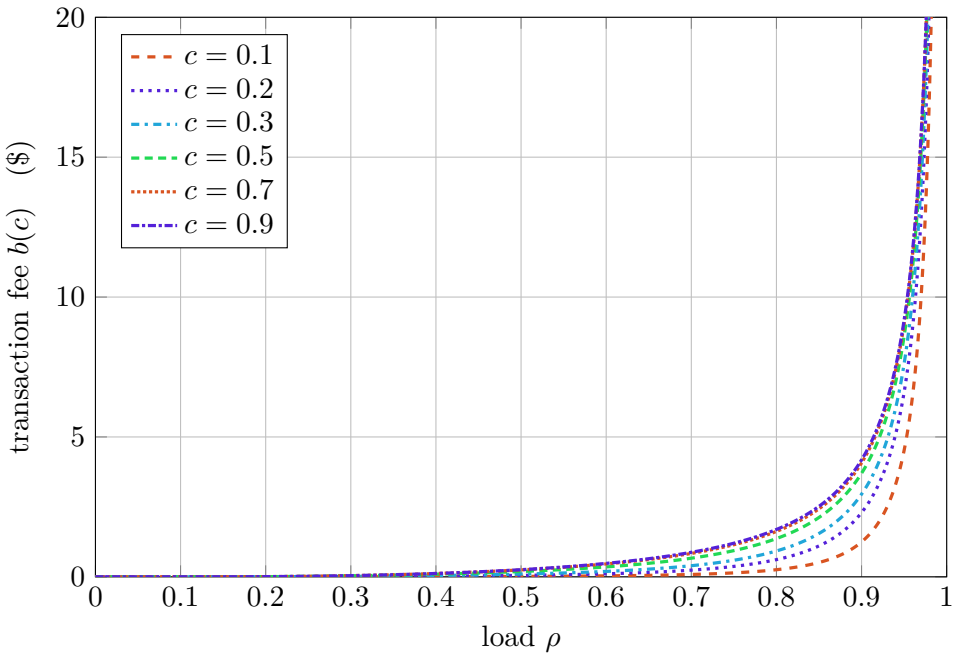

Figure 1: The dependence of equilibrium transaction fees on congestion ρ for fixed user's delay cost c. Block size is taken to be $K = 2,000$, block arrival rate $\mu = 1$ and delay costs are distributed according to $c \sim U[0,1]$.

Corollary 4. *The transaction fees paid in equilibrium coincide with the payments that result from selling priority of service in a VCG auction.*

Without an auctioneer, the Bitcoin protocol indirectly entails a priority auction. Users' bids have the VCG property that each user bids an amount equal to the externality he imposes on others. All the auction's proceeds are dissipated on competition among the service providers, i.e. the miners. In particular, the equilibrium allocation of priority is efficient. However, 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 a change in λ, μ, K will affect transaction fees.

Figure 1 and 2 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 1 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 become arbitrarily high as ρ approaches 1.

Figure 2 shows that the transaction fees increase with the user's delay cost, but do not vary much among users with high delay cost. To understand why, observe that the

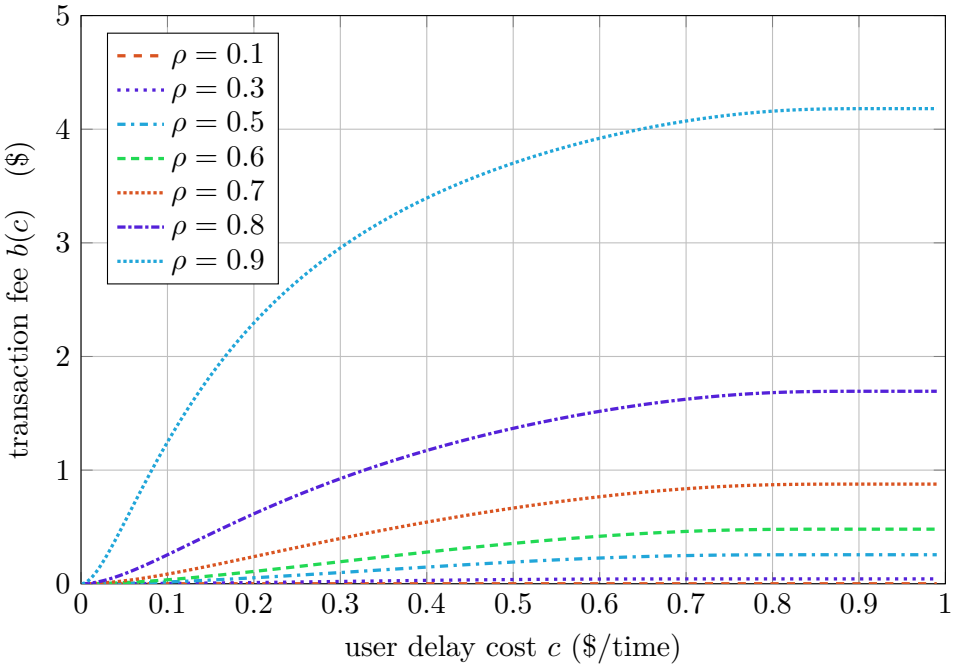

Figure 2: The dependence of equilibrium transaction fees on the user's delay cost c for fixed congestion ρ. Block size is taken to be $K = 2,000$, block arrival rate $\mu = 1$ and delay costs are distributed according to $c \sim U[0,1]$.

expected wait for a user with cost c_j 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 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 $c > 0.7$, impose the same externality to other users with delay costs $[0, 0.7]$, plus a relatively small externality to other users with delay costs $(0.7, c)$.

The following corollary of Proposition 3 establishes that all users receive positive net reward if congestion ρ is below a threshold $\bar{\rho}$ that depends on $R \cdot \mu$, K, and F.

Corollary 5. *The net reward for user i with delay cost c_i is*

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

All users receive positive net reward if $\rho < \bar{\rho}$ where $\bar{\rho}$ is the unique solution to

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

The possibility that all users are net benefactors of the system highlights its distinction from a profit maximizing monopolist. Under the latter it is always the case that some users receive no net benefit.

4.3 Total revenue and infrastructure

The analysis from the previous two subsections allows us to characterize the system under equilibrium. Theorem 6 establishes the first main result. It gives the amount of miner infrastructure in the system by calculating the total revenue per unit time. The revenue's source consists of users' transaction fees. The revenue is paid as reward to miners.

Theorem 6. *The total revenue per unit time raised from users is*

$$\mathrm{Rev}_K(\rho) = K\rho^2 \int_0^{\bar{c}} cf(c)\bar{F}(c)W_K'\left(\rho\bar{F}(c)\right)\, dc \tag{1}$$

$$= K\rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_K\left(\rho\bar{F}(c)\right)\, dc. \tag{2}$$

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

$$N = \frac{\mathrm{Rev}_K(\rho)}{c_m}.$$

The system raises revenue from users by offering them differentiated service quality based on their transaction fee, where differentiation stems from variation in delay. The delay function $W\left(b \mid G\right)$ parallels the role of the differentiated quality offered by a price-discriminating monopolist: users with low valuation (low delay cost) suffer from low service quality (long delays), forcing users with high valuation (high delay cost) to make higher payments for high service quality (short delays). Thus, in addition to paying transaction fees that fund the system's infrastructure, users incur costly delays. The following results give the total costs borne by users.

Theorem 7. *The total delay cost per unit time incurred by users is*

$$\mathrm{DelayCost}_K(\rho) = K\rho \int_0^{\bar{c}} cf(c)W_K\left(\rho\bar{F}(c)\right)\, dc. \tag{3}$$

The total cost per unit time borne by users is

$$\mathrm{TotalCost}_K(\rho) \triangleq \mathrm{Rev}_K(\rho) + \mathrm{DelayCost}_K(\rho) = K\rho \int_0^{\bar{c}} \bar{F}(c)W_K\left(\rho\bar{F}(c)\right)\, dc. \tag{4}$$

The system cannot arbitrarily set the delay $W_K(\cdot)$. Rather, the rules embedded in the protocol determine how miners process transactions in equilibrium. Miner incentives imply higher priority for transactions with a higher fee. The values μ, K determine the processing capacity of the system, and together with the arrival rate of users λ determine congestion $\rho = \lambda/\mu K$. The following corollary shows how delay cost, revenue and therefore also infrastructure, vary with ρ.

Corollary 8. *In equilibrium, if $\rho = 0$, both revenue and delay cost are zero. For all $\rho \in (0,1)$,*

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

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

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

An increase in the block arrival rate μ or a decrease in the transaction arrival rate λ will reduce the congestion ρ. Corollary 8 shows that this will reduce delay costs, revenue and therefore also available infrastructure. Proposition 3 shows that the transaction fee paid by a user also depends on the congestion level ρ. Thus, pricing, revenue and infrastructure vary with the congestion in the system.

An implication of Corollary 8 is that congestion and delays are necessary for the system to function. Low congestion ρ leads to low delay costs, as blocks are rarely full and each transaction is likely to be processed in the next block. But when blocks are rarely full users have little incentive to pay transaction fees to gain priority, and the system raises little revenue. Without sufficient revenue the number of miners N can become too small, making the system unreliable.

Figure 3 shows how delay costs and revenue in the system vary with congestion $\rho = \lambda/\mu K$. When ρ is low, users do not have to wait long for their transactions to be processed, regardless of their priority, and delay costs and revenue are low. Delay cost increases with ρ, but revenue remains small until there is significant congestion in the system. When the system becomes significantly congested users have larger incentives to gain priority, and revenue grows quickly with ρ.

Average block size in MB can be used as measure of the actual congestion in the Bitcoin system. In practice, the Bitcoin limits blocks to 1MB of data per block, which 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.

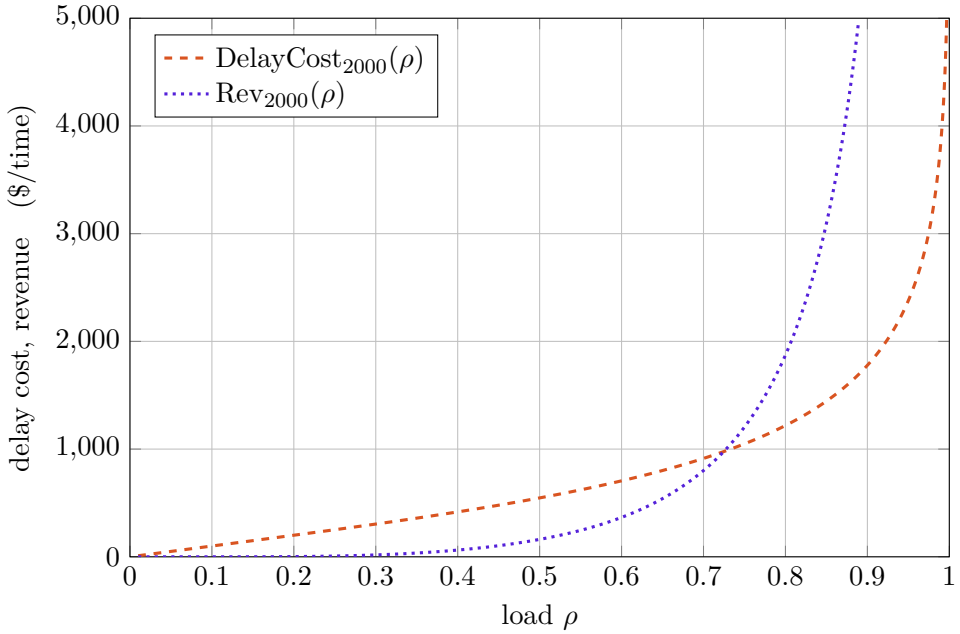

Figure 3: Revenue and delay cost for varying congestion level ρ. Delay costs are distributed according to $c \sim U[0,1]$ and the block size is $K = 2,000$.

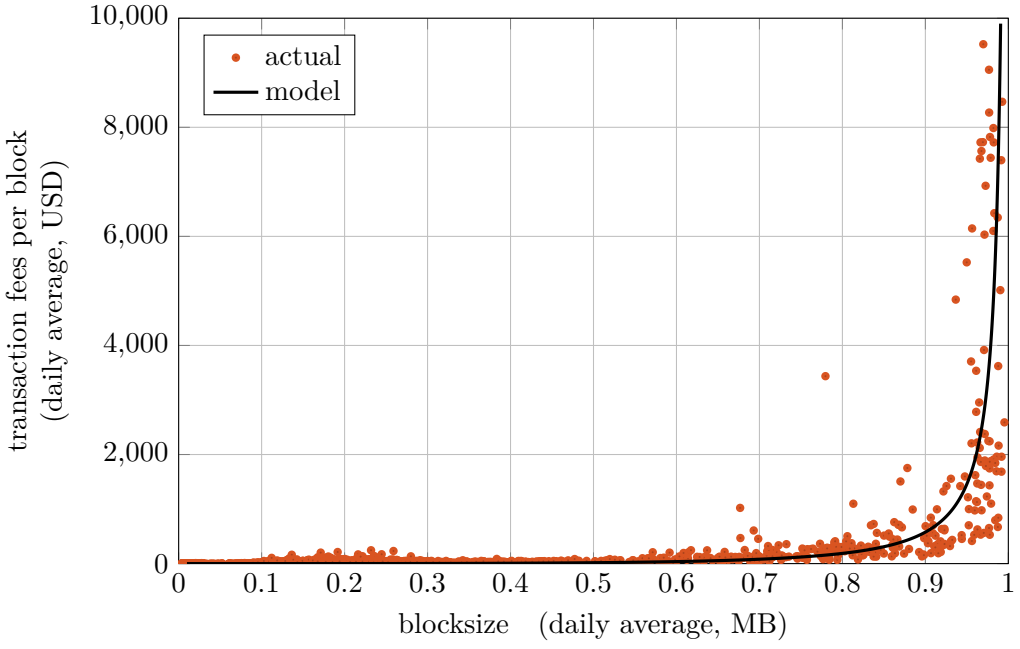

Figure 4: A comparison of actual bitcoin transaction fees per block and block size (daily averages, 4/1/2011–7/2/2017) versus model predictions.

Analogously, we interpret the average size of a block relative to the 1MB limit as a proxy for congestion ρ. Each point in Figure 4 corresponds to one day in the Bitcoin system, displaying daily average transaction fees per block and daily average block size.[15] 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 4.

Note that the solid line produced by our model matches the broad patterns in the data. Figure 4 shows that transaction fees are negligible when congestion is low. As congestion approaches 1 transaction fees increase rapidly, even though the system has excess capacity.

4.4 Behavior for large block size K and for small service rate ρ

Subsections 4.2 and 4.3 present closed form formulas for the system's attributes. This subsection considers limiting cases to better understand the dependence of the system's outcomes on its parameters. Lemma 9 studies the behavior of expected waiting time for large values of block size K. Theorem 10 studies the behavior of $\text{Rev}_K(\cdot)$ and $\text{DelayCost}_K(\cdot)$ for large values of K. Theorem 11 studies their behavior for small values of ρ. Theorem 12 considers a target revenue level, or equivalently a desired infrastructure level, and asks what is the required congestion and associated delay costs. For a fixed target revenue level, delay cost will be unboundedly high as K goes to infinity.

Lemma 9. *Holding fixed $\hat{\rho} \in (0,1)$, 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

$$W_\infty(\hat{\rho}) \triangleq \frac{1}{1 - \big(1 + \alpha(\hat{\rho})\big)e^{-\alpha(\hat{\rho})}}, \tag{5}$$

[15] Transaction fee and block size data is from `http://blockchain.info`. Each point is a daily average over the interval 4/1/2011–7/2/2017. The starting date 4/1/2011 was selected as this is roughly when the fees per block started exceeding USD1. In computing the daily average fees per block, we divide the total transaction fees in a given day by an assumed constant number $24 \times 6 = 144$ of blocks per day.

where $\alpha(\hat{\rho}) > 0$ is the unique strictly positive root of the transcendental algebraic equation

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

For $\hat{\rho} = 0$, define $W_\infty(\hat{\rho}) \triangleq 1$ to coincide with the limiting value.

Moreover, the asymptotic expected delay satisfies

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

To illustrate Lemma 9, fix λ and consider a user i with waiting cost c_i. Given block size K and block rate μ, the expected delay of user i is $\mu^{-1}W_K(\hat{\rho})$ for $\hat{\rho} = (\lambda/\mu K)\,\bar{F}(c_i)$. Consider a modification to the system that doubles the block size to $2K$ and reduces the block rate to $\mu/2$, thereby keeping the system's load ρ constant. Lemma 9 implies that, for sufficiently large K, the modification does not change the expected number of blocks until user i's transaction is processed. Because the modification doubles the wait for each block, the delay of user i roughly doubles and becomes $(\mu/2)^{-1} W_{2K}(\hat{\rho}) \approx 2\cdot\mu^{-1}W_\infty(\hat{\rho}) \approx 2\cdot\mu^{-1}W_K(\hat{\rho})$.

Lemma 9 allows us to give a simple approximate expression for revenue and delay costs when K is large. The following Theorem is an immediate corollary from Lemma 9, it shows that both revenue and delay costs grow approximately linearly with block size K when the congestion ρ is held fixed.

Theorem 10. *For a fixed load $\rho \in [0,1)$, as the block size $K \to \infty$, we have that*[16]

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

[16]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.

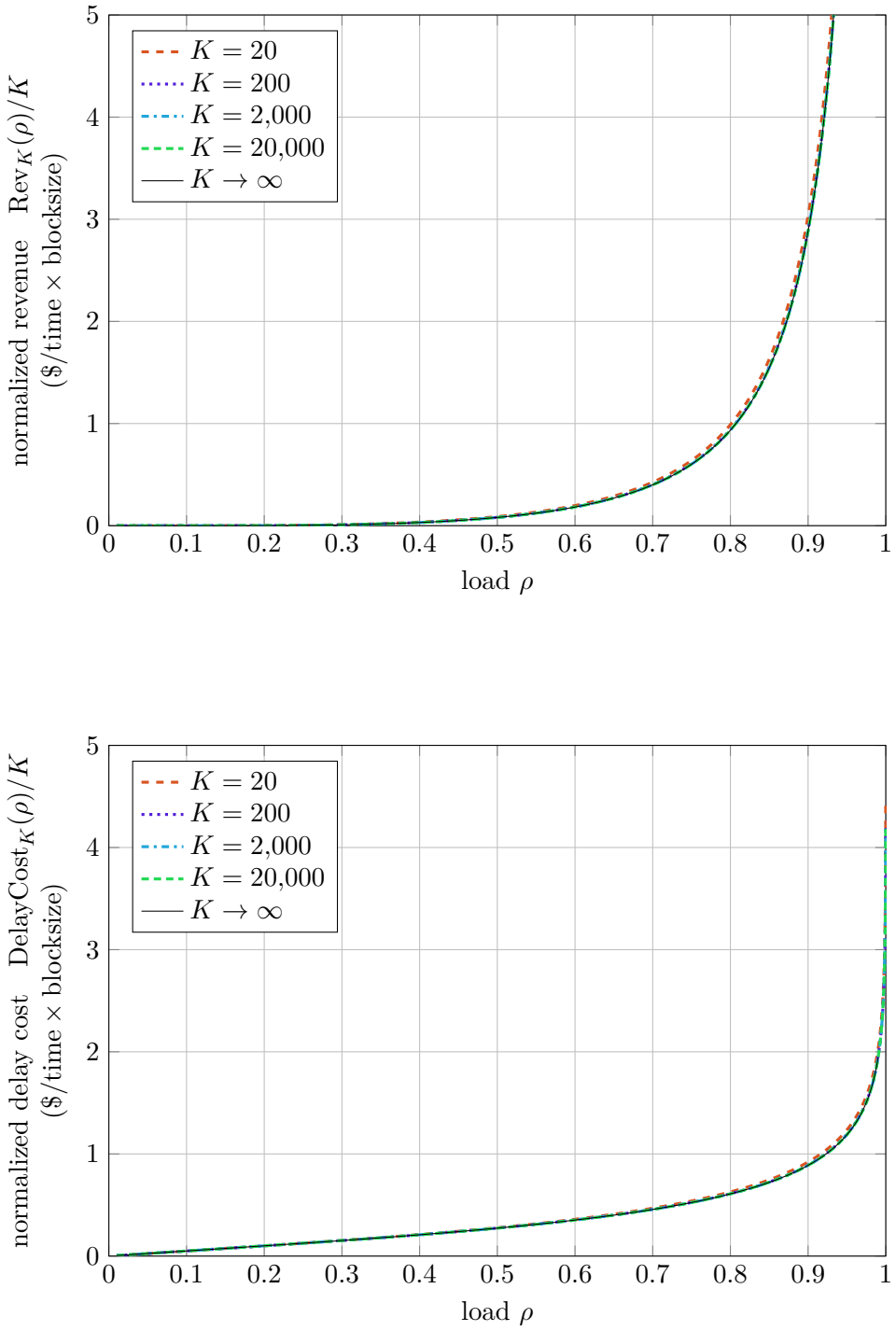

Figure 5: Normalized revenue $\mathrm{Rev}_K(\rho)/K$ *and normalized delay costs* $\mathrm{DelayCost}_K(\rho)/K$ *when* $c \sim U[0,1]$ *and* $K \in \{20,\ 200,\ 2,000,\ 20,000\}$*, compared to the limiting values obtained from the approximation using* $W_\infty(\cdot)$*.*

where

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

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.

The expressions Rev_∞ and $\mathrm{DelayCost}_\infty$ depend on ρ and F but are independent of K. These expressions allow us to approximate revenue and delay costs by a simple function of K. Figure 5 shows that this approximation is fairly good even for $K = 20$. We proceed to characterize the expressions Rev_∞ and $\mathrm{DelayCost}_\infty$ for small ρ.

Theorem 11. *As* $\rho \to 0$, *we have that*[17]

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

therefore,

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

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

[17]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.

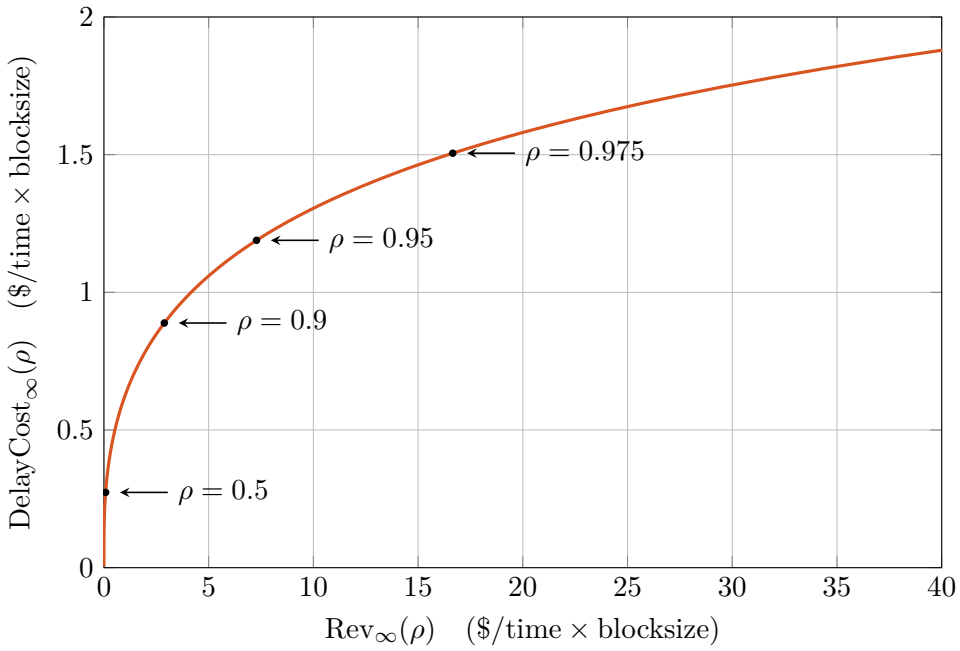

Figure 6: The parametric curve $(\text{Rev}_\infty(\rho), \text{DelayCost}_\infty(\rho))$ *for* $\rho \in [0,1)$*, describing (up to a scaling by blocksize) the achievable combinations of revenue and delay cost in for systems with large blocksize. The distribution of delay costs is taken to be* $c \sim U[0,1]$.

Theorem 11 implies that the system must impose significant delay costs on users to raise revenue. For $\rho \approx 0$ all transactions are likely to be processed in the next block regardless of their priority, because a block is unlikely to reach its maximal size. This implies that users have little incentive to choose higher transaction fees to buy priority, and the system raises little revenue. A marginal increase in user arrival rate λ or a marginal decrease in the block rate μ lead to a marginal increase in ρ. Both do little to increase revenue, because the chance a block reaches its maximal size is still low, but increase delay costs, as more transactions wait for the next block to arrive.

Figure 6 shows the possible trade-offs between revenue and delay cost in the system. It plots the curve composed of the points $(\text{Rev}_\infty(\rho), \text{DelayCost}_\infty(\rho))$ for $\rho \in [0,1)$. The figure shows that raising higher revenue requires imposing higher delay cost, and that significant delay costs are required for raising even small amounts of revenue.

By Theorem 10, the trade-off between revenue and delay costs for any block size K is a scaled version of Figure 6. Notice that Figure 6 exhibits an unfavorable trade-off for small values, implying that using a larger K would yield unfavorable results. We formally state this as the following theorem. A discussion of its implications is in Section 6.

Theorem 12. *Fix a target level of revenue* $R^* > 0$ *and a block size* K*. Define* $\text{DelayCost}_K^*(R^*)$

to be the delay cost required to achieve revenue R^*, i.e.,

$$\text{DelayCost}_K^*(R^*) \triangleq \text{DelayCost}_K\left(\text{Rev}_K^{-1}(R^*)\right),$$

where

$$\text{Rev}_K^{-1}(R^*) \triangleq \inf\left\{\rho > 0 \ : \ \text{Rev}_K(\rho) \geq R^*\right\}$$

is the minimal load required to acheive revenue R^*. Then,

$$\lim_{K\to\infty} \text{DelayCost}_K^*(R^*) = \infty.$$

5 Bitcoin as a Self Regulating System

The analysis above gives the equilibrium behavior of the system, allowing us to evaluate its performance and compare the behavior of Bitcoin to that of a monopolist. In this section we focus on the implications for the system under its current design. Section 6 discusses alternative designs

5.1 Pricing

In the distributed blockchain system, transaction fees are determined in equilibrium, as discussed in Section 4. In particular, transaction fees are not set by a profit maximizing firm.

To appreciate the novelty of the economics of Bitcoin as a payment system, compare it with an alternative owned by a monopolist. Assume that the monopolist can provide its service at zero marginal cost and at no delay,[18] and that users' benefit from the service is uniformly distributed $[\underline{R}, \bar{R}]$ independently of the delay cost c. The profit maximizing monopolist will then charge each user a fee equal to $\max\left\{\underline{R}, \bar{R}/2\right\}$ and process all transactions willing to pay the fee without delay. Two cases are of special interest: (i) when $\underline{R} = \bar{R}$, the fee is equal to the users' benefit and the users' surplus is zero; (ii) When $\bar{R}/2 > \underline{R}$ the monopolist's fee is so high that some potential users choose to avoid the system, and thereby the monopolist's price is associated with deadweight loss.

In general, a monopolistic provision of a good implies that at least some users enjoy no consumer surplus (in the example no users enjoy surplus when $\underline{R} = \bar{R}$.) Moreover, it is possible that the monopolist's price is high enough that some users will opt out of the

[18]The monopolistic firm is not constrained to use a blockchain design.

service although they would pay the cost of its provision, i.e., the monopolistic pricing entails a deadweight loss.

In contrast, the Bitcoin system can raise revenue without excluding any users and without eliminating users' surplus. The system can operate under sufficiently low congestion that allows all transactions to be processed. Even transactions that pay no fees will be processed eventually, and users need to pay transaction fees only to avoid costly delays. Transaction fees are determined by the cost of delay increase to other users, independently of users' willingness to pay for service. In particular, Corollary 5 shows that, unlike a monopolist, it is possible for all users to have a strictly positive net reward.

In addition to protecting users from monopolistic pricing, the system allocates service priority efficiently. Users with high delay costs are prioritized and processed quickly. Each transaction is charged the externality it imposes on other users (given that the system's processing capacity is held fixed). Transaction fees are not fixed by the protocol, but are determined endogenously from users' choices and their delay costs.

Setting transaction fees via equilibrium of the congestion game entails some disadvantages. First, transaction fees and total revenue depend on the congestion in the system, and may differ from the soically desired level. Congestion is a function of the system's fixed parameters μ and K as well as the transaction arrival rate λ. A judicious choice of μ, K in the protocol can induce users to pay appropriate transaction fees for a given transaction arrival rate λ. However, the transaction arrival rate may change over time, leading to undesirable transaction fees. Second, transaction fees are used to fund infrastructure provision by miners, but are determined without any regard for users' value for additional miners. Thus, equilibrium transaction fees are unlikely to match those that would be chosen by a social surplus-maximizing planner. Third, in order to raise revenue the system must impose costly delays on users. Last, instead of offering users a transparent fee, users are required to be strategic in choosing their transaction fee, depending on behavior of others and system congestion. If users fail to choose appropriate transaction fees, the system will fail to efficiently prioritize transactions.

5.2 Stability of the system

The revenue, which depends on the congestion level, determines the miners' infrastructure provision. Thus, changes in the arrival rate of transactions to the Bitcoin system will change the number of miners providing infrastructure. In particular, in the absence of congestion users pay almost no transaction fees, generating low revenue which funds only a small number of miners. When the number of miners is very low, users' assumed

indifference to the number of miners N does not hold, as the system becomes unreliable when there are few miners. With only a handful of miners the system becomes susceptible to occasional service disruptions due to network or computer failures.

Expand the analysis in Section 4 to allow potential users to avoid the system when they deem it insufficiently reliable. An equilibrium in reliability-congestion space requires that the congestion is sufficiently high that all actual users find the system reliable enough. Such an equilibrium, aside from complete abandonment of the system, need not exist.

Absence of an equilibrium can be associated with the system's collapse, but the collapse is not unavoidable. Bitcoin may survive a period of low or zero transaction fees if some miners choose to provide infrastructure to the system although the fees they receive are low. For example, it may be in the best interest of users with large coin balances to provide mining services without receiving direct compensation.

In contrast, a monopolist-run system is more stable in several ways. First, the monopolist does not rely on congestion for generating revenue, and therefore is not susceptible to the risk of implosion due to lack of congestion. Moreover, the monopolist can select the infrastructure level to maximize his long-term profits, and can offer a consistently reliable service even if demand fluctuates. Finally, the monopolist can adjust the fee he charges as demand conditions change or infrastructure requirements changes.

5.3 Social cost of the system and potential waste

The equilibrium characterized in Section 4 entails the following costs for users and miners. Users bear the cost of paying transaction fees, as well as bearing costly delays. Transaction fees are transfers to miners, and in total are equal to miner's revenue. Free entry implies that miners costs are equal to miner's revenue. Therefore, the social cost of the system is equal to the users' total cost given in Theorem 7. We proceed to discuss both costs associated with the system.

Delay costs are necessary in order to raise revenue from users. The analysis in Section 4 sheds light on the relation between the revenue and delay costs. Delay costs and revenue are both increasing with congestion, and therefore a system with higher revenue will also have higher delay costs. Consider a system with a fixed and large block size K. The system generates different combinations of revenue and delay costs for different congestion levels ρ. Theorem 10 shows the possible combinations of revenue and delay costs are approximately the combinations given by the curve in Figure 6, except that both axes need to be scaled by the block size K. Therefore, delay costs are much greater than revenue when revenue is small relative to K.

Delay costs are potentially wasteful. Delays serve no purpose other than to create incentives for users to pay higher transaction fees. But alternative system designs may be able to raise revenue from users without the need for delays. For example, a monopolist may charge all transactions a fixed fee, and process all transaction immediately.

There are several kinds of potential waste in miners' efforts. First, miners spend substantial resources in the tournament to declare a block legal and thereby receive the reward. This effort is wasteful in that it consumes real resources (such as electricity), but gives no benefit except for the random selection of a miner. Such tournaments would not be necessary in a traditional monopolist system. Second, all miners spend resources ascertaining that the transactions conform to the rules. While this effort accounts for only a small portion of miners' cost Croman et al. (2016), this duplication can be avoided in a traditional monopolist system.

Last, there can be waste in the system in that the amount of infrastructure is determined by congestion pricing, regardless of its value to users. It is possible for the Bitcoin system to operate at a congestion level that implies high revenue and a larger number of miners, even though all users prefer to have a lower number of miners.[19] A judicious choice of μ, K in the protocol can induce an appropriate number of miners for a given transaction arrival rate λ. However, the transaction arrival rate may change over time, leading to undesirable infrastructure levels.

In contrast, a monopolist-run system will avoid the tournament-waste and redundant duplication of effort, eliminate all delay costs and can set the amount of infrastructure. On the other hand, fee-setting by the monopolist may entail social cost from deadweight loss, as the monopoly price may inhibit some users from transacting in the system.

6 Design Suggestions and Alternative Pricing Mechanisms

Frustration with Bitcoin's limited throughput capacity has generated a heated discussion of protocol modifications to scale Bitcoin. In addition, hundreds of altcoins (shorthand for alternate crypto-currencies) have been proposed and more are being designed. This paper points out that under the current design of Bitcoin and other crypto-currencies, congestion is imperative to raise revenue. In this section we examine modifications of the original protocol and their implications. First, without a radical departure from Bitcoin's

[19] For example, users may be concerned with the environmental impact of the vast electricity consumption of miners.

design, some congestion is integral to the workings of Bitcoin and its scaled-up versions. We discuss how the system's parameters μ, K should be set to trade off the need for congestion against the implied waste.

Second, we pose the question of identifying the class of revenue generating mechanisms that can be implemented under the distributed system, and briefly examine some possible alternatives.

By the revelation principle, a revenue generating mechanism can be equivalently described as an incentive compatible menu $\{(b(\cdot), W(\cdot))\}$ of possible combinations of payments and delay.[20] In contrast to a traditional payment system, the distributed system cannot directly determine the delay schedule $W(\cdot)$. The protocol sets the rules of the game played by miners, and miners' equilibrium behavior generates the menu of options offered to users. Thus, the distributed system can offer a menu of options $\{(b(\cdot), W(\cdot))\}$ only if it can arise from the miners' equilibrium behavior. In particular, the protocol rules must maintain that the legality of the ledger can be verified by a third party and that miners are incentivized to process transactions rather than ignore them.

6.1 Block size increase

Transaction volume on the Bitcoin system increased over recent years with the increased popularity of the currency. As the transaction volume approached the system's capacity, high delays and fees became major concerns for the Bitcoin community. In response, a number of proposals argued for a system modification that would allow Bitcoin to support higher transaction volume, seeking to increase the block size by a factor of 2 or more.[21] If the transaction arrival rate remains fixed, such an increase in the block size should dramatically lower congestion and therefore lead to low transaction fees.[22] The following corollary gives a simple bound for any user's transaction fee as a function of congestion.

Corollary 13. *For any distribution of users' delay cost $c \sim F[0, \bar{c}]$, the transaction fee paid by a user with delay cost c_i is bounded by*

$$b(c_i) \leq \mu^{-1} c_i \cdot (W_K(\rho) - 1).$$

$\mu^{-1} c_i$ is the user's average cost for being delayed one block.

[20] Exclusion from service can be denoted by zero payment and infinite delay.

[21] For a summary of the various proposals see `https://en.wikipedia.org/wiki/Bitcoin_scalability_problem`, retrieved 7/23/2017.

[22] Currently the Bitcoin system rewards miners with newly minted coins, which account for the major part of the payment to miners.

	K (tx)	μ^{-1} (m)	λ (tx/10m)	revenue (\$/10m)	delay cost (\$/10m)
status quo	2,000	10min	1,500	\$1,205	\$1,049
big blocks	20,000	10min	1,500	\$0.002	\$750
frequent blocks	2,000	1min	1,500	\$0.0002	\$75

Table 1: Comparison between increased block size and block rate under the assumption c˜U[0,1]

A simple intuition for the result is that user c_i may choose to pay a transaction fee equal to 0, which entails the lowest service priority and a delay cost equal to $c_i \cdot \mu^{-1} W_K(\rho)$. If user c_i receives the highest service priority there is still a delay cost of $c_i \cdot \mu^{-1} W_K(0) = c_i \cdot \mu^{-1}$, as the transaction still needs to wait for the next block. The difference between the two expressions bounds the user's willingness to pay for any intermediate priority.

As an illustration, consider multiplying the block size by 10, from $K = 2,000$ transactions to $K = 20,000$ transactions. Suppose that prior to the change the system processed all potential transactions. With the increased block size by 10 the system congestion is at most $\rho = 1/10$, and $W_{20,000}(1/10) = 0.0005$. Therefore, even a user with delay cost of $\mu^{-1} c_i = \$10$ would pay at most a transaction fee of at most \$0.005.

While multiplying the block size by 10 causes revenue to collapse, it does not eliminate delay costs. For example, consider the system with $K = 2,000$ with $\mu^{-1} = 10$ minutes and an average arrival of $\lambda = 1,500$ transaction per block. Table 1 presents revenue and delay cost for the system under the assumption that users' cost of 10min delay costs per are distributed uniformly between \$0 and \$1. Increasing the block size to $K = 20,000$ causes revenue to collapse, but delay cost per block remain substantial. This is because even though virtually all transactions are processed in the next block, there is still delay until the next block arrives. If instead capacity was increased by keeping $K = 2,000$ and making blocks 10 times more frequent, then delay costs are substatially reduced.

6.2 Adjusting throughput to control congestion

The Bitcoin protocol can generate artificial congestion in the system. Congestion in the Bitcoin system is imposed by the rules of system that control the processing capacity of the system. The current Bitcoin protocol fixes the block size K and block arrival rate μ, setting a fixed capacity for the system.[23] As discussed in Section 5, this implies the congestion in the system varies with the transaction arrival rate λ, and may result in an undesirable level of revenue and infrastructure.

Consider an alternative protocol which sets the system's capacity parameters μ, K in

[23]Technological constraints may limit the values of K and μ that are feasible.

response to the transaction arrival rate λ to achieve a desired congestion level ρ. While no individual miner can affect the congestion level, the capacity parameters μ, K can affect congestion and transaction fees in a similar way to how a quantity restrictions set by a monopolist affect prices. By adjusting capacity the protocol can adjust delay cost, revenue, and infrastructure.

Holding fixed the arrival rate λ, Theorems 6 and 7 characterize the possible values of delay cost, revenue, and infrastructure that the system can achieve by adjusting μ, K. Under the assumptions specified in Section 4, adjusting μ and K is equivalent to choosing $\rho = \lambda/\mu K$ and K. Figure 7 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. All curves are (approximately) a scaled version of the curve 6 (note the logarithmic scale for the vertical axis), as implied by Theorem 10. Each curve's two main features are (i) its monotonicity – more 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 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. A comparison between the curves shows that a larger block size K is bad in that the required delay costs to raise a certain amount of revenue is larger for larger K.

A judicious choice of μ and K should provide the system with a sufficient number of miners, while minimizing the delay costs and transaction fees borne by users. For simplicity, assume that the system requires at least $\bar{N}$ miners and additional miners add no benefit. In such case, the system should raise revenue exactly equal to $R^* \triangleq c_m \cdot \bar{N}$. Users will have to pay the required revenue R^* as well as bear the necessary delay costs. As illustrated by Figure 7, the necessary delay cost to sustain the target level of revenue is increasing in K. For example, when $c \sim U[0, 1]$, adjusting congestion ρ to attain revenue of $R^* = 500$ requires delay cost of 170 when $K = 200$, of 780 when $K = 2,000$, and of 3930 when $K = 20,000$.[24] More generally, Theorem 12 shows that the required delay costs to attain a given target revenue go to infinity as K grows large.

This analysis suggests the following simple adaptation to the current protocol. First, select the smallest block size K that is feasible.[25] Second, adjust the block rate μ according

[24]The required congestion ρ to attain revenue of 500 when K is equal to 20, 200, 2,000, 20,000 is 0.98, 0.89, 0.64, and 0.38 respectively.

[25]Clearly, there are communication and other limitations that may require the block size to exceed certain levels. This paper ignores these engineering challenges.

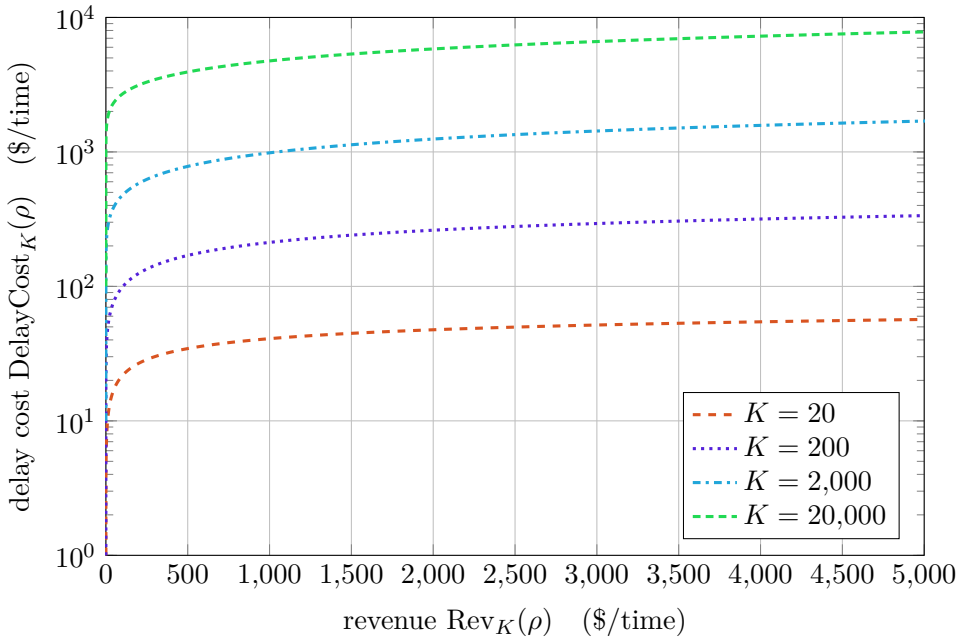

Figure 7: Possible pairs of revenue and delay cost as ρ varies, for different values of K, where delay costs are distributed according to $c \sim U[0, 1]$.

to the arrival rate of transactions λ so that ρ remains constant at the level that generates the required revenue level.

This adaptation can be implemented with small changes to the protocol. Although the parameter λ is not directly observable, it can be estimated from the ledger which is publicly available. The protocol can specify the function that calculates μ from the ledger, making μ publicly available. The calculated block arrival rate can be maintained in the same manner the block rate is currently maintained.

6.3 Mandating a fixed transaction fee paid by users to miners

A required fixed transaction fee implies that only users willing to pay the fee will contemplate using the system. The system can easily enforce this, by setting the protocol rules to render any transaction without the required fee to be illegal. The presence of a fixed fee can render congestion unnecessary. However, the determination of the desired level of the fixed fee is challenging; it depends on users' willingness to pay (which is difficult to estimate) and its aggregate should garner a sufficient, but not excessive level of miner effort.

A fixed fee can be easily included when the protocol is first introduced. However, a change of circumstances may cause the fee to be no longer appropriate. For exam-

ple, technological changes may reduce the miners' cost or change the cost of alternative transaction methods. If the protocol specifies a fixed fee, changing the transaction fee requires a switch to a new protocol. The question of designing protocol rules that allow an adjustable minimal fee is left for future research.

6.4 Direct VCG pricing

Instead of asking users to determine their transaction fees, the direct mechanism introduced by Dolan (1978) asks users to specify their delay cost. Transactions with higher delay cost are given priority. The mechanism charges a payment (transaction fee) from each user equal to the realized externality imposed on other transactions. The realized externality can be calculated after all the delayed transactions are processed (once a block is not full, no delay is imposed on any following transactions), using the delay cost specified by the delayed transactions. Dolan (1978) shows that this pricing mechanism is incentive compatible, i.e., it is optimal for each user to declare his true cost.

Section 4 showed that if each user knows the distribution of transaction fees and in response optimally chooses his transaction fee, then each user pays his expected externality. Thus, this dynamic VCG mechanism would yield the same transaction fees in expectation, without requiring users to know the distribution of transaction fees or calculate their optimal response. In addition, under this mechanism it does not matter whether users observe the current state of the system.

Although the calculation of transaction fees under this mechanism is more involved, fees can be eventually calculated from information on the ledger. Thus balances and transaction fees can be verified by miners or any third party. However, there are some difficulties in implementing this mechanism. First, payments may be unbounded, and a user cannot be charged more than his available balance. Second, while miners will receive the same reward in expectation, any transaction that is processed in a block that is not full imposes no externality and thus pays no transaction fee. If miners collect only transaction fees from transaction in a block they mined, then a miner has no incentive to mine a block that is not full.

7 Conclusion

Starting with the simple questions of who pays for the Bitcoin payment system, why and how much, this paper proceeds to analyze the economics underlying that distributed

system. Transaction fees are paid by users who wish to gain processing priority over other users and avoid delays. The system's infrastructure is provided by miners, who compete and provide their services at cost. Our analysis identifies a relation between congestion and transaction fees, which matches features of the empirical data, as seen in Figure 4. Congestion is essential for raising revenue from users to fund miners' provision of infrastructure.

The paper draws a comparison between the economic structures of the distributed Bitcoin system and a traditional electronic payment systems operated by a monopolist. Several additional differences should be noted. As opposed to traditional systems, the Bitcoin system does not require trust in any entity. However, the Bitcoin system cannot provided some services: transaction cannot be reversed in case of error or fraud, and users who lose the credentials to their account have no way of retrieving their balance. As such, Bitcoin may be more comparable to cash than to a modern electronic payment system.

Bitcoin is a monopoly run by a protocol, not by a managing organization. Familiar monopolies are run by managing organizations with discretion to determine and then change prices, offerings and rules. Monopolies are often regulated to prevent or at least mitigate their abuse of power.

Bitcoin is not regulated. It cannot be regulated. There is no need to regulate it because as a system it is committed to the protocol as is and the transaction fees it charges the users are determined by the users independently of the miners' efforts.

Bitcoin's design as an economic system is revolutionary and therefore would merit an economist's attention and scrutiny even if it had not been functional. Its apparent functionality and usefulness should further encourage economists to study this marvelous structure.

References

Athey, S., Parashkevov, I., Sarukkai, V. & Xia, J. (2016), 'Bitcoin pricing, adoption, and usage: Theory and evidence'.

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.

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. (2016), Some simple economics of the blockchain, Technical report, National Bureau of Economic Research.

Chiu, J. & Koeppl, T. (2017), 'The economics of cryptocurrencies–bitcoin and beyond'.

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'.

Dolan, R. J. (1978), 'Incentive mechanisms for priority queuing problems', *The Bell Journal of Economics* pp. 421–436.

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

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.

Gandal, N. & Halaburda, H. (2014), 'Competition in the cryptocurrency market'.

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.

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

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

Kasahara, S. & Kawahara, J. (2016), 'Priority mechanism of Bitcoin and its effect on transaction-confirmation process', *Working paper* .

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* 'Proceedings of WEIS', Vol. 2013, Citeseer.

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

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.

Posner, R. A. (1975), 'The social costs of monopoly and regulation', *Journal of political Economy* **83**(4), 807–827.

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.

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.

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

Yermack, D. (2013), Is bitcoin a real currency? an economic appraisal, Technical report, National Bureau of Economic Research.

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

A Endogenous Entry

The analysis in Section 4 assumed that the reward R is sufficiently high for all users receive positive net reward. Corollary 5 shows that all users receive positive net reward if

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

To extend the analysis to values of R for which the inequality is not satisfied, let $c^* \in [0, \bar{c}]$ be the unique solution to

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

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

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

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

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

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

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

$$N = \frac{\mathrm{Rev}_K(\rho)}{c_m}.$$

Note that these expressions coincide with their counterparts in Section 4 when $c^* = \bar{c}$.

B Proofs

Proof of Lemma 2: Consider a queueing system consisting of 'high-priority' transactions arriving according to a Poisson process rate $\hat{\lambda}$. Transactions are processed according to exponential service times with parameter μ, and block size of K. Standard analysis of bulk service systems (e.g., Section 4.6, Kleinrock 1975) yields that, if $\hat{\rho} \triangleq \hat{\lambda}/\mu K \in [0, 1)$, this queuing system is stable and the steady-state queue length for high-priority transactions has the geometric distribution

$$\pi_\ell = (1 - z_0) z_0^\ell, \quad \ell = 0, 1, \ldots,$$

where $z_0 \in [0, 1)$ is the aforementioned polynomial root.

Now, suppose a single, distinguished 'low-priority' transaction arrives, whose service is preempted by any high-priority transaction. Let T_ℓ be the expected remaining time until the low-priority transaction is processed, give $\ell \geq 0$ high-priority transactions awaiting processing. Then, we must have

$$T_\ell = \frac{1}{\mu + \hat{\lambda}} + \frac{\hat{\lambda}}{\mu + \hat{\lambda}} T_{\ell+1} + \frac{\mu}{\mu + \hat{\lambda}} \mathbb{I}_{\{\ell \geq K\}} T_{\ell-K}, \quad \ell = 0, 1, \ldots. \tag{6}$$

The first term in (6) is the expected waiting time until the next transaction arrival or service, the second term is the additional waiting time if there is a transaction arrival, while the final term is the additional waiting time if there is a service. We can rewrite (6) as

$$(\mu + \hat{\lambda}) T_\ell = 1 + \hat{\lambda} T_{\ell+1} + \mu \mathbb{I}_{\{\ell \geq K\}} T_{\ell-K}, \quad \ell = 0, 1, \ldots. \tag{7}$$

Define $T(z)$ to be the generating function

$$T(z) \triangleq \sum_{\ell=0}^{\infty} T_\ell z^\ell.$$

Applying (7), we have that

$$(\mu + \hat{\lambda}) T(z) = \frac{1}{1-z} + \hat{\lambda} z^{-1} \big[T(z) - T_0 \big] + \mu z^K T(z).$$

Solving for $T(z)$ and simplifying,

$$T(z) = \frac{\hat{\lambda} T_0 (1-z) - z}{(1-z)\big[\mu z^{K+1} - (\hat{\lambda} + \mu) z + \hat{\lambda}\big]} = \frac{\hat{\lambda} T_0 (1-z) - z}{\mu (1-z)\big[z^{K+1} - (K\hat{\rho} + 1) z + K\hat{\rho}\big]}.$$

We are interested in the steady-state average waiting time

$$\bar{W} = \sum_{\ell=0}^{\infty} \pi_\ell T_\ell = (1 - z_0) T(z_0).$$

Clearly $\bar{W} < \infty$ (since the system is stable), but, by construction, z_0 is a root of the denominator of $T(z)$. Therefore, it also must be a root of the numerator, and this implies the boundary condition

$$T_0 = \frac{z_0}{\hat{\lambda}(1 - z_0)}.$$

Furthermore, denote by $Q_K(z,\hat{\rho})$ the degree K polynomial in z defined by

$$z^{K+1} - (K\hat{\rho}+1)z + K\hat{\rho} = \big(z_0(\hat{\rho},K) - z\big)Q_K(z,\hat{\rho}), \quad \forall\ (z,\hat{\rho}) \in \mathbb{R}\times[0,1). \tag{8}$$

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. Then, we have that

$$T(z) = \frac{z_0 - z}{\mu(1-z_0)(1-z)\big[z^{K+1} - (K\hat{\rho}+1)z + K\hat{\rho}\big]} = \frac{1}{\mu(1-z_0)(1-z)Q_K(z,\hat{\rho})},$$

and the expected waiting time can be written as

$$\bar{W} = \frac{1}{\mu(1-z_0)Q_K(z_0,\hat{\rho})}.$$

In order to simplify this expression, we will apply the implicit function theorem and differentiate (8) with respect to $(z,\hat{\rho}) \in \mathbb{R}\times[0,1)$ to obtain

$$(K+1)z^K - (K\hat{\rho}+1) = -Q_K(z,\hat{\rho}) + \big(z_0(\hat{\rho},K) - z\big)\partial_z Q_K(z,\hat{\rho}), \tag{9}$$

$$-Kz + K = \partial_{\hat{\rho}} z_0(\hat{\rho},K)Q_K(z,\hat{\rho}) + \big(z_0(\hat{\rho},K) - z\big)\partial_{\hat{\rho}} Q_K(z,\hat{\rho}). \tag{10}$$

Substituting $z = z_0(\hat{\rho},K)$ into (9), we have that

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

Therefore, the expected waiting time is

$$\bar{W} = \mu^{-1} W_K(\hat{\rho}),$$

where

$$W_K(\hat{\rho}) \triangleq \frac{1}{(1-z_0)\big(1 + K\hat{\rho} - (K+1)z_0^K\big)}, \tag{11}$$

as desired.

We will now show that $W_K'(\hat{\rho}) > 0$. Differentiating (11),

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

Substituting $z = z_0(\hat{\rho}, K)$ into (9), 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,

$$\begin{aligned} W_K'(\hat{\rho}) &= K \frac{\left(Q_K(z_0,\hat{\rho}) + K(K+1)(1-z_0)z_0^{K-1}\right) - Q_K(z_0,\hat{\rho})}{(1-z_0)Q_K(z_0,\hat{\rho})^3} \\ &= \frac{K^2(K+1)z_0^{K-1}}{Q_K(z_0,\hat{\rho})^3} \\ &= K^2(K+1)z_0^{K-1}(1-z_0)^3 W_K(\hat{\rho})^3. \end{aligned} \tag{12}$$

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 (11), $W_K(\hat{\rho}) \to \infty$, and also from (12),

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

□

Proof of Proposition 3: 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

$$R - b - c_i \cdot W\left(b \mid G\right),$$

with $W\left(b \mid G\right)$ denoting the expected delay given transaction fee b and the CDF G. By Lemma 2 the expected delay is decreasing with b, and standard arguments (see Lui (1985), Hassin & Haviv (2003)) imply that $b\left(c_i\right)$ is increasing in c_i and $b\left(0\right) = 0$. Monotonicity of $b\left(\cdot\right)$ implies that $G\left(b\left(c\right)\right) = F\left(c\right)$. Therefore we have that

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

and

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

Each agent is bidding optimally if and only if

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

The first order condition implies

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

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

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

or

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

Integration together with the fact that $b\left(0\right) = 0$ yields

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

□

Proof of Corollary 5: Integration by parts yields that

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

Therefore,

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

From the last expression we have that $U\left(c_i\right)$ is decreasing with c_i, and therefore if

$$
U\left(\bar{c}\right) = R \cdot \mu - \int_0^{\bar{c}} W_K \left(\bar{\rho} \bar{F}\left(c\right)\right) dc \geq 0
$$

we have that $U\left(c_i\right) \geq 0$ for any $c_i \in [0, \bar{c}]$ given $\bar{\rho}$. Because W_K is an increasing function, if the inequality holds given $\bar{\rho}$, it also holds for any $\rho < \bar{\rho}$. □

Proof of Theorem 6: 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 established (1). For (2), we integrate by parts with

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

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.

□

Proof of Theorem 7: Transactions arrive per unit time at rate λ, and the expected delay cost per transaction is

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

Therefore, the total expected revenue per unit time is

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

as desired. The expression for total cost per unit time (4) follows by combining (2) and (3).

□

Proof of Lemma 9: The result is trivial for $\hat{\rho} = 0$.

Fix $\hat{\rho} > 0$. Define the transcendental function

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

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. Since $T''(\alpha) > 0$ for all $\alpha \geq 0$, the root must be unique. Thus,

$$T(\alpha) < 0, \quad \forall\ 0 < \alpha < \alpha(\hat{\rho}); \qquad T(\alpha) > 0, \quad \forall\ \alpha > \alpha(\hat{\rho}). \tag{13}$$

Next, we wish to prove that, as $K \to \infty$,

$$z_0(\hat{\rho}, K) = 1 - \alpha(\hat{\rho})/K + o(1/K). \tag{14}$$

Recall the polynomial defining z_0,

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

Note that

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

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

$$P_K(z) > 0, \quad \forall\ 0 \leq z < z_0(\hat{\rho}, K); \qquad P_K(z) < 0, \quad \forall\ z_0(\hat{\rho}, K) < z < 1. \tag{15}$$

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

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

Then,

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

where (13) is used for the final inequality. Thus, for all K sufficiently large, $P_K(\underline{\nu}_K) > 0$. By (15), this implies that, for all K sufficiently large, $z_0(\hat{\rho}, K) > \underline{\nu}_K$. Combining this with an analogous argument applied to $\overline{\nu}_K$, we have that, for all K sufficiently large,

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

or equivalently,

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

Since ϵ is arbitrary, we have established (14).

Finally, we are ready to analyze the asymptotic waiting time. Equation (14) 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, using the fact that $\log(1-x) = -x + O(x^2)$ as $x \to 0$,

$$\begin{aligned}\lim_{K\to\infty} K\log z_0 &= \lim_{K\to\infty} K\log\left(1 - \frac{\alpha(\hat{\rho})+\epsilon_K}{K}\right) \\ &= \lim_{K\to\infty} -\big(\alpha(\hat{\rho})+\epsilon_K\big) + O\left(\frac{\big(\alpha(\hat{\rho})+\epsilon_K\big)^2}{K}\right) = -\alpha(\hat{\rho}).\end{aligned}$$

This implies that $z_0^K \to e^{-\alpha(\hat{\rho})}$. 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,

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

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 (5), we have that

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

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 $W'_\infty(0) = 0$.

□

Proof of Theorem 10: Note that, from (2),

$$\frac{\mathrm{Rev}_K(\rho)}{K} = \rho \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_K\left(\rho\bar{F}(c)\right)\, dc. \tag{16}$$

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 9, 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 (16) 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 limits for delay cost and total cost can be established using similar dominated convergence theorem arguments. Finally, the derivative expressions can be derived directly by differentiation.

□

Proof of Theorem 11: 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

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

For the asymptotic revenue,

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

where we have used the fact that

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

Then,

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

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

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

□

Proof of Theorem 12: Define $\rho_K \triangleq \text{Rev}_K^{-1}(R^*)$, so that $\text{Rev}_K(\rho_K) = R^*$ for all K. Then,

$$\begin{aligned}\text{DelayCost}_K^*(R^*) &= \text{DelayCost}_K\left(\rho_K\right) \\ &= K\rho_K \int_0^{\bar{c}} cf(c)W_K\left(\rho_K \bar{F}(c)\right)\, dc \\ &\geq K\rho_K \mathbb{E}[c],\end{aligned}$$

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

$$\lim_{K\to\infty} K\rho_K = \infty. \tag{17}$$

We will proceed by contradiction. Fix $\epsilon > 0$. Suppose (17) does not hold. Then, there must exist a infinite subsequence

$$1 \leq K_1 < K_2 < \ldots$$

so that $K_i\rho_{K_i}$ is bounded, i.e.,

$$M \triangleq \sup_{i\geq 1}\ K_i\rho_{K_i} < \infty.$$

Define $\bar{\rho} > 0$ so that

$$W_\infty(\rho) \leq 1 + \epsilon, \quad \forall\ \rho \in (0, \bar{\rho}).$$

This is possible since $W_\infty(\rho) \to 1$ as $\rho \to 0$. Since $\rho_{K_i} \leq M/K_i \to 0$, there exists $I_1 \geq 1$ so that

$$\rho_{K_i} \leq \bar{\rho}, \quad \forall\ i \geq I_1.$$

From Lemma 9, there exists $I_2 \geq 1$ such that

$$W_{K_i}(\bar{\rho}) \leq W_\infty(\bar{\rho}) + \epsilon, \quad \forall\ i \geq I_2.$$

Then, we have that, for $i \geq \max\{I_1, I_2\}$,

$$\begin{aligned}
R^* &= \mathrm{Rev}_{K_i}\left(\rho_{K_i}\right) \\
&= K_i \rho_{K_i} \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) W_{K_i}\left(\rho_{K_i}\bar{F}(c)\right)\, dc \\
&= K_i \rho_{K_i} \int_0^{\bar{c}} \left(\bar{F}(c) - cf(c)\right) \left(W_{K_i}\left(\rho_{K_i}\bar{F}(c)\right) - 1\right)\, dc \\
&\leq M \int_0^{\bar{c}} \left|\bar{F}(c) - cf(c)\right| \cdot \left|W_{K_i}\left(\rho_{K_i}\bar{F}(c)\right) - 1\right|\, dc \\
&\leq M \int_0^{\bar{c}} \left(\bar{F}(c) + cf(c)\right) \left(W_{K_i}\left(\rho_{K_i}\bar{F}(c)\right) - 1\right)\, dc \\
&\leq 2M\mathbb{E}[c]\left(W_{K_i}\left(\rho_{K_i}\right) - 1\right) \\
&\leq 2M\mathbb{E}[c]\left(W_{K_i}\left(\bar{\rho}\right) - 1\right) \\
&\leq 2M\mathbb{E}[c]\left(W_{\infty}\left(\bar{\rho}\right) - 1 + \epsilon\right) \\
&\leq 4M\mathbb{E}[c]\epsilon.
\end{aligned}$$

Since $\epsilon > 0$ is arbitrary but $R^* > 0$, we have a contradiction.

□

Proof of Corollary 13: Using integration by parts we have that

$$\begin{aligned}
b\left(c_i\right) &= \mu^{-1} \int_0^{c_i} \left(W_K\left(\rho\bar{F}\left(c\right)\right) - W_K\left(\rho\bar{F}\left(c_i\right)\right)\right) dc \\
&\leq \mu^{-1} \int_0^{c_i} \left(W_K\left(\rho\right) - 1\right) dc \\
&\leq \mu^{-1} c_i \cdot \left(W_K\left(\rho\right) - 1\right)
\end{aligned}$$

□