An Economic Model of Blockchain: The Interplay between Transaction Fees and Security

Jiahao He, Guangyuan Zhang, Jiheng Zhang, and Rachel Q. Zhang

The Hong Kong University of Science and Technology

May 21, 2020

Abstract

A blockchain system, such as Bitcoin or Ethereum, validates electronic transactions and stores them in a chain of blocks without a central authority. *Miners* with computing power compete for the right to create blocks according to a pre-set protocol and in return earn fees paid by *users* who submit transactions. Such a system essentially operates as a single server queue with batch services based on a fee-based priority discipline, albeit with distinctive features due to the security concerns caused by *decentralization*. That is, a transaction is confirmed only after a number of additional blocks are subsequently extended to the block containing it, which complicates the interplay between miners and users. In our study, we build a model to analyze how miners' participation decisions interact with users' participation and fee decisions in equilibrium, and identify the optimal protocol design when the goal is to maximize total throughput or users' utility. Our analyses show that miners and users may end up in either a vicious or virtuous cycle, depending on the initial system state. We validate our model and analytical results using data from Bitcoin.

Key words: blockchain; queueing model; decentralization; transaction fee; security.

1 Introduction

In the past decade, *cryptocurrencies*, or digital assets that function as a means of exchange enabled by the blockchain technology, have exploded as a significant global phenomenon. While Bitcoin and Ethereum are the most commonly known, there are almost five thousand cryptocurrencies in the global market with a daily trading volume exceeding $100 billion[1]. The market capitalization of these cryptocurrencies is close to $200 billion, representing a fifty-fold increase since 2015, and the estimated number of unique active users has grown from between $0.3 \sim 1.3$ million in 2013 to over $40 million by April 2020[2]. Furthermore, these cryptocurrencies have

[1] https://coinmarketcap.com/charts/

[2] https://www.statista.com/statistics/647374/worldwide-blockchain-wallet-users/

generated an entire financial ecosystem comprising exchanges, financial derivatives and mining businesses, making them viable investment assets. For example, Bitcoin alone had generated over \$14[3] billion in mining revenue by August 2019, not to mention the additional revenue generated from the sales of mining hardware and the provision of cloud mining and remote hosting services as well as price appreciation gains. Cryptocurrencies have also led to the development of sizable new business platforms and new forms of peer-to-peer economic activities.

These economic activities differ from traditional *centralized* payment systems which are processed through a single trusted central agency. For instance, fiat cash is issued by central banks that possess reliable anti-counterfeiting technology while credit cards and digital payment services are provided by trusted financial institutions. In contrast, cryptocurrencies are based on a *decentralized* participant-level exchange following a pre-specified protocol. Since these exchanges are not conducted under the auspices of a credible agency, they are open to potential adversarial attacks and thus the underlying protocol must be able to ensure consensus in the presence of such an adversary. One protocol that has been established to ensure consensus is *proof-of-work* (PoW). PoW is the most popular consensus mechanism and supports several mainstream cryptocurrencies such as Bitcoin and Ethereum, representing over two thirds of the cryptocurrency market. In our paper, we examine the impact of PoW on user participation and fees in a cryptocurrency setting.

Under a proof-of-work mechanism, users submit transactions to a public buffer called a *mempool*. These transactions then await processing by miners, who process the transactions in batches according to the following procedure. At any time, each miner selects a number of transactions from the mempool not exceeding a pre-specified upper limit (1 megabyte for Bitcoin). Each miner then packages the selected transactions into a block and identifies an existing block on the blockchain as the new block's predecessor. To be allowed to append the new block to its predecessor, the miner needs to solve a cryptographic puzzle before other miners do. Solving the puzzle requires a specifically-designed computing machine such as GPU or ASIC and considerable computing power (Antonopoulos, 2014). In this system, miners essentially engage in a competition to earn the right for a miner's block to be accepted by other miners, as once a block is created, all its transactions are considered to be processed and hence are removed from the mempool. Given the computing power required to compete in arriving at the solution, miners equipped with more computing power have a greater chance of winning the competition.

Note that this process of appending a new block to an existing one leads to a *tree* of blocks. Blocks that are carried on the longest chain indicate the most extensive proof-of-work and are included in a public ledger. They are also an indication of honest miner activity, as opposed to blocks not carried on the longest chain, which reflect malicious work done by adversaries. Thus, we refer to miners who append their blocks to the end of the longest chain as *honest* ones and those who don't as *adversaries*. Since the decentralized nature of the process creates

[3]https://cointelegraph.com/news/bitcoin-miners-made-14-billion-to-date-securing-the-network

possibility of adversaries, blocks within the longest chain are considered confirmed only after a certain number of blocks are attached to them, referred to as the *confirmation latency*. This confirmation latency provides a means of preventing adversary activity, as it would take a disproportionally large amount of computing power for an adversary to fork a branch fast enough to overrun the honest chain and invalidate a transaction after its confirmation. Hence, as long as the computing power of honest miners exceeds that of adversarial ones, consensus can be established with a high probability. Thus, the records for transaction contained within the longest chain of blocks form a secure and irrevocable public ledger, as demonstrated by the successful transaction management of various cryptocurrencies.

As indicated, the successful operation of a blockchain system depends on the participation of honest miners, who incur costs in terms of computing power required to compete. These miners receive a *block reward*, or fixed fee, in the form of newly-issued coins, in addition to transaction fees associated with all the transactions in the block provided by the users. To prevent inflation and limit the total supply of new coins, mainstream cryptocurrencies have instituted exponentially-diminishing block reward policies. For example, Bitcoin starts with 50 bitcoins as a block reward and halves the amount every $210,000$ blocks or roughly every four years according to the block production speed. Eventually, block rewards will completely disappear, leaving miners to rely solely on transaction fees as their mining income.

On the user side, users who need their transactions processed by miners will decide to participate based on transaction queueing time, the transaction fee and the confirmation latency. To obtain a fast queueing time, users may need to pay a higher transaction fee to motivate miners to select their transactions over others. The confirmation latency to guarantee a high probability that the system will not be attacked by adversaries relies on the collective computing power of honest miners. At a high level, the higher computing power honest miners have, the harder it is for adversaries to attack the system and the shorter the confirmation latency is needed.

Thus, a user's participation and fee decisions are affected by the collective ability of honest miners to validate transactions, while a miner's computing power expenditure decision is determined by transaction fees and block rewards, although the latter will disappear eventually. A greater number of users willing to participate and pay higher fees incentivizes honest miners to provide a greater amount of computing power, leading to ultimately a healthier system and shorter confirmation latency. Thus, the interplay between users and miners in this decentralized system exhibits an intricate dynamic and one which has received little attention in the academic literature. This paper attempts to fill this void in the literature by first building an economic model that captures the unique features of cryptocurrency systems, and then using this model to analyze participant behavior and the optimal system design.

To do so, we first study the queueing dynamics in the mempool under a homogeneous user utility function to capture users' rational behavior in equilibrium, assuming away the block reward which is planned to disappear in the future. We first note that the activity of solving a cryptographic puzzle is essentially a continuous flipping of a coin with an extremely

small success probability. Thus, the number of trials needed to mine a block is geometrically distributed, which can be well approximated by an exponential random variable. To ensure a stable block production rate, the protocol dynamically adjusts the mining difficulty, i.e., the small success probability, according to the total computing power in the system. Therefore, the block processing rate remains constant even though the honest miners' participation level varies over time and the honest miners collectively work as a single server with exponential service time. If we assume transactions arrive according to a Poisson process, then the mempool essentially operates as an $M/M/1$ queue with prioritized batch service. While honest miners cannot increase the block production rate, their total computing power can affect the confirmation latency, which in turn impacts users' participation and fee decisions and thus impacts the honest miners' participation decisions. We characterize the equilibrium behavior of both users and miners and delineate the optimal system design using the model. We then verify our model with cryptocurrency data from Bitcoin and discuss our results.

Our study contributes to the existing literature on blockchain systems in several important ways. To the best of our knowledge, it is the first to incorporate the security features of such systems in analyzing the intricate interplay between users and miners, leading to important findings and insights into how a blockchain system works. Specifically, we provide three important insights.

1. Assuming that honest miners' participation is proportional to the level of transaction fees, we show how the equilibrium behavior of the users and miners is interdependent, and how the ultimate health of the system depends on the initial participation of honest mining power from an evolutionary point of view. Thus, our results suggest that it is critical for a blockchain system to attract a sufficient number of honest miners at the beginning.

2. Our findings suggest that the blockchain design to achieve maximal throughput or user welfare, in terms of mining rate, block size, and minimum transaction fee requirement, entails running the system at its full capacity, which contradicts some existing research that recommends holding back capacity in order to generate higher transaction fees. Our results further confirm a current trend in practice that suggests setting a block size as small as possible. We further obtain optimal design parameters in the presence of a block reward.

3. We analytically identify user behavior under heterogeneous user utility and conduct numerical experiments using real block rewards and transaction fees from Bitcoin. We show that classifying users into multiple types leads to a better fit of user behavior to real data.

The rest of the paper is organized as follows. After a literature review in Section 2, we introduce our detailed model in Section 3 and derive the equilibrium behavior in Section 4. The optimal system parameters are derived in Section 5. We discuss some extensions of our model in Section 6 and conduct a numerical study using real data in Section 7. The paper concludes in Section 8.

2 Literature Review

Since the inception of Bitcoin, the first blockchain system designed by and documented in Nakamoto (2008), a number of systems have evolved to enable users to establish trust in a decentralized setting. These systems seek to develop alternative mechanisms to achieve the same functionality as Bitcoin with better performance. For instance, Algorand Gilad et al. (2017) use a modification of the Byzantine agreement algorithm by Feldman and Micali (1988) to reach consensus, while Conflux Li et al. (2018) and Prism Bagaria et al. (2018) utilize a graph structure rather than a simple chain structure to store transaction contents. To reach consensus efficiently, Conflux relies on the weights of the graph vertices while Prism incorporates a sortition and group voting mechanism to improve throughput and reduce latency. Blockchain systems have also been extended to applications beyond the processing of transaction payments. For instance, Ethereum implements state machines on a Bitcoin-like system that allows users to sign and fulfill contracts in a decentralized manner (Wood (2014)). Since these blockchain systems are focused on real-world applications, the design reliability issue has only been discussed with informal arguments, e.g., the recorded transaction history can hardly be modified or all users agree on the same transaction history in a reasonable time.

Garay et al. (2015) are the first to analytically define system"reliability" using the concepts *common prefix property* and *chain quality property*. They show that Bitcoin possesses the two properties under the assumption that communication among participants is highly synchronized. In another study, Pass et al. (2017) allow for asynchronous communication with a bounded delay and find similar results. The subsequent discussion examines several streams of research related to our study of the interaction of participant decisions in a blockchain system.

2.1 Incentives and Participant Behavior

Miner Incentives and Decision Strategies: In the first study on blockchain miners' participation incentives, Kroll et al. (2013) show that the impact of transaction fees on miners' participation decisions is low when the block reward is high. They also find that transaction fees function as a reward substitute and impact miners prioritization of transactions in the mempool.

Subsequent papers explore other determinants of miners' decisions. Prat and Walter (2018) conduct an empirical study on Bitcoin and establish a model to verify that miners' decisions are influenced by the exchange rate of the cryptocurrency to US dollars. In another study, Arnosti and Weinberg (2018) show that miners' participation decisions are affected by the required investment costs for mining machine and electricity. They further derive an equilibrium of miners' decisions under asymmetric investment costs and show how cost asymmetry leads to a market oligopoly. Finally, Cong et al. (2019) examine the impact of miner collaboration decisions, (i.e., miners pool their computing power together to form a so-called mining pool in order to reduce the risk of mining) on the extent of computing power decentralization.

User Behavior and Miner Participation Decisions: Another stream of research examines users' participation decision and bids on the transaction fees and how their decisions affect the miners' participation decisions. Huberman et al. (2019) and Easley et al. (2019) characterize the equilibrium of users' strategy under a priority queueing model. The difference between the two studies is that Huberman et al. (2019) assume that the block size can be any integer and optimize the block size to achieve the maximum total transaction fees. By contrast, Easley et al. (2019) only consider block size of one. Our study extends these models by modeling the interplay of user and miner decisions and the design of blockchain system in greater detail and examining the reliability of decentralized payment systems under the possibility of an adversary attack.

Auction Mechanisms for Transaction Fees: Another stream of work related to our study focuses on the mechanism for determining transaction fees, comparing the performance of various auction mechanisms with the current "pay your bid" transaction fees mechanism. Within this area, Lavi et al. (2019), Yao (2018) and Basu et al. (2019) consider different auction mechanisms for determining transaction fees, while Lavi et al. (2019) show that their new mechanism can extract higher transaction fees from users.

Blockchain Related Research in Operations Management: Research on blockchain technology is still quite new in the field of operations management. One study in this area (Babich and Hilary (2019)) identifies five strengths and weaknesses in blockchain applications and points out several potential areas for future research. In another study, Cui et al. (2018) examine how improved traceability due to blockchain technology influences quality decisions and supply chain contracts in parallel and serial supply chains. Our study can lend potential insight into this emerging field within operations management research.

2.2 Priority Queues with Rational Behavior

In addition to the research on participant incentives, our study is related to the queueing literature, as we model the operations of Bitcoin as a priority queue and with rational participants. Within this area, Hassin (2016) provides a comprehensive review. Meanwhile, several studies on queues with priority and rational behavior are closely related. In an early study, Kleinrock (1967) investigates an unobservable $M/M/1$ queueing system in which a relative position in the queue is determined by a customer's bribe and establish the relationship between the bribe amount distribution and the average waiting time. When the cost is a heterogeneous linear combination of bribe amount and waiting time, they establish certain monotonicity in the optimal deterministic bribing under a constant average bribe constraint.

In two additional studies, Lui (1985) and Hassin (1995) analyze an $M/G/1$ queue in a similar setting while assuming a linear waiting cost and additive positive utility from receiving the service, each with a different linear coefficient distribution, to determine the impact of the

process rate on revenue and social welfare. Since a low process rate creates less competition but a greater number of entrants, they aim for an optimal balance in their respective models. Our queueing model differs from theirs in that our transaction fee affects both the position of the customers and the equilibrium behavior of the miners (server), leading to a more involved influence on the waiting time of users.

3 Model Description

In this section, we outline a model that incorporates the operational features of blockchain systems, the interplay between miners and users, and the security issue associated with the decentralized nature of the blockchain system.

In our model, transactions arrive to the system over time and are immediately placed in the mempool. Miners then select transactions from the mempool and process the transactions in blocks up to K transactions in a process referred to as hashing or mining, which is essentially a series of Bernoulli trials until one success. Thus, the number of trials needed to mine a block follows a geometric distribution and the service time can be described as an exponential random variable with rate μ. If we assume that transactions submitted by users arrive to the system according to a Poisson process with rate λ, then a blockchain system essentially operates as an $M/M/1$ queue with arrival rate λ, service rate μ, and batch size K. The service discipline is prioritized by the fee b that a user is willing to pay for his transaction to be processed, such that the higher the fee, the more quickly the transaction will be selected and processed.

For each transaction processed, a user will gain R, pay a transaction fee b, and incur a waiting cost that is an increasing convex function $c(\cdot)$ of the total waiting time, i.e., the mempool waiting time plus the confirmation latency. Thus, some potential users may not join the system and the fees users are willing to pay may be different, depending on the system status upon arrival. Hence, we model a user's behavior by (p, G), where p is the probability a potential user will join the system and G is the distribution from which the fee b is sampled. Assuming that the total market size is Λ, the arrival rate to the service system $\lambda = p\Lambda$ if every user joins the system with the same probability p. For generality, we allow the system to specify a minimum entrance fee $\underline{b}$, so G is a cumulative distribution function on $[\underline{b}, \infty)$. In Section 6, we will extend our basic model to accommodate heterogeneous users with different waiting cost functions.

In our basic model, we assume away the block reward due to its planned disappearannce in the mainstream cryptocurrencies. We further ignore potential miner incentives based on the market value of cryptocurrencies given the difficulty in modeling the volatility of these currencies. Thus, we assume that miners' total computing power is proportional to the total fee paid by users, Φ. Without loss of generality, we then use Φ to represent the total computing power provided by miners. Note that miners' total computing power Φ does not affect the block production rate μ in practice as mining difficulty is dynamically adjusted with Φ. Hence, Φ affects neither the system capacity μK nor the mempool waiting time.

By contrast, Φ does affect the confirmation latency required to guarantee that there is an overwhelmingly small probability of the system being attacked by an adversary, as represented by α, e.g., $\alpha = 10^{-4}$. Assume that the total adversary mining power is known and fixed at A at all times, as adversary also needs to acquire expensive mining machines and incur electricity cost for its operations. Then, a higher the computing power Φ makes it more difficult for an adversary to overtake the longest chain and hence a shorter confirmation latency is required. If we let z denote the number of blocks required to be extended on the same branch to ensure a sufficiently low probability α that a newly-mined block on the longest chain is confirmed, then it takes $\frac{z}{\mu}$ to confirm a block of transactions in expectation. For convenience, we refer to z as the *confirmation latency* and note that it is decreasing in miners' computing power Φ for a given α. While z is an integer in practice, we treat it as a real number in our basic model for the ease of presentation. In subsequent discussions, we incorporate integer constraints for z.

Figure 1 summarizes our basic model of how user behavior (p, G) and miners' total computing power Φ influence each other through total transaction fees and confirmation latency z. Miners' computing power Φ affects the confirmation latency z, which impacts users' waiting costs and behavior (p, G). On the other hand, users' behavior (p, G) determines the total transaction fees which in turn incentivize miners' computing power Φ.

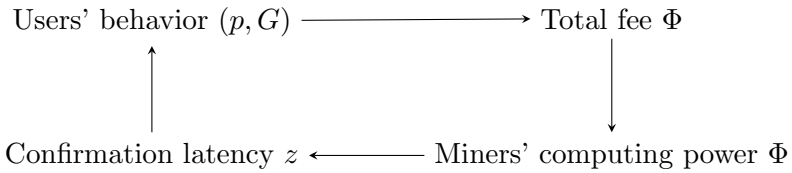


Figure 1: Interplay between users and miners through total transaction fees and confirmation latency

4 Interplay between Users and Miners

Given the complexity of a blockchain system, we focus on the equilibrium behavior as in most existing economic studies of complex systems including blockchains and use superscript "$*$" to represent equilibrium values and functions. Note that we may add different arguments to the notation as needed when deriving equilibrium behavior to make the dependence of outcomes and parameters explicit. In Section 4.1, we first investigate users' equilibrium behavior (p^*, G^*) under any given confirmation latency z, which leads to an equilibrium miner computing power $\Phi^*(z)$. In Section 4.2, we examine the minimum required confirmation latency $z^*(\Phi)$ to achieve a certain security level for given total computing power Φ. The system equilibria, defined as (z^*, Φ^*), are obtained and presented in Section 4.3.

4.1 Users' Equilibrium Behavior

To analyze users' behavior for a given confirmation latency z, we need to derive their expected total waiting time and define their utility.

4.1.1 Users' Expected Waiting Time and Utility Function

Since it takes $\frac{z}{\mu}$ to confirm a block in expectation, we need to derive only the queueing latency. As described in Section 3, queueing latency is the expected waiting time in an $M/M/1$ queue with arrival rate λ assuming all users follow the same strategy, service rate μ, block sizes of K transactions, and a fee-based priority. To derive the queueing latency, we begin by considering a user with a transaction fee b. Given a transaction fee distribution $G(\cdot)$ adopted by all users, only $1 - G(b)$ portion of all transactions will have fees above b and hence a higher priority. This means that, assuming that transactions with the same fee will be selected by miners on a first-come-first-serve basis [4], our user will be bumped in the priority only by those who arrive at a rate of $\lambda[1 - G(b)]$. Thus, his expected waiting time is the expected time it takes for an $M/M/1$ batch service queue with arrival rate $\tilde{\lambda} = \lambda[1 - G(b)]$ to become empty for the first time after his arrival, as given in the following proposition.

Proposition 1. *The expected time it will take an $M/M/1$ queue with arrival rate $\tilde{\lambda}$, service rate μ, and batch size K to become empty for the first time after a user's arrival is*

$$W_q(\tilde{\lambda}) = \begin{cases} \frac{1}{(1-\theta)[\tilde{\lambda}-\mu(K+1)\theta^K]}, & \text{if } \tilde{\lambda} < \mu K, \\ \infty, & \text{otherwise,} \end{cases} \tag{1}$$

where $\theta \in (0, 1)$ is the unique solution to $(\tilde{\lambda} + \mu)\theta - \tilde{\lambda} - \mu\theta^{K+1} = 0$.

Therefore, the expected queueing latency of a transaction with fee b is $W_q([1 - G(b)]\lambda)$ if $G(\cdot)$ is continuous. We later demonstrate that the users' equilibrium fee distribution $G^*(\cdot)$ is indeed continuous and thus the above proposition applies to our equilibrium solutions. The next proposition states how the queueing latency changes in users' behavior (p, G) and establishes that the inverse function $W_q^{-1}(\cdot)$ is well-defined, which is critical for identifying users' equilibrium behavior in Theorem 1.

Proposition 2. *$W_q(\tilde{\lambda})$ is strictly increasing convex for $\tilde{\lambda} \in [0, \mu K)$.*

Based on Proposition 2, the total expected waiting time for a transaction with fee b given other users' behavior (p, G) and confirmation latency z is

$$W(b|(p\Lambda, G), z) = W_q\left(p\Lambda[1 - G(b)]\right) + \frac{z}{\mu}. \tag{2}$$

[4] In fact, any tie-breaking rule yields an identical analysis as long as the distribution function G is continuous.

For a given z, the expected utility of a user who adopts the strategy (p, G) given everyone else's strategy (p', G') is then

$$U((p,G)|(p',G'),z) = p\int_{\underline{b}}^{\infty}[R-b-c(W(b|(p'\Lambda,G'),z))]dG(b), \tag{3}$$

assuming the utility of those users who balk is zero. For a given confirmation latency z, we define the users' equilibrium strategy (p^*, G^*) as one that maximizes a user's expected utility given that all other users apply the same strategy, i.e., it is the solution to the following equation:

$$U((p^*,G^*)|(p^*,G^*),z) = \sup_{(p,G)} U((p,G)|(p^*,G^*),z). \tag{4}$$

4.1.2 Users' Equilibrium Behavior

Before deriving users' equilibrium strategy in Theorem 1, we demonstrate in Proposition 3 that the equilibrium fee distribution $G^*(\cdot)$ is continuous and hence Proposition 1 applies.

Proposition 3. *The equilibrium fee distribution $G^*(\cdot)$ is continuous on $[\underline{b}, \infty)$.*

Intuitively, if the equilibrium fee distribution $G^*(\cdot)$ were not continuous and instead exhibited a jump at b, then a positive proportion of the transactions would incur b as a fee. However, this is impossible as an infinitesimal increase at b would allow a transaction to jump ahead of a positive proportion of the transactions and reduce its queueing latency by a non-infinitesimal amount; hence, no user would bid at b. The continuity of $G^*(\cdot)$ and monotonicity of $W_q(\cdot)$ in Proposition 2 lead to a unique equilibrium user strategy $(p^*(z), G^*(\cdot|z))$ for a given z as stated in our first theorem.

Theorem 1. *For a given z, there exists a unique equilibrium user strategy $(p^*(z), G^*(\cdot|z))$, as defined in (4), that is represented by the following:*

$$p^*(z) = \min\left\{\frac{1}{\Lambda}W_q^{-1}\left(c^{-1}(R-\underline{b})-\frac{z}{\mu}\right), 1\right\}, \tag{5}$$

$$G^*(b|z) = 1-\frac{1}{p^*(z)\Lambda}W_q^{-1}\left(c^{-1}\left(c\left(W_q(p^*(z)\Lambda)+\frac{z}{\mu}\right)-(b-\underline{b})\right)-\frac{z}{\mu}\right). \tag{6}$$

Users' equilibrium strategy as a function of the confirmation latency reveals some interesting properties in Proposition 4. For instance, a shorter confirmation latency z will attract more users to join the system, which intensifies user competition and increases queueing latency, resulting in higher transaction fees. We state these formally in our next proposition. As one will see later in our numerical study in Section 7, users' equilibrium strategy and its properties obtained from our simple utility function fits the Bitcoin data nicely.

Proposition 4. *The equilibrium solution given in Theorem 1 has the following properties.*

1. *$p^*(z)$ decreases in z;*

2. $G^*(\cdot|z)$ stochastically decreases in z;

3. For a given z, $G^*(b|z)$ is strictly increasing convex in b before it reaches 1.

By Theorem 1, a user's expected equilibrium utility then becomes:

$$U((p^*, G^*)|(p^*, G^*), z) = \max\left\{0, R - \underline{b} - c\left(W_q(\Lambda) + \frac{z}{\mu}\right)\right\}. \tag{7}$$

If $R - \underline{b} - c\left(W_q(\Lambda) + \frac{z}{\mu}\right) \geq 0$, a user can achieve a positive utility by bidding the minimum entrance fee $\underline{b}$ even if all users choose to participate, i.e., the arrival rate is Λ. In this case, all users will indeed participate, i.e., $p^*(z) = 1$, and achieve a positive utility $R - \underline{b} - c\left(W_q(\Lambda) + \frac{z}{\mu}\right)$. Otherwise, $p^*(z) < 1$ and all users will have a zero utility. Thus, users' total expected utility is also a function of z in equilibrium and can be expressed as:

$$U^*(z) = \Lambda \max\left\{0, R - \underline{b} - c\left(W_q(\Lambda) + \frac{z}{\mu}\right)\right\}. \tag{8}$$

4.1.3 Total Fee Rate

In equilibrium, transactions arrive to the system at the rate $p^*(z)\Lambda$ as all users join the system with the same probability $p^*(z)$ with fees that follow the distribution $G^*(\cdot)$. By Theorem 1, the expected total fee rate paid by users is expressed as:

$$\begin{aligned}\Phi^*(z) &= p^*(z)\Lambda \int_{\underline{b}}^{\infty} b dG^*(b) \\ &= p^*(z)\Lambda \min\left\{R, \underline{b} + c\left(W_q(\Lambda) + \frac{z}{\mu}\right)\right\} - \int_0^{p^*(z)\Lambda} c\left(W_q(\tilde{\lambda}) + \frac{z}{\mu}\right) d\tilde{\lambda}. \end{aligned} \tag{9}$$

This leads to Proposition 5, as expressed below:

Proposition 5. *$\Phi^*(z)$ is decreasing and $\ln[\Phi^*(z)]$ is decreasing concave in z.*

Note that the expected total fee rate exhibits monotonicity since the joining probability $p^*(z)$ decreases and the fee distribution $G^*(\cdot|z)$ decreases stochastically in the confirmation latency z, by Proposition 4. While $\Phi^*(z)$ is not concave in general, it is log-concave, which helps in establishing the system equilibria discussed in Section 4.3.

4.2 Confirmation Latency

To obtain the equilibrium confirmation latency z^* for a given total fee rate, or equivalently the total miner computing power, Φ, we first derive the probability of a successful attack. To do so, we follow the blockchain literature and model the attack process as a random walk. Since the exact expression for the probability of a successful attack is quite complex and difficult to analyze, we instead use a simple yet accurate approximation.

4.2.1 Probability of a Successful Attack

An attack is successful when an adversary is able to fork another chain from a confirmed block in the longest chain, referred to as double spending, and eventually overtake the longest chain following Nakamoto (2008). Since a confirmed block in the longest chain, by definition, has already been followed by at least z blocks, an adversary needs to catch up with the longest chain from at least z blocks behind. Thus, the number of blocks by which the adversary chain is behind the longest one is a random walk with a one-step transition probability $\frac{\Phi}{A+\Phi}$ if the next block is added to the longest chain and $\frac{A}{A+\Phi}$ otherwise. Thus, the probability the adversary will ever catch up within the longest chain from at least z blocks behind is given by

$$\gamma(\beta, z) = e^{-z\beta}\left[\beta^z \sum_{k=0}^{z} \frac{z^k}{k!} + \sum_{k=z+1}^{\infty} \frac{(z\beta)^k}{k!}\right], \tag{10}$$

where $\beta \triangleq \frac{A}{\Phi}$ is the adversary-to-miner computing power ratio. The higher the β and/or the smaller the z are, the higher the probability that an adversary will be able to launch a successful attack. Lemma 1 provides the lower and upper bounds for this probability.

Lemma 1. *For any given* $0 \leq \beta < 1$, $\underline{\gamma}(\beta, z) \leq \gamma(\beta, z) \leq \bar{\gamma}(\beta, z)$ *where*

$$\underline{\gamma}(\beta, z) = \frac{1}{2}\beta^z e^{z(1-\beta)},$$
$$\bar{\gamma}(\beta, z) = \left[\frac{1}{2} + \frac{1}{\sqrt{2\pi z}}\left(\frac{2}{3} + \frac{1}{1-\beta}\right)\right]\beta^z e^{z(1-\beta)}.$$

4.2.2 Confirmation Latency and Its Approximations

The confirmation latency for a given security level α and ratio of computing power β is the smallest integer z that satisfies $\gamma(\beta, z) \leq \alpha$. As mentioned, due to the complexity of $\gamma(\beta, z)$, we will look for approximations inspired by the bounds in Lemma 1. We first establish that the accuracy of confirmation latency approximation increases as α decreases.

Lemma 2. *Denote* $\underline{z}$ *and* $\bar{z}$ *as the smallest* z *such that* $\underline{\gamma}(\beta, z) \leq \alpha$ *and* $\bar{\gamma}(\beta, z) \leq \alpha$*, respectively. Then, the difference between* $\bar{z}$ *and* $\underline{z}$ *decreases as* α *becomes smaller.*

Numerical experiments for various values of β when $\alpha = 0.001$ in Table 1 demonstrate the accuracy of our approximations when we use the bounds provided in Lemma 1. From Table 1, we see that our approximations are very accurate especially when the proportion of adversary computing power β is not too high, which is in general true in reality. Similar results are observed for various values of α.

Since both the bounds work well, we will use the lower bound $\underline{\gamma}(\beta, z)$ as a proxy for $\gamma(\beta, \alpha)$ for its simplicity. Furthermore, we will treat z as a continuous variable in our basic model for cleaner presentation. We will present the analytical results when z is an integer in the extensions in Section 6 and numerical studies in Section 7. With z being a non-negative real number and

β	z	$\underline{z}$	$z-\underline{z}$	$\bar{z}$	$\bar{z}-z$
0.10	5	5	0	5	0
0.15	7	6	1	7	0
0.20	8	8	0	9	1
0.25	11	10	1	11	0
0.30	13	13	0	14	1
0.35	17	16	1	17	0
0.40	21	20	1	21	0
0.45	26	26	0	27	1
0.50	34	33	1	34	0
0.55	44	43	1	45	1
0.60	58	57	1	59	1
0.65	80	77	3	81	1
0.70	114	110	4	115	1
0.75	170	165	5	172	2
0.80	277	269	8	279	2

Table 1: Confirmation latency z and their approximations $\underline{z}$ and $\bar{z}$

using $\underline{\gamma}(\beta, z)$ as a proxy for $\gamma(\beta, \alpha)$, for a given security level α and miners' computing power Φ, the equilibrium confirmation latency $z^*(\Phi)$ satisfies $\underline{\gamma}\left(\frac{A}{\Phi}, z^*(\Phi)\right) = \alpha$ which results in the following:

$$z^*(\Phi) = \frac{\ln 2\alpha}{1 - \frac{A}{\Phi} + \ln \frac{A}{\Phi}}. \tag{11}$$

4.3 System Equilibria

We define system equilibrium as the state when (9) and (11), which describe the dependency between the confirmation latency z and computing power Φ in equilibrium, are satisfied. That is, a system equilibrium is a pair (z^*, Φ^*) that satisfies (9) and (11) simultaneously. This leads us to state the following interesting property.

Proposition 6. *$\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)$ is quasi-convex in z, where $\Phi^*(z)$ is given by* (9).

While it is harder for an adversary to attack a system successfully when the confirmation latency z is higher given a fixed level of computing power Φ by Proposition 5, the probability of a successful attack $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)$ is not monotonic in equilibrium. This is because enhancing security with a higher confirmation latency z increases users' total waiting time, discouraging users from joining the system or paying higher fees by Proposition 4. This in turn will reduce the total fee rate and equivalently lower computing power Φ^*. Thus, $\underline{\gamma}(\beta, z)$ may increase or decrease in z, depending on whether the loss of computing power dominates the enhancement of security via an increase in the confirmation latency z. As $\underline{\gamma}(\beta, z)$ is an exponential function in z for a fixed β, the former (latter) dominates and $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)$ increases (decreases) in z when z is large (small). Furthermore, the quasi-convexity of $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)$ leads directly to the possibility of the existence of up to two system equilibria. Thus a unique equilibrium occurs if and only if $\min_z \underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right) = \alpha$. This leads to our second theorem.

Theorem 2. *A system equilibrium* (z^*, Φ^*) *exists if and only if* $\min_z \underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right) \leq \alpha$*, in which case there can exist up to two equilibria.*

We next examine how the system evolves. Suppose that an equilibrium exists and the system starts with Φ^0 amount of computing power from miners. Then, users will respond with a strategy which results in a required confirmation latency $z^1 = z^*(\Phi^0)$, according to (9). Miners in turn will then adjust their computing power to $\Phi^1 = \Phi^*(z^1)$, following (11), and the process continues as $z^{n+1} = z^*(\Phi^n)$ and $\Phi^{n+1} = \Phi^*(z^{n+1})$, $n = 0, 1, \cdots$. It is obvious that, a system that begins in equilibrium remains so. Otherwise, the following proposition reveals the evolution of the blockchain system before it reaches an equilibrium.

Proposition 7. *Suppose that there exist at most two equilibria* (z_1^*, Φ_1^*) *and* (z_2^*, Φ_2^*) *with* $z_1^* \leq z_2^*$ *and* $\Phi_1^* \geq \Phi_2^*$*. Then, the series* (z^n, Φ^n) *converges to* (z_1^*, Φ_1^*) *if* $\Phi^0 > \Phi_2^*$ *and to* $(\infty, 0)$ *if* $\Phi^0 < \Phi_2^*$*.*

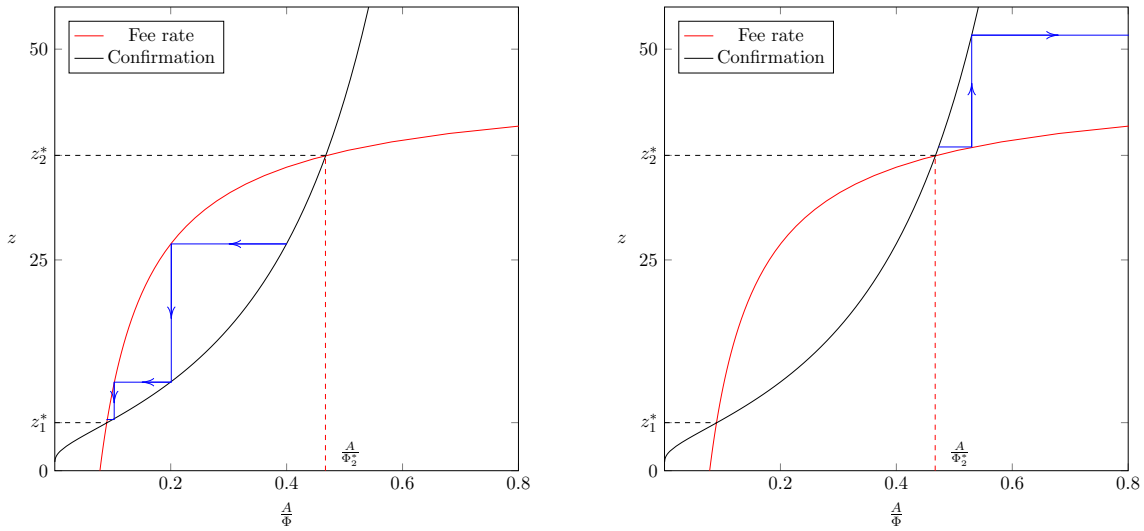
Here, z_1^* and z_2^* are the solutions to $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right) = \alpha$. That is, a system will converge to (z_2^*, Φ_2^*) only if the system starts with it, making (z_2^*, Φ_2^*) an unstable equilibrium. Figure 2 plots the evolution of series (z^n, Φ^n) when there are two equilibria: z^n as a function of $\frac{A}{\Phi^{n-1}}$ (from (11)) and Φ^n as a function of z^n (from (9)). As defined earlier, the equilibria are the solutions to (9) and (11), represented by the intersecting lines in Figure 2. If $\Phi^0 > \Phi_2^*$, i.e., the system begins with sufficient computing power, it will converge to a stable equilibrium (z_1^*, Φ_1^*) through a virtuous cycle, as seen in Figure 2(a). Otherwise, the system will be locked in a vicious cycle and eventually dissolve, as seen in Figure 2(b). A system that begins with insufficient computing power requires a long confirmation latency, which in turn discourages users from participating or being willing to pay high fees, in turn discouraging miners participation. Thus, key to a successful launch of a new blockchain system is the ability to secure a sufficient amount of initial mining power.

5 System Designs to Optimize Equilibrium Performance

While the potential market Λ, security level requirement α, and amount of adversary computing power A are all exogenous, system designers are able to decide the rate μ at which blocks are created, the number of transactions K in a block, and the entrance fee $\underline{b}$. Since each block must contain both transaction data and headers that identify the block in the entire blockchain, a small block size requires more headers and greater data storage. Coupled with other engineering concerns, we impose a lower bound K_m on the block size K and require that

$$K \geq K_m. \tag{12}$$

Furthermore, whenever a new block is mined, it is required to be broadcasted in the system, and the system capacity, or the maximum number of transactions that can be broadcasted by



(a) Virtuous cycle ($\Phi^0 > \Phi_2^*$)

(b) Vicious cycle ($\Phi^0 < \Phi_2^*$)

Figure 2: The evolution of a blockchain system for different levels of initial computing power Φ^0

the system, η, per unit time, is fixed. Thus, given the rate at which blocks are created μ, the rate at which transactions can be processed is μK and we specify that

$$\mu K \leq \eta. \tag{13}$$

Recall that, for any system design $(\mu, K, \underline{b})$, there may be two equilibria $(z_1^*, \Phi^*(z_1^*))$ and $(z_2^*, \Phi^*(z_2^*))$, and that $(z_2^*, \Phi^*(z_2^*))$ can be reached only if the system beings with it (Proposition 7). It is easy to show that $(z_1^*, \Phi^*(z_1^*))$ always yields a better performance. Thus, we focus on the equilibrium $(z_1^*, \Phi^*(z_1^*))$ for any given system design. To indicate the dependence of performance on the design, we replace $p^*(z)$ in (5) with $p^*(z|\mu, K, \underline{b})$ and replace $\Phi^*(z)$ in (9) with $\Phi^*(z|\mu, K, \underline{b})$. Since an explicit expression for z_1^* as a function of $(\mu, K, \underline{b})$ is not available, we include z as a decision along with the design parameters $(\mu, K, \underline{b})$ and require that z satisfies the following security constraint

$$\underline{\gamma}\left(\frac{A}{\Phi^*(z|\mu, K, \underline{b})}, z\right) = \alpha. \tag{14}$$

Thus, a feasible design $(\mu, K, \underline{b})$ and the corresponding confirmation latency z in equilibrium must satisfy (12), (13), and (14). We investigate optimal design with the goals of maximizing the system equilibrium throughput in Section 5.1 and users' total utility in Section 5.2.

5.1 Maximizing the Throughput

Denoting the throughput rate as $\lambda^*(z|\mu, K, \underline{b}) = p^*(z|\mu, K, \underline{b})\Lambda$, we can express the optimization problem as follows:

$$\begin{aligned} \max_{(z,\mu,K,\underline{b})\geq 0} \quad & \lambda^*(z|\mu, K, \underline{b}) \\ \text{s.t.} \quad & (12), (13), (14). \end{aligned}$$

We can now partially characterize an optimal solution in Lemma 3

Lemma 3. *Suppose that the feasible region is not empty. Then, for any feasible z, there exits $\underline{b}(z) \in \left[0, \left[R - c\left(W_q(\Lambda|\frac{\eta}{K_m}, K_m) + \frac{zK_m}{\eta}\right)\right]^+\right]$ such that $(\mu, K, \underline{b}) = \left(\frac{\eta}{K_m}, K_m, \underline{b}(z)\right)$ maximizes the throughput.*

While Lemma 3 does not rule out other potential optimal designs, it establishes that there exists at least one optimal solution such that constraints (12) and (13) are binding, assuming the feasible set is not empty. Essentially, the solution entails setting a block size as small as possible and running the system at full capacity. Indeed, we observe smaller block sizes in practice, e.g., IOTA even sets $K = 1$ (Popov (2016)). Note that our finding that an optimal system should run at full capacity $\mu K = \eta$ differs from the conclusion in Huberman et al. (2019) that a blockchain system should withhold some capacity to create longer queues. This difference in conclusions reflects differences in how we model system goals and decisions. First, while we maximize the throughput rate and specify that congestion discourages user participation, they maximize the transaction fees given a fixed arrival rate and specify that congestion motivates users to bid high fees. Second, they implicitly set $\underline{b} = 0$ and ignore the security issue, while we treat $\underline{b}$ as a decision and require a confirmation latency z to address the security requirement, both of which can influence the fees. Their model can thus be viewed as an example of a limiting case of ours when $R = \infty$ and $z = 0$, under which maximizing the throughput is equivalent to maximizing the total fees and their solution is also feasible to our problem. We further note that while they maximize the total fees by congesting the system, we do so by setting a positive $\underline{b}$ to extract users' utility such that it motivates sufficient miner participation to ensure a secure system without diminishing the throughput rate.

By Lemma 3, we can now reduce the above optimization problem to the following with decision variable z and an implicit function $\underline{b}(z)$ as

$$\max_{\underline{b}(z),z\geq 0} \quad \lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, \underline{b}(z)\right) \tag{15}$$

$$\text{s.t.} \quad \underline{\gamma}\left(\frac{A}{\Phi^*\left(z|\frac{\eta}{K_m}, K_m, \underline{b}(z)\right)}, z\right) = \alpha, \tag{16}$$

$$\underline{b}(z) \in \left[0, \left[R - c\left(W_q\left(\Lambda|\frac{\eta}{K_m}, K_m\right) + \frac{zK_m}{\eta}\right)\right]^+\right]. \tag{17}$$

If $z \geq z_0$ where $R = c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m}, K_m\right) + \frac{z_0 K_m}{\eta}\right)$, then $R \leq c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m}, K_m\right) + \frac{zK_m}{\eta}\right)$ and $\underline{b}(z) = 0$ at which the objective function $\lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, 0\right)$ decreases in z from Λ. Otherwise, all users will participate and $\lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, \underline{b}(z)\right) = \Lambda$, as discussed in Section 4.1.2. Thus, the optimal objective value (15) is exactly $\lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, 0\right)$. Since $\underline{\gamma}\left(\frac{A}{\Phi^*\left(z|\frac{\eta}{K_m}, K_m, \underline{b}\right)}, z\right)$ decreases in $\underline{b}$, feasibility of z in Problem (15)-(17) can be expressed by constraints (19) and (20) as shown below, and Problem (15)-(17) becomes equivalent to the following problem with a single decision variable z:

$$\max_{z} \quad \lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, 0\right) \tag{18}$$

$$\text{s.t.} \quad \underline{\gamma}\left(\frac{A}{\Phi^*\left(z\middle|\frac{\eta}{K_m}, K_m, \left[R - c\left(W_q(\Lambda|\frac{\eta}{K_m}, K_m) + \frac{zK_m}{\eta}\right)\right]^+\right)}, z\right) \leq \alpha, \tag{19}$$

$$\underline{\gamma}\left(\frac{A}{\Phi^*\left(z|\frac{\eta}{K_m}, K_m, 0\right)}, z\right) \geq \alpha. \tag{20}$$

We first establish that the left-hand sides of (19) and (20) are quasi-convex, as expressed in Lemma 4.

Lemma 4. *Both* $\underline{\gamma}\left(\frac{A}{\Phi^*\left(z\middle|\frac{\eta}{K_m}, K_m, \left[R - c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m}, K_m\right) + \frac{zK_m}{\eta}\right)\right]^+\right)}, z\right)$ *and* $\underline{\gamma}\left(\frac{A}{\Phi^*\left(z\middle|\frac{\eta}{K_m}, K_m, 0\right)}, z\right)$ *are quasi-convex in* z.

Assuming a non-empty feasible region, if equalities hold for (19) and (20) at $z_3 \leq z_4$ and $z_3' \leq z_4'$, respectively, then it is easy to verify that $z_3 \leq z_3' \leq z_4' \leq z_4$ (if all exist). By extension, the feasible region is either $[z_3, z_3'] \cup [z_4', z_4]$ or $[z_3, z_4]$ and all $z \leq z_0$ are optimal as long as they are feasible. We summarize the structure of the optimal solutions in the next proposition.

Proposition 8. *If the feasible region of* (18)–(20) *is not empty,* z_3 *is always optimal.*

1. *If* $z_0 < z_3$, $(z, \mu, K, \underline{b}) = (z_3, \frac{\eta}{K_m}, K_m, 0)$ *is the unique optimal solution and the optimal* $\lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, 0\right) < \Lambda$.
2. *Otherwise, the set of optimal solution is* $[z_3, z_0 \wedge z_3']$ *or* $[z_3, z_0 \wedge z_4]$ *and* $\lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, 0\right) = \Lambda$.

Note that there may be multiple solutions that lead to the maximum throughput Λ. Indeed, with a sufficiently high utility gain R, $K > K_m$, $\mu = \frac{\eta}{K}$, and $\underline{b} = 0$ may also yield the maximum throughput Λ. However, when the objective is to maximize users' total utility, there exists a unique optimal solution, as we discuss next.

5.2 Maximizing Users' Total Utility

Since the miners achieve zero utility in a completely competitive environment, maximizing users' total utility (8) and maximizing the social welfare are equivalent in our setting. Furthermore, users' total utility is also zero when $\lambda^* < \Lambda$, say $p^* < 1$, as shown in Section 4.1. Thus, the designs that maximize users' utility must be among those that achieve the throughput Λ, i.e.,

$$p^*(z|\mu, K, \underline{b}) = 1. \tag{21}$$

To indicate the dependence on the design decisions, we use $U^*(z|\mu, K, \underline{b})$ and $W_q(\Lambda|\mu, K)$ to represent users' total utility and waiting time, respectively. At $p^* = 1$, the objective function can be written as:

$$U^*(z|\mu, K, \underline{b}) = \Lambda \left[R - \underline{b} - c \left(W_q(\Lambda|\mu, K) + \frac{z}{\mu} \right) \right]. \tag{22}$$

We can now describe the problem as follows:

$$\begin{aligned} \max_{(z,\mu,K,\underline{b})\geq 0} \quad & U^*(z|\mu, K, \underline{b}) \\ \text{s.t.} \quad & (12),\ (13),\ (14),\ (21). \end{aligned}$$

The feasible region of the above problem is a subset of the feasible region when maximizing throughput. While a solution where constraints (12) and(13) are not binding may also be optimal when the goal is to maximize throughput, when the goal is to optimize user utility, a system designer must use up all system capacity and set the block size as small as possible, as indicated in Lemma 5.

Lemma 5. *Suppose that the feasible region is not empty. Then, an optimal solution must exist and be of the form* $(z^*, \mu^*, K^*, \underline{b}^*) = \left(z^*, \frac{\eta}{K_m}, K_m, \underline{b}(z^*)\right)$.

This is because, for any feasible solution $(z, \mu, K, \underline{b})$, a feasible solution $\left(z, \frac{\eta}{K_m}, K_m, \underline{b}(z)\right)$ will increase user utility by reducing the waiting cost while maintaining the total transaction fee. Thus, there exists a unique optimal design for a given z, which is different when the goal is to maximize the throughput, by Lemma 3. By Proposition 8, the problem when the goal is to maximize users' utility is reduced to the following unconstrained form:

$$\max_{z\geq 0} \quad U^* \left(z \middle| \frac{\eta}{K_m}, K_m, \underline{b}(z) \right) \tag{23}$$

$$\text{s.t.} \quad z \in [z_3, z_0 \wedge z_3']. \tag{24}$$

By (9) at $p^*(z) = 1$ and (11), the objective function $U^* \left(z \middle| \frac{\eta}{K_m}, K_m, \underline{b}(z) \right)$ can be rewritten as a function of Φ^* as

$$U^*(\Phi^*) = \Lambda R - \Phi^* - \int_0^{\Lambda} c \left(W_q \left(\tilde{\lambda} \middle| \frac{\eta}{K_m}, K_m \right) + \frac{\ln(2\alpha)}{\frac{\eta}{K_m}\left[1 - \frac{A}{\Phi^*} + \ln(\frac{A}{\Phi^*})\right]} \right) d\tilde{\lambda}. \tag{25}$$

This leads to Lemma 6.

Lemma 6. *$U^*(\Phi^*)$ is quasi-concave in Φ^* with a unique maximizer $\hat{\Phi}^*$ and $z^*(\hat{\Phi}^*) \geq z_3$.*

When the computing power Φ^* is low, the required confirmation latency is long and the waiting cost in (25) is large, resulting in lower user utility. However, when Φ^* is high, users pay high fees, which also yields lower utility. The quasi-concavity of $U^*(\Phi^*)$ leads to a unique optimal confirmation latency for Problem (23)-(24), as stated in Proposition 9.

Proposition 9. *The unique optimal $z^* = z_0 \wedge z_3' \wedge z^*(\hat{\Phi}^*)$.*

Here, z_0 enforces a throughput rate of Λ, z_3' fulfils the security requirement, and $z^*(\hat{\Phi}^*)$ is the unique maximizer of the utility function.

6 Extensions

In this section, we present the results when the confirmation latency z is an integer, extend our basic model to allow a block reward to each winning miner for creating a block besides the transaction fees, and derive users' equilibrium strategy under heterogeneous user waiting costs.

6.1 Confirmation Latency z as an Integer

We first consider the implications for our model when we treat the confirmation latency z as an integer rather than a real number. In this case, the smallest integer z such that $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right) \leq \alpha$ must satisfy $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right) \leq \alpha$ and $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z-1\right) > \alpha$, and vice versa. Theorem 3 presents the set of equilibria confirmation latencies obtained through establishing the quasi-convexity of $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)$ and $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z-1\right)$. Here, the equilibrium behaviors identified in Theorem 2 and Proposition 7, are replaced by Theorem 3 and Proposition 10, respectively. Recall that z_1^* and z_2^* are defined in Proposition 7 and the solutions to $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right) = \alpha$, given that these solutions exist. Let z_1' and z_2' be the solutions to $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z-1\right) = \alpha$ if they exist, and $\lceil z_1^* \rceil < z_1' \leq z_2' \leq z_2^*$.

Theorem 3. *A system equilibrium z^* exists if and only if $\min_{z \in \mathcal{N}^+} \underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right) \leq \alpha$, in which case the system equilibria are all the integers in $[z_1^*, z_2^*]$ if $\min_{z \in \mathcal{N}^+} \underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z-1\right) > \alpha$ or $[z_1^*, z_2^*]/[z_1', z_2']$ otherwise.*

By contrast, when z is treated as a real number, the set of equilibria, if exists, is reduced by up to two points z_1^* and z_2^* as stated in Theorem 2. As long as an equilibrium exists when z is treated as a real number, $\lceil z_1^* \rceil < z_1'$ guarantees the existence of an equilibrium when it is specified to be an integer. Analogous to Proposition 7, we now consider the series defined by $z^{n+1} = \left\lceil z^*(\Phi^{n+1}) \right\rceil = \left\lceil \frac{\ln 2\alpha}{1 - \frac{A}{\Phi^n} + \ln \frac{A}{\Phi^n}} \right\rceil$ and $\Phi^{n+1} = \Phi^*(z^{n+1})$ with initial computing power Φ^0.

When the set of equilibria is comprised of all the integers in $[z_1^*, z_2^*]/[z_1', z_2']$, the system will evolve through either a vicious or virtuous cycle, as stated in Proposition 10. The case where z_1' and z_2' do not exist is treated as a special case.

Proposition 10. *Suppose the set of equilibria z is all the integers in $[z_1^*, z_2^*]/[z_1', z_2']$.*

1. *If $z_1^* \leq \lceil z^*(\Phi^0) \rceil < z_1'$ or $z_2' < \lceil z^*(\Phi^0) \rceil \leq z_2^*$, the system begins and remains at an equilibrium.*
2. *Otherwise, the series z^n converges to $\lceil z_1^* \rceil$ if $\lceil z^*(\Phi^0) \rceil < z_1^*$, to $\lceil z_1' - 1 \rceil$ if $z_1' \leq \lceil z^*(\Phi^0) \rceil \leq z_2'$ and to ∞ in which case $\Phi^* = 0$ if $\lceil z^*(\Phi^0) \rceil > z_2^*$.*

We next derive the optimal design when the confirmation latency z is an integer and miners are able to receive a block reward.

6.2 Existence of Block Rewards and Confirmation Latency as an Integer

In our basic model, we allow miners to receive only transaction fees. Here, we consider in our model that miners also receive a reward B for each block mined or $B_0 = \mu B$ per unit time. Then, the miners' total fee, or equivalently their total computing power, in equilibrium becomes $B_0 + \Phi^*$ per unit time. Furthermore, the probability of a successful attack $\underline{\gamma}\left(\frac{A}{\Phi^*(z)+B_0}, z\right)$ is no longer quasi-convex in z in general, as shown numerically in Section 7.3. While a structural analysis of the equilibrium behavior in this context is very difficult, we are able to derive the structure of the optimal design in Propositions 11 and 12 as counterparts of the designed outlined in Propositions 8 and 9.

Here, it is easy to verify that Lemma 3 still holds, with the exception that $\underline{b}(z)$ may not be unique, so the problem with the goal of maximizing throughput rate can still be reduced to a problem with a single decision variable z.

Proposition 11. *When the feasible region of the problem with the goal of maximizing the throughput rate in the presence of a block reward is non-empty and $\tilde{z}_3$ is the smallest feasible integer, the following will hold.*

1. *If $\lambda^*\left(\tilde{z}_3|\frac{\eta}{K_m}, K_m, 0\right) < \Lambda$, then $(z, \mu, K, \underline{b}) = \left(\tilde{z}_3, \frac{\eta}{K_m}, K_m, 0\right)$ is the unique optimal solution.*
2. *Otherwise, there exists $\tilde{z}_4$, $\tilde{z}_4 \geq \tilde{z}_3$, such that $\left(z, \frac{\eta}{K_m}, K_m, \underline{b}(z)\right)$ is optimal for all $z \in [\tilde{z}_3, \tilde{z}_4]$ and $\lambda^*\left(z, \frac{\eta}{K_m}, K_m, \underline{b}(z)\right) = \Lambda$.*

In the absence of a block reward, $\tilde{z}_3 = \lceil z_3 \rceil$ where z_3 is defined as in Proposition 8. Recall that $z^*(\hat{\Phi}^*)$ is defined in Lemma 6.

Proposition 12. *In the presence of a block reward, an optimal design that maximizes users' utility must be in the form $(z, \mu, K, \underline{b}) = \left(z, \frac{\eta}{K_m}, K_m, \underline{b}(z)\right)$ where the optimal z is $\tilde{z}_4 \wedge \lceil z^*(\hat{\Phi}^*) \rceil$ and/or $\tilde{z}_4 \wedge \lfloor z^*(\hat{\Phi}^*) \rfloor$.*

6.3 Heterogeneous Users

In our final extension of our model, we consider the impact of heterogeneous users. Specifically, we follow Huberman et al. (2019) and allow users to have different waiting costs, i.e., a user's waiting cost is a linear function $c(w) = Cw$ where C follows a general distribution $F(\cdot)$ and is not required to be continuous. Since there may be infinite types of users, it is difficult to derive an equilibrium behavior for each type of user. Thus, we derive an aggregated joining probability p^* and fee distribution $G^*(\cdot)$ in equilibrium for a given z, and the resulting computing power Φ^*. This is also sufficient for our numerical study for heterogeneous users in Section 7.2 as we only have access to aggregated data from Bitcoin. We outline the aggregated equilibrium user strategy in Proposition 13. The key insight of the proposition is that the $q\%$ of users who are most patient will pay the $q\%$ lowest fees.

Proposition 13. *Let $C(q)$ and $b(q)$ be the q-quantile of the distributions $F(\cdot)$ and $G^*(\cdot)$, respectively. Then,*

$$\begin{aligned} p^* &= \max_{p' \le 1} \left\{ p' : R - \underline{b} - \int_0^{p'} C(p' - p) dW_q(p\Lambda) - C(p') \left(\frac{1+z}{\mu} \right) \ge 0 \right\}, \\ b(q) &= \underline{b} + \int_{(1-q)p^*}^{p^*} C(p^* - p) dW_q(p\Lambda), \end{aligned}$$

and the resulting computing power is

$$\Phi^* = \underline{b} p^* \Lambda + \int_0^{p^*\Lambda} p\lambda C(p^* - p) dW_q(p\Lambda).$$

7 Numerical Study

In this section, we first use data from Bitcoin to verify our users' utility function (4) in Section 7.1 and equilibrium behavior obtained in Section 4 in Section 7.2. We then continue with a comprehensive study of optimal system design when the confirmation latency is an integer and miners are able to receive a block reward in Section 7.3.

In a blockchain system, transaction records are kept by all the miners and can be obtained from any miner. To obtain our data, we crawl from a miner's website (`https://www.blockchain.com`) all the transaction data from the period of 16:28:18 Jan 5 to 23:59:59 January 31, 2018. We select this timeframe for our sample as it represents Bitcoin's most congested transaction period to date and hence contains transactions with the most significant fees. For each transaction, we obtain its *arrival time*, *size* in bytes, and *fee* in Satoshi, the monetary unit in Bitcoin, as well as the time that the block containing it was created. We extract the following system parameters from the data.

1. Arrival rate: Our dataset is comprised of a total of $6,674,639$ transaction arrivals, of which $6,669,963$ were successfully packed into a block by the end of the considered period,

yielding an effective arrival rate $\lambda^* \approx 3.0518$ per second.

2. Process rate: Within our dataset, there is a total of $4,073$ created blocks, reflecting an average mining rate $\mu \approx 0.0018754$ blocks per second, or one every 9 minutes on average.

3. From Figure 3, we see that the size of most transactions are concentrated around the median value of 226 bytes. Since the size limit of a block is 10^6 bytes, by design, the block size $K \approx 10^6/226 = 4,425$ transactions.

4. Between 2016 and 2020, miners are also rewarded $B = 12.5$ Bitcoin, or 12.5×10^8 Satoshi, newly-minted Bitcoin at the moment they add a new block to the system. To be consistent with the unit of transaction fee data, which is in Satoshi per byte, the block reward rate $B_0 = \mu B = 12.5 \times 10^8 \times 0.0018754/226 = 10,370.57522$ Satoshi per second per byte, where 226 is the median transaction size.

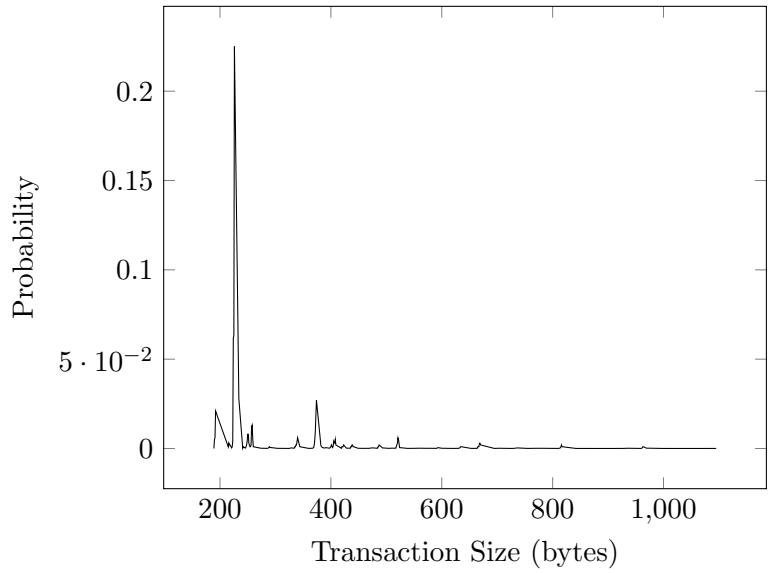


Figure 3: Distribution of the transaction size

With the above estimated parameters, we can estimate the parameters in the users' utility function in section 7.1. Since we do not have access to more data needed to estimate other parameters, we will make the following assumptions for illustration purposes.

1. $z = 6$, as suggested in Nakamoto (2008).

2. The system runs at its capacity during our sample period, i.e., $\eta \approx \mu K = 8.24622$ transactions per second.

3. $\Lambda = 5.5$ per second so that $\lambda^* = 3.0518 < \Lambda < \eta = 8.24622$, the system capacity.

4. $\alpha = 10^{-4}$ and $10 \leq K_m \leq 4,000$.

7.1 Model Validation

Note that we describe users' utility with a very simple utility function:

$$U((p,G)|(p,G),z) = p\int_{\underline{b}}^{\infty}[R-b-c(W(b|(p\Lambda,G),z))]dG(b).$$

To verify whether this function captures the users' behavior within the Bitcoin system, we further limit the waiting cost to be a linear function as $c(W) = CW$, i.e, our utility function has only two parameters (C, R). We estimate $C = 2.079$ Satoshi per second per byte and $R = 8,270$ Satoshi per byte by Theorem 1. Figure 4 plots the fee distribution $G^*(b)$ from the data marked as Empirical.

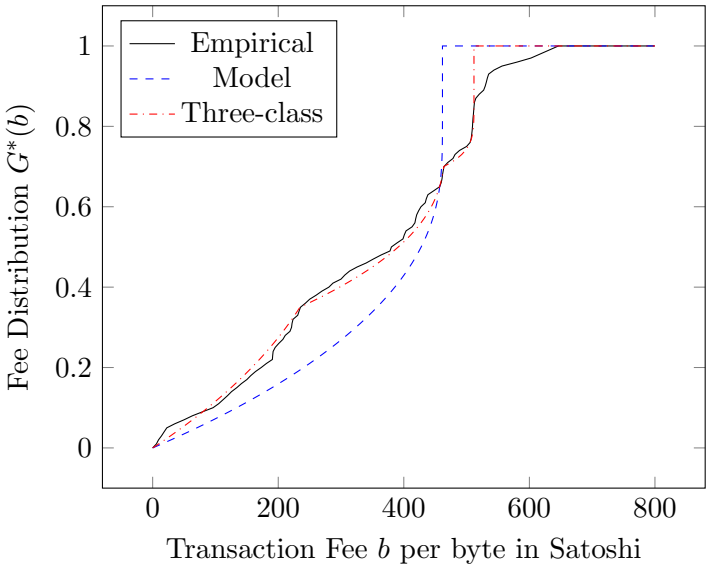


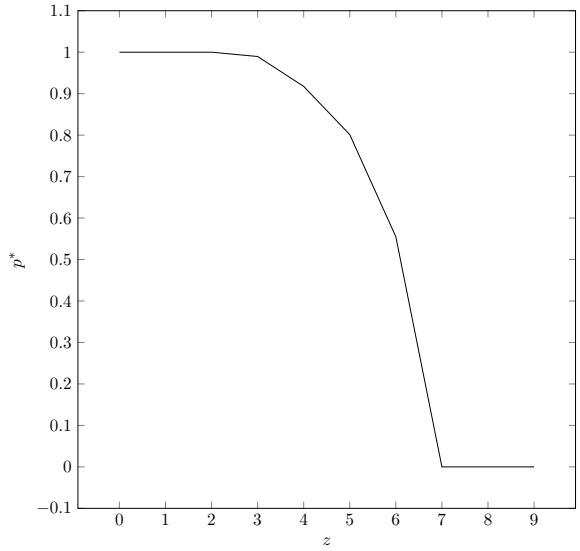
Figure 4: Distributions of fees

We then plot in Figure 4 the equilibrium fee distribution predicted by our model labeled as Model. As one can see, our simple model with only two degrees of freedom fits the data with only slight discrepancies which may be caused by user heterogeneity and behavior not captured by the model. Figure 4 further plots the fee distribution predicted by our model with three user classes (high with a waiting cost rate of 4.6281; medium, 2.424; and low, 1.3650), labeled as Three-class, and it clearly fits the data better. However, for simplicity, we will assume homogeneous users with $C = 2.079$ in our subsequent numerical study in this section.

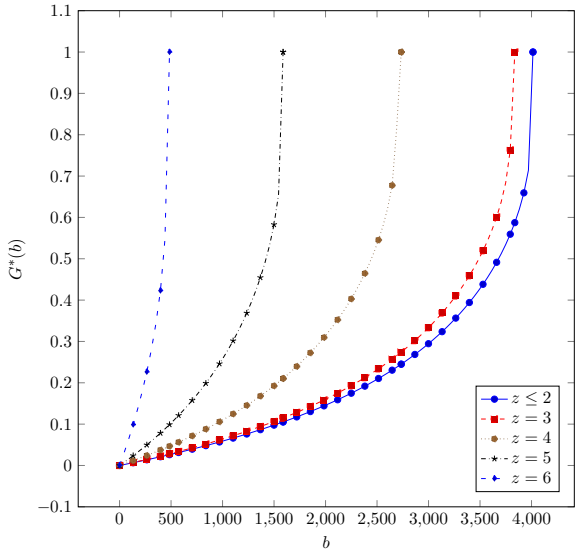
7.2 Equilibrium Behavior without Block Rewards

In this section, we illustrate the equilibrium behavior predicted by our model in section 4. Specifically, given the utility function with $(C, R) = (2.079, 8,270)$, $\mu = 0.0018754$, and $K = 4,425$, as estimated from the data, we can illustrate the extent to which users' equilibrium behavior $(p^*, G^*(b))$ and miners' computing power Φ^* change in z, predicated qualitatively in Proposition 4. From Figure 5(a), we see that p^* hovers briefly at 1 before decreasing sharply to

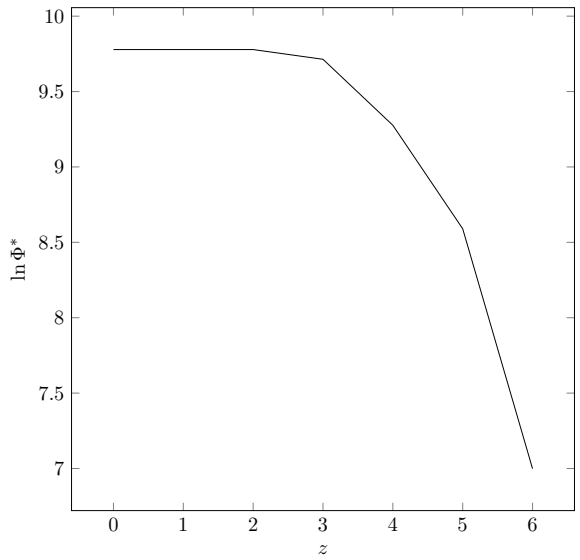
zero. Figure 5(b) illustrates how the increasing convex function $G^*(b)$ stochastically decreases in z. Finally, Figure 5(c) shows the log-concavity of Φ^*, while Figure 5(d) reveals that Φ^* at first exhibits a concave shape but becomes convex as z increases. Since Bitcoin operates at $z^* = 6$, our model yields $\Phi^*(6) \approx 1,095.3802$ Satoshi per second per byte.



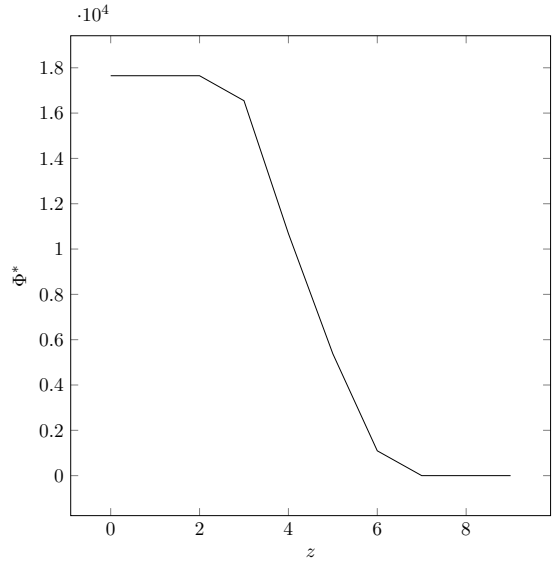
(a) Joining probability



(b) Fee distribution



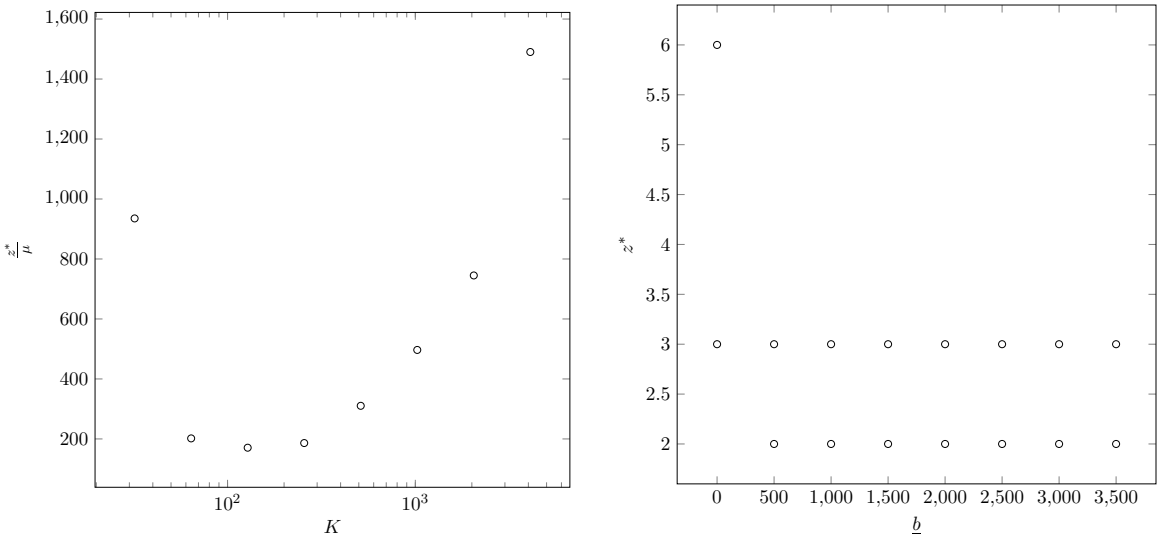
(c) Log-concavity of the computing power



(d) Computing power

Figure 5: Equilibria including user joining probability, fee distribution, and computing power as functions of z

We then illustrate the system equilibria z^* using data from Bitcoin and assume $A = 100$ and $B = 0$. From Figure 6(a), we see that the confirmation time z^*/μ first decreases and then increases in the block size K. Figure 6(b) presents the equilibria z^* for various $\underline{b}$. Note that the equilibria are achieved at $\lceil z_1^* \rceil$ and $\lfloor z_2^* \rfloor$ at $\underline{b} = 0$, while they are $\lceil z_1^* \rceil$ and $\lfloor z_1' \rfloor$ for other $\underline{b}$ values.



(a) Confirmation latency vs block size

(b) Confirmation latency vs. $\underline{b}$

Figure 6: Confirmation latency for different parameters

7.3 The Impact of a Block Reward to the Probability of a Successful Attack

We next use the Bitcoin data to examine the impact of a block reward on the probability that an adversary will launch a successful attack. At $B_0 = 10,370.57522$ Satoshi per second per byte, the total hashing power becomes $\Phi^* + B_0$. Here, we examine how the block reward affects the shape of the probability of a successful attack $\underline{\gamma}\left(\frac{A}{\Phi^*(z)+B_0}, z\right)$ as well as its sensitivity as a function of A and z.

Although we cannot observe A, we can derive the upper bound allowed for any given α. When there is a high block reward ($z^* = 6$ at which $\Phi^*(6) = 1,095.3802$ Satoshi per second per byte, $\underline{\gamma}\left(\frac{A}{\Phi^*(z)+B_0}, z\right) \leq \alpha = 10^{-4}$), $A \leq 9.81\%[\Phi^*(z) + B_0] = 1,124.8032$, indicating an extremely reliable system. By contrast, when there is no block reward, $A \leq 107.456798$, which indicates a much stronger requirement.

Figure 7 plots $\underline{\gamma}\left(\frac{A}{\Phi^*(z)+B_0}, z\right)$ as a function of z for various values of A. From Figure 7, we can see that the attack probability function is no longer convex. Figure 8 plots the equilibrium z^* in integers required as a function of α for $A = 700$ and $1,000$. Here, the equilibrium z^* is always unique for $A = 700$; for $A = 1000$, it is unique except when $4.88 \times 10^{-5} \leq \alpha \leq 4.96 \times 10^{-5}$ (between the dashed lines). As the figure shows, while z^* is quite sensitive to the hashing power of the adversary A for a given α, it is not as sensitive to the security requirement α for a given A. For instance, we see that z^* changes from 8 to 5 as α changes from 3×10^{-7} to 3×10^{-5} when $A = 700$ and from 6×10^{-6} to 10^{-4} when $A = 1,000$.

However, $\underline{\gamma}\left(\frac{A}{\Phi^*(z)+B_0}, z\right)$ is no longer quasi-convex in general in the presence of a block reward. Finally, we depict the attack probability since Bitcoin's inception in Figure 9. Here, $B_0 = 41,480$ (2008 - 2012, the initial reward), $10,370$ (2016 - 2020), $5,185$ (2020 - 2024), and

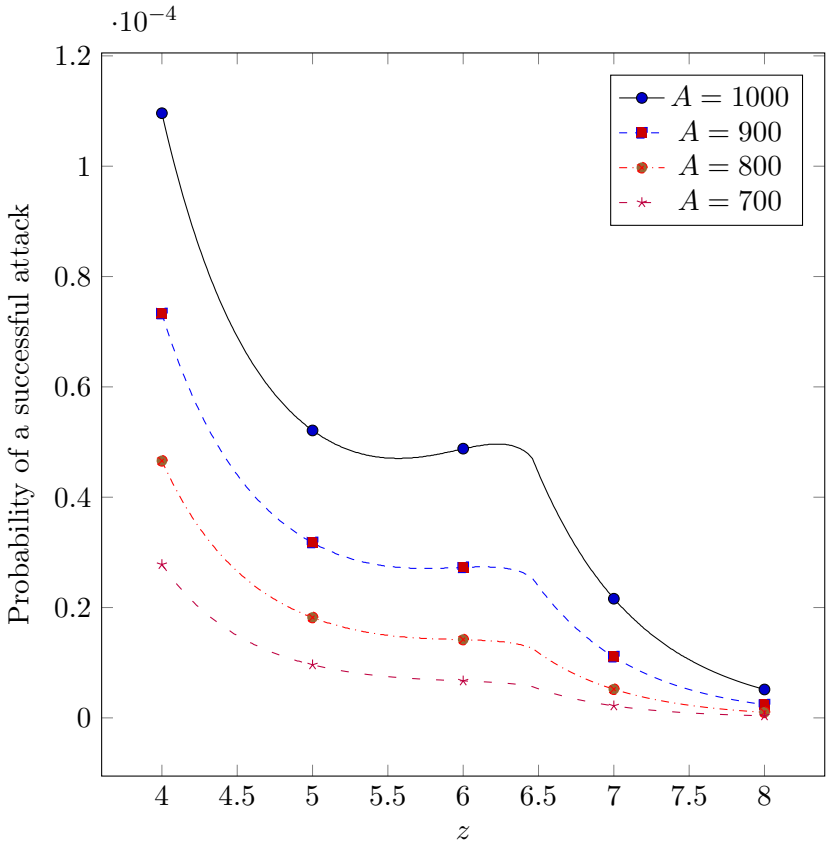
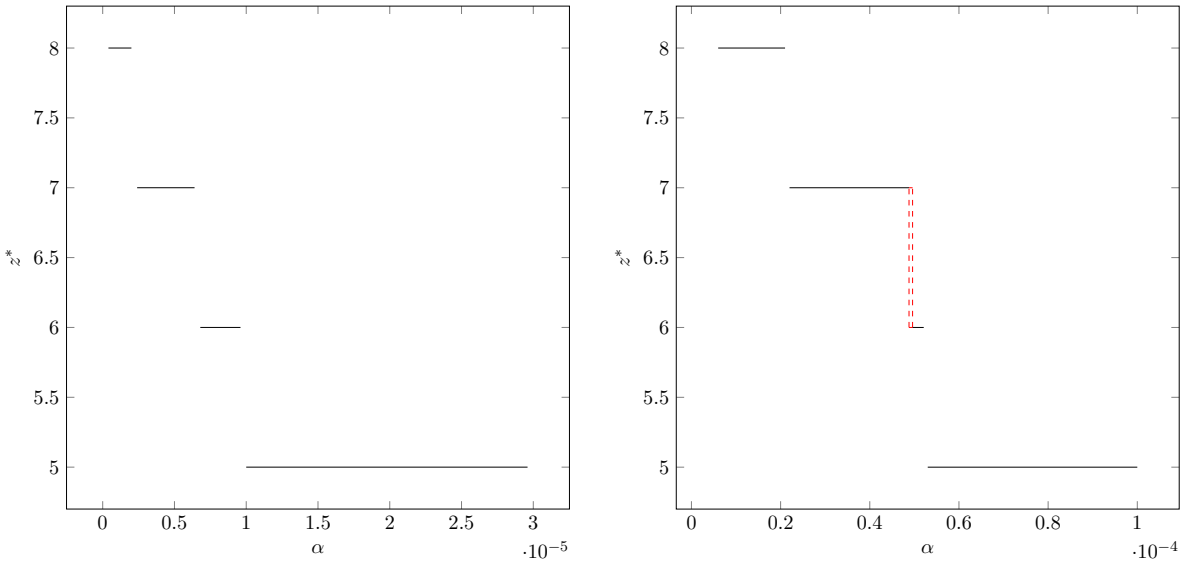


Figure 7: Probability of a successful attack as a function of z for different A

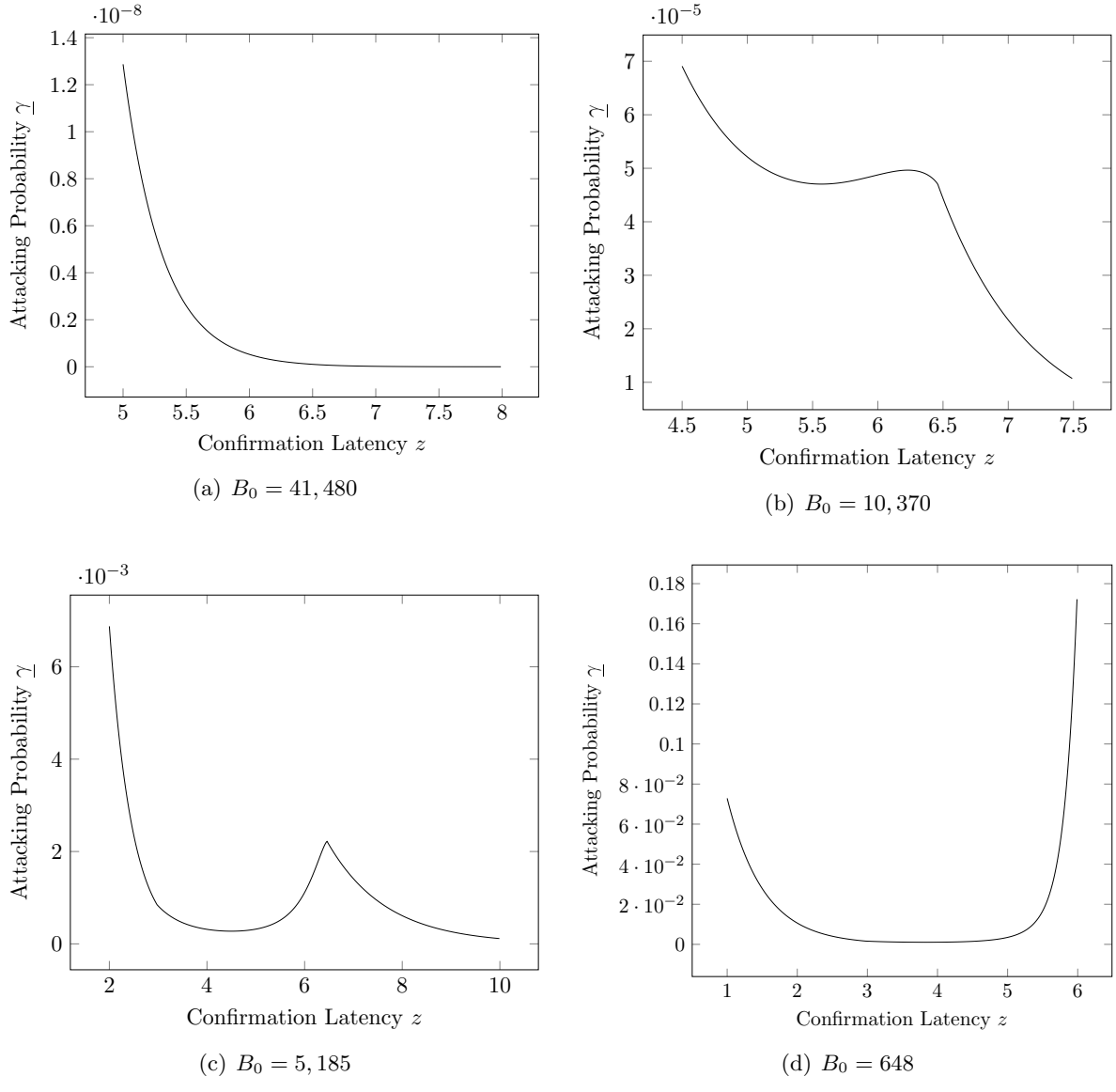


(a) $A = 700$ (b) $A = 1000$

Figure 8: All integers z^* versus α for different A

648 (2032 - 2036) Satoshi per second per byte for $A = 1,000$. Thus, we see that Proposition 6 and Theorem 2 no longer hold. From Figure 9, we identify the following properties of the probability function.

1. When the block reward is large enough, $\frac{A}{\Phi^*(z)+B_0}$ is dominated by $\frac{A}{B_0}$ and increasing z decreases $\underline{\gamma}\left(\frac{A}{\Phi^*(z)+B_0}, z\right)$. Thus, the probability of a successful attack is a decreasing function in z.
2. When the block reward is small enough, $\frac{A}{\Phi^*(z)+B_0}$ is dominated by $\frac{A}{\Phi^*(z)}$ and $\underline{\gamma}\left(\frac{A}{\Phi^*(z)+B_0}, z\right)$ is quasi-convex.
3. Otherwise, $\underline{\gamma}\left(\frac{A}{\Phi^*(z)+B_0}, z\right)$ may have several reflection points.



(a) $B_0 = 41,480$

(b) $B_0 = 10,370$

(c) $B_0 = 5,185$

(d) $B_0 = 648$

Figure 9: The probability of a successful attack for various block rewards B_0

7.4 System Design

In our final numerical analysis, we examine our system design recommendations using Bitcoin Data. Assuming that Bitcoin runs at its capacity at $\eta = 8.24622$, we calculate the optimal equilibrium z^* and the corresponding $\underline{b}(z^*)$ for K_m from 10 to 4,000 for the data-current block reward ($B_0 = 10,370$) in Figure 10 and for the May 2020 level ($B_0 = 5,185$) in Figure 11. Note that the shaded areas represent multiple equilibrium solutions that maximize the throughput while the solid lines represent the unique solutions that maximize the user utility. From Figures 10 and 11, we see that as K_m increases and $\mu^* = \frac{\eta}{K_m}$ decreases, the shaded areas become narrower. As μ^* decreases, the time to process a block and the queueing latency both increase. This discourages miner participation and pushes down $\underline{b}(z^*)$, resulting in an increase in the lower bands of the areas and a decrease in the upper bands. Due to the high block reward and low $\underline{b}(z^*)$ (at zero most of the time), our z^* that maximizes the user utility occurs more often in the upper band.

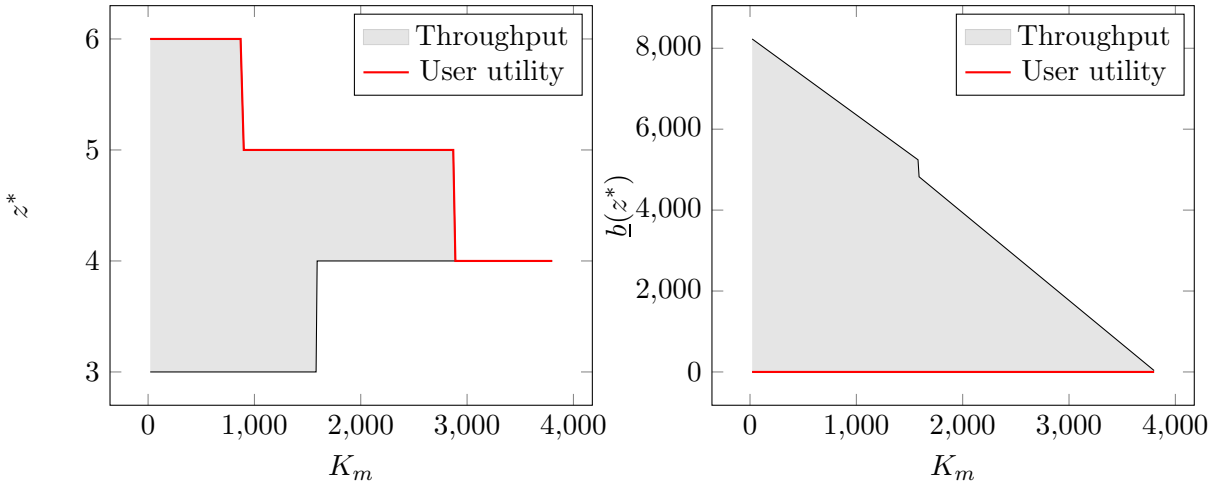


Figure 10: The optimal equilibrium z^* and the corresponding $\underline{b}(z^*)$ when $B_0 = 10,370$

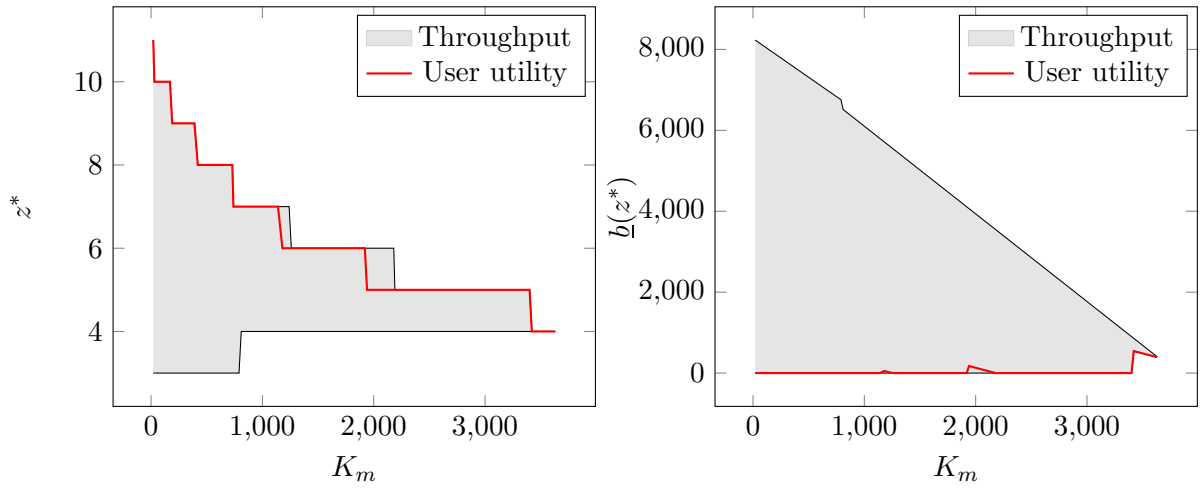


Figure 11: The optimal equilibrium z^* and the corresponding $\underline{b}(z^*)$ when $B_0 = 5,185$

8 Conclusions

In this paper, we develop a blockchain model to describe the interplay between user participation, transaction fees, and miner participation. Our model focuses on the transaction blockchain system used by cryptocurrencies and includes system security, as reflected by confirmation latency. We also test our model using a data sample from Bitcoin. Our results from our numerical analysis show that the Bitcoin data validates our simple user utility function. In particular, we analyze the equilibrium behavior of users and miners and identify the optimal design of system parameters for maximizing both system throughput and users' total utility under limited system capacity.

Our equilibrium analysis indicates the system must attract a sufficient level of initial computing power to encourage users participation and willingness to pay sufficiently high transaction fees. Doing so creates enough miner participation to stabilize the system and allow it to reach equilibrium. In this sense, the total transaction fee, which is assumed to be proportional to the total computing power and thus reflects system maintenance energy costs, can be seen as the cost for decentralization, and it is finite.

Our analysis also provides insight into the optimal system design. In particular, we find that the optimal design entails setting the block size as small as possible, consistent with practice. We also find that the optimal design entails running a system at full capacity, which differs from previous research that does not model the goal of maximizing users total utility. Smaller block sizes lead to faster processing speed, which attracts more users and hence more computing power to the system. Indeed, when it has exhausted the whole user market, a blockchain system may still be able to increase miners' revenue with a higher entrance fee while maintaining user participation at the highest level.

Our paper is one of the first to study the interplay between users and miners in a blockchain system and incorporate the security feature brought about by decentralization. Future research could extend our model in various way, for instance, incorporating more complicated miner behavior by considering heterogeneous mining efficiency and costs due to prices of different mining equipment and electricity consumption. It could also extend our analysis to blockchain systems use similar protocols but with functions beyond those used by cryptocurrencies as well as to systems with multiple cryptocurrencies in order to gain further understanding of the operational features underneath blockchain systems and optimal system designs.

References

Antonopoulos, A. M. (2014). *Mastering Bitcoin: unlocking digital cryptocurrencies.* " O'Reilly Media, Inc.".

Arnosti, N. and S. M. Weinberg (2018). Bitcoin: A natural oligopoly. *arXiv preprint arXiv:1811.08572*.

Babich, V. and G. Hilary (2019). Distributed ledgers and operations: What operations management researchers should know about blockchain technology. *Manufacturing & Service Operations Management*.

Bagaria, V., S. Kannan, D. Tse, G. Fanti, and P. Viswanath (2018). Deconstructing the blockchain to approach physical limits. *arXiv preprint arXiv:1810.08092*.

Basu, S., D. Easley, M. O'Hara, and E. Sirer (2019). Towards a functional fee market for cryptocurrencies. *Available at SSRN 3318327*.

Cong, L. W., Z. He, and J. Li (2019). Decentralized mining in centralized pools. Technical report, National Bureau of Economic Research.

Cui, Y., M. Hu, and J. Liu (2018). Values of traceability in supply chains. *Available at SSRN 3291661*.

Easley, D., M. O'Hara, and S. Basu (2019). From mining to markets: The evolution of bitcoin transaction fees. *Journal of Financial Economics*.

Feldman, P. and S. Micali (1988). Optimal algorithms for byzantine agreement. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pp. 148–161. ACM.

Garay, J., A. Kiayias, and N. Leonardos (2015). The bitcoin backbone protocol: Analysis and applications. In *Advances in Cryptology - EUROCRYPT*, Berlin, Heidelberg, pp. 281–310. Springer.

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

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

Hassin, R. (2016). *Rational queueing*. Chapman and Hall/CRC.

Huberman, G., J. Leshno, and C. C. Moallemi (2019). An economic analysis of the bitcoin payment system. *Columbia Business School Research Paper* (17-92).

Kleinrock, L. (1967). Optimum bribing for queue position. *Operations Research 15*(2), 304–318.

Kroll, J. A., I. C. Davey, and E. W. Felten (2013). The economics of bitcoin mining, or bitcoin in the presence of adversaries. In *Proceedings of WEIS*, Volume 2013, pp. 11.

Lavi, R., O. Sattath, and A. Zohar (2019). Redesigning bitcoin's fee market. In *The World Wide Web Conference*, pp. 2950–2956. ACM.

Li, C., P. Li, D. Zhou, W. Xu, F. Long, and A. Yao (2018). Scaling nakamoto consensus to thousands of transactions per second. *arXiv preprint arXiv:1805.03870*.

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. `https://bitcoin.org/bitcoin.pdf`.

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

Popov, S. (2016). The tangle. *cit. on*, 131.

Prat, J. and B. Walter (2018). An equilibrium model of the market for bitcoin mining.

Watson, G. N. (1929). Theorems stated by ramanujan (v): Approximations connected with ex. *Proceedings of the London Mathematical Society 2*(1), 293–308.

Wolff, R. W. (1982). Poisson arrivals see time averages. *Operations Research 30*(2), 223–231.

Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper*. `https://ethereum.github.io/yellowpaper/paper.pdf`.

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

A Appendix

A.1 Preliminary results on $M/M^K/1$ queues

Proof of Proposition 1. We first summarize some basic properties of an $M/M^K/1$ queue with arrival rate $\tilde{\lambda}$ and service rate μ. The queue length process is an ergodic continuous time Markov chain. Let Q be the random variable representing the stationary queue length. Then $\pi_n = P(Q = n) = (1-\theta)\theta^n$ and $E(Q) = \frac{\theta}{1-\theta}$ where $\theta \in (0,1)$ is the unique solution to $(\tilde{\lambda}+\mu)\theta - \tilde{\lambda} - \mu\theta^{K+1} = 0$.

Due to the ergodicity (or Poisson Arrival See Time Average, see Wolff (1982)), upon an arrival, the observed queue length has the same distribution as Q. Thus, if we let w_n be the expected hitting time of 0 for a queue of length n, then the expected time to clear a queue after an arrival is given by

$$W_q(\tilde{\lambda}) = \sum_{n\geq 0} \pi_n w_{n+1} = \frac{1-\theta}{\theta} \sum_{n=1}^{\infty} w_n \theta^n = \frac{1}{(1-\theta)[\tilde{\lambda} - \mu(K+1)\theta^K]} \tag{26}$$

where the last equality follows from the standard generating function method. The above argument serves as a proof for Proposition 1. Substituting $\tilde{\lambda} = \mu\frac{\theta-\theta^{K+1}}{1-\theta}$, we can write W_q as a function of θ

$$W_q(\theta) = \frac{1}{\mu[1-(1+K)\theta^K + K\theta^{K+1}]}., \tag{27}$$

□

Proof of Proposition 2. Since

$$\frac{dW_q(\theta)}{d\tilde{\lambda}} = W_q'(\theta)\frac{d\theta}{d\tilde{\lambda}} = \frac{K(K+1)}{\mu^2} \frac{(1-\theta)^3\theta^{K-1}}{[1-(1+K)\theta^K + K\theta^{K+1}]^3}$$

and θ is increasing in $\tilde{\lambda}$, W_q is increasing in $\tilde{\lambda}$ and it suffices to show that $\frac{(1-\theta)^3\theta^{K-1}}{\eta^3(\theta)}$, where $\eta(\theta) = 1-(1+K)\theta^K + K\theta^{K+1}$, is increasing in θ. Since

$$\frac{d}{d\theta}\left(\frac{(1-\theta)^3\theta^{K-1}}{\eta^3(\theta)}\right) = \frac{(1-\theta)^2\theta^{K-2}}{\eta^4(\theta)}[(K-1-(K+2)\theta)\eta(\theta) - 3\eta'(\theta)(1-\theta)\theta] \triangleq \frac{(1-\theta)^2\theta^{K-2}}{\eta^4(\theta)}h(\theta),$$

and $h(\theta) > 0$ when $K \leq 3$, it suffices to show that $h(\theta) > 0$ for $\theta \in (0,1)$ and hence $h'(\theta) < 0$ given that $h(1) = 0$ for $K \geq 4$. We show this by establishing that $h'(\theta)$ has a unique global maximum $\theta = 1$ in $(0,1]$ at $h'(1) = 0$. Taking the derivatives, we obtain:

$$\begin{aligned} h'(\theta) =& -(K+2) + K(K+1)(2K+1)\theta^{K-1} + (K+1)(-4K^2-4K+2)\theta^K + K(K+2)(2K+1)\theta^{K+1}, \\ h''(\theta) =& K(K+1)\theta^{K-2}[(K-1)(2K+1) + (-4K^2-4K+2)\theta + (K+2)(2K+1)\theta^2]. \end{aligned}$$

It can be easily shown that the term in "[ ]" in $h''(\theta)$ is quadratic with exactly two roots $\theta_1 < \theta_2$ in $(0,1)$. Thus, $h'(\theta)$ must achieve its global maxima at either θ_1 or 1. Since θ_1 is a root of $h''(\theta) = 0$, it satisfies $(K-1)(2K+1) + (-4K^2-4K+2)\theta_1 = -(K+2)(2K+1)\theta_1^2$ and $h'(\theta_1)$

can be reduced to

$$h'(\theta_1) = -(K+2) + 2K(2K+1)\theta_1^{K-1} - (4K^2+4K-2)\theta_1^K,$$

which is bounded from above by $\left\{ \frac{4K+2}{K+2} \left[\frac{(2K+1)(K-1)}{2K^2+2K-1} \right]^{K-1} - 1 \right\} (K+2)$. Applying $\ln(1-x) < -x - \frac{x^2}{2}$, we have:

$$\begin{aligned}
&\frac{d}{dK} \ln \left[\frac{4K+2}{K+2} \left(\frac{(2K+1)(K-1)}{2K^2+2K-1} \right)^{K-1} \right] \\
=&\frac{6K^3+18K^2+9K+3+(K+2)(2K+1)(2K^2+2K-1)\ln(1-\frac{3K}{2K^2+2K-1})}{(K+2)(2K+1)(2K^2+2K-1)} \\
<&\frac{6K^3+18K^2+9K+3+(K+2)(2K+1)(-3K-\frac{9K^2}{2K^2+2K-1})}{(K+2)(2K+1)(2K^2+2K-1)} \\
=&\frac{-3(4K^4+11K^3+3K^2-K+1)}{(K+2)(2K+1)(2K^2+2K-1)^2} < 0
\end{aligned}$$

for $K \geq 1$. Since $\frac{4K+2}{K+2}(\frac{(2K+1)(K-1)}{2K^2+2K-1})^{K-1}|_{K=4} < 1$, $h'(\theta) < 0$ $\theta \in [0,1]$ for $K \geq 4$. □

A.2 Proof of Proposition 3

Proof. We first claim that $G^*(b)$ must be continuous for $b > \underline{b}$ for a given p^*. Suppose that $G^*(b+) > G^*(b-)$ at some b. Then, for ϵ sufficiently small, the cost difference for bidding at b and $b+\epsilon$ is $b + c\left(W(b-\epsilon|(p^*\Lambda, G^*), z)\right) - (b+\epsilon) - c(W(b+\epsilon|(p^*\Lambda, G^*), z)) > 0$. Thus, bidding at $b+\epsilon$ is preferred to bidding at b, and hence $G^*(\cdot)$ must be continuous.

Second, if $\underline{b}$ is the lowest bid allowed, the lowest bid must be $\underline{b}$ as it would otherwise cost users more to bid the lowest bid without lowering the queueing latency otherwise. □

A.3 Proof for Theorem 1

Proof. By Proposition 3, the support of the equilibrium fee distribution $G^*(\cdot)$ includes $\underline{b}$ and $G^*(\cdot)$ is continuous. Thus, (6) follows as the users' cost is the same for any bid b in the support, i.e.,

$$\underline{b} + c\left(W_q(p^*\Lambda) + \frac{z}{\mu}\right) = b + c\left(W_q(p^*\Lambda(1-G^*(b))) + \frac{z}{\mu}\right).$$

Since the equilibrium joining probability $p^* \leq 1$, $p^* = 1$ if the users' utility $R - \underline{b} - c\left(W_q(\Lambda) + \frac{z}{\mu}\right) \geq 0$. Otherwise, p^* is given by $c\left(W_q(p^*\Lambda) + z/\mu\right) = R - \underline{b}$ and users' utility is 0 in equilibrium. Thus, we have (5).

Since the highest possible bid is the smallest solution to $\bar{G}^*(b) = 0$ or $\underline{b} + c\left(W_q(p^*\Lambda) + \frac{z}{\mu}\right) -$

$c\left(\frac{z+1}{\mu}\right)$, (9) holds as

$$\begin{aligned}
\Phi^*(z) &= p^*\Lambda\underline{b} + \int_{\underline{b}}^{\underline{b}+c\left(W_q(p^*\Lambda)+\frac{z}{\mu}\right)-c\left(\frac{z+1}{\mu}\right)} p^*\Lambda\bar{G}^*(b)db \\
&= p^*\Lambda\underline{b} + \int_{c\left(\frac{1+z}{\mu}\right)}^{c(W_q(p^*\Lambda)+\frac{z}{\mu})} W_q^{-1}\left(c^{-1}(s) - \frac{z}{\mu}\right) ds \\
&= p^*\Lambda\underline{b} + \int_{1/\mu}^{W_q(p^*\Lambda)} W_q^{-1}(t)dc\left(t + \frac{z}{\mu}\right) \\
&= p^*\Lambda\underline{b} + \int_0^{p^*\Lambda} \tilde{\lambda}dc\left(W_q(\tilde{\lambda}) + \frac{z}{\mu}\right) \qquad (28) \\
&= p^*\Lambda\underline{b} + p^*\Lambda c\left(W_q(p^*\Lambda) + \frac{z}{\mu}\right) - \int_0^{p^*\Lambda} c\left(W_q(\tilde{\lambda}) + \frac{z}{\mu}\right) d\tilde{\lambda} \\
&= p^*\Lambda \min\left\{R, \underline{b} + c\left(W_q(\Lambda) + \frac{z}{\mu}\right)\right\} - \int_0^{p^*\Lambda} c\left(W_q(\tilde{\lambda}) + \frac{z}{\mu}\right) d\tilde{\lambda}
\end{aligned}$$

and the expected utility of the users are given by (7). □

A.4 Proof for Proposition 4

Proof. $p^*(z)$ is decreasing in z and $G^*(b|z)$ is strictly increasing convex in b as $W_q(\cdot)$ and $c(\cdot)$ are both strictly increasing and convex, which imply that $W_q^{-1}(\cdot)$ and $c^{-1}(\cdot)$ are decreasing concave. Thus, $G^*(b|z) = 1 - \frac{W_q^{-1}\left(c^{-1}(R-b)-\frac{z}{\mu}\right)}{W_q^{-1}(c^{-1}(R-\underline{b})-\frac{z}{\mu})}$ is increasing in z for $p^* < 1$ and $G^*(b|z)$ is a constant otherwise. □

A.5 Proof for Proposition 5

Proof. We first establish the log-concavity of $\Phi^*(z)$ for piece-wise linear $c(\cdot)$ functions. That is, for $0 = s_0 < s_1 < \cdots$, $k_1 < k_2 < \cdots$ and $w \in [s_{i-1}, s_i]$,

$$c(w) = d_0 + \sum_{j=0}^{i-1} k_j(s_j - s_{j-1}) + k_i(w - s_{i-1}). \qquad (29)$$

Suppose that $W_q(0) + \frac{z}{\mu} \in [s_{m-1}, s_m)$ and $W_q(p^*(z)\Lambda) + \frac{z}{\mu} \in [s_{n-1}, s_n)$. By (28),

$$\begin{aligned}
\Phi^*(z) &= \underline{b}p^*(z)\Lambda + \int_0^{p^*(z)\Lambda} \tilde{\lambda}dc\left(W_q(\tilde{\lambda}) + \frac{z}{\mu}\right) \\
&= \underline{b}p^*(z)\Lambda + \sum_{j=m}^{n-1}(k_{j+1} - k_j)\int_{s_j-\frac{z}{\mu}}^{W_q(p^*(z)\Lambda)} W_q^{-1}(\tilde{\lambda})d\tilde{\lambda} + k_m\int_0^{W_q(p^*(z)\Lambda)} W_q^{-1}(\tilde{\lambda})d\tilde{\lambda}
\end{aligned}$$

and is differentiable even if $W_q(0) + \frac{z}{\mu} = s_{m-1}$ as

$$\lim_{\delta \downarrow 0} \frac{\Phi^*(z+\delta) - \Phi^*(z)}{\delta} - \lim_{\delta \downarrow 0} \frac{\Phi^*(z) - \Phi^*(z-\delta)}{\delta} = \frac{-(k_m - k_{m-1}) W_q^{-1}\left(s_{m-1} - \frac{z}{\mu}\right)}{\mu} = 0.$$

Since

$$\begin{aligned} \frac{d \ln \Phi^*(z)}{dz} = \frac{\Phi^{*\prime}(z)}{\Phi^*(z)} &= \frac{-\underline{b}}{\mu W_q'(p^*(z)\Lambda)\Phi^*(z)} + \frac{-k_m p^*(z)\Lambda}{\mu \Phi^*(z)} \\ &\quad + \sum_{j=m}^{n-1} (k_{j+1} - k_j) \frac{W_q^{-1}\left(s_j - \frac{z}{\mu}\right) - W_q^{-1}\left(c^{-1}(R - \underline{b}) - \frac{z}{\mu}\right)}{\mu \Phi^*(z)} \leq 0 \end{aligned}$$

by applying $p^{*\prime}(z) = -\frac{1}{\Lambda \mu W_q'(p^*(z)\Lambda)}$ from (5), $\Phi^*(z)$ decreases in z. Furthermore, both the first term and the summands in the third term are all decreasing in z, by the concavity of $W_q^{-1}(\cdot)$. Note that $\frac{\Phi^*(z)}{p^*(z)\Lambda}$ is the expected fee paid by a user who joins the system and $G^*(\cdot|z)$ is the fee distribution in equilibrium. Since $G^*(\cdot|z)$ is stochastically decreasing in z, the second term is also decreasing in z. Thus, $\ln \Phi^*(z)$ is concave and, by the Weierstrass' approximation, remains concave for general increasing convex $c(\cdot)$ functions. □

A.6 Proof for Lemma 1

Proof. By Watson (1929) and Stirling's formula,

$$\frac{1}{2} \leq \frac{1}{2} + \frac{n^n e^{-n}}{2n!} \leq \sum_{k=0}^{z} \frac{z^k e^{-z}}{k!} \leq \frac{1}{2} + \frac{2n^n e^{-n}}{3n!} \leq \frac{1}{2} + \frac{2}{3\sqrt{2\pi z}}.$$

Since

$$0 < \sum_{k=z+1}^{\infty} \frac{(z\beta)^k}{k!} \leq \sum_{k=z+1}^{\infty} \frac{(z\beta)^{z+1}}{(z+1)!} \left(\frac{z\beta}{z+2}\right)^{k-z-1} = \frac{(z\beta)^{z+1}}{(z+1)!} \frac{1}{1 - \frac{z\beta}{z+2}} \leq \frac{1}{1-\beta} \frac{1}{\sqrt{2\pi z}} \beta^z e^{z(1-2\beta)},$$

we are able to obtain our bounds. □

A.7 Proof for Lemma 2

Proof. Since $\bar{\gamma}\left(\beta, \bar{z}(\beta,\alpha)\right) = \underline{\gamma}(\beta, \underline{z}(\beta,\alpha)) = \alpha$,

$$\left[1 + \frac{\frac{4}{3} + \frac{2}{1-\beta}}{\sqrt{2\pi \bar{z}(\beta,\alpha)}}\right] \left[\beta e^{(1-\beta)}\right]^{\bar{z}(\beta,\alpha) - \underline{z}(\beta,\alpha)} = 1.$$

Since $0 < \beta e^{(1-\beta)} < 1$ and $\bar{z}(\beta,\alpha)$ is decreasing in α for $0 \leq \beta < 1$, $\bar{z}(\beta,\alpha) - \underline{z}(\beta,\alpha)$ is increasing in α. □

A.8 Proof for Proposition 6

Proof. When $p^*(z) = 1$, $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)$ is decreasing. Thus, it suffices to show that the log of the function is quasi-convex when $p^*(z) < 1$. Letting

$$\frac{d}{dz}\left[\ln\left(\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)\right)\right] = 0 \tag{30}$$

yields

$$[\ln \Phi^*(z)]' = 1 + \frac{\ln\left(\frac{A}{\Phi^*(z)}\right)}{1 - \frac{A}{\Phi^*(z)}}. \tag{31}$$

The left hand side is decreasing in z by Proposition 5 and the right hand side is increasing in z. Therefore, (30) has at most one solution and the function is quasi-convex. □

A.9 Proof for Theorem 2

Proof. The results hold since $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)$ is quasi-convex in z. □

A.10 Proof for Proposition 7

Proof. Due to the quasi-convexity of $\underline{\gamma}\left(\frac{A}{\Phi^*(z)}, z\right)$, the sequence is monotonic and hence either converges to an equilibrium or diverges to infinity. The initial conditions determine whether the sequence increases or decreases. See Figure 2 for a graphical illustration. □

A.11 Proof for Lemma 3

Proof. It suffices to show that

$$\Phi^*\left(z\middle|\frac{\eta}{K_m}, K_m, \left[R - c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m}, K_m\right) + \frac{zK_m}{\eta}\right)\right]^+\right) \geq \Phi^*(z|\mu, K, \underline{b}) \geq \Phi^*\left(z\middle|\frac{\eta}{K_m}, K_m, 0\right)$$

which implies that there is a desired $\underline{b}(z)$ such that $(z, \frac{\eta}{K_m}, K_m, \underline{b}(z))$ is feasible, and

$$\lambda^*(z|\mu, K, \underline{b}) \leq \lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, \left[R - c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m}, K_m\right) + \frac{zK_m}{\eta}\right)\right]^+\right) = \lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, 0\right).$$

By (5), λ^* is maximized when $\underline{b} = \left[R - c\left(W_q(\Lambda|\mu, K) + \frac{z}{\mu}\right)\right]^+$ for given (μ, K, z) and

$$\begin{aligned}
\lambda^*(z|\mu, K, \underline{b}) &\leq \lambda^*\left(z\middle|\mu, K, \left[R - c\left(W_q(\Lambda|\mu, K) + \frac{z}{\mu}\right)\right]^+\right) = \min\left\{W_q^{-1}\left(c^{-1}(R) - \frac{z}{\mu}\middle|\mu, K\right), \Lambda\right\} \\
&\leq \lambda^*\left(z\middle|\frac{\eta}{K}, K, \left[R - c\left(W_q\left(\Lambda\middle|\frac{\eta}{K}, K\right) + \frac{zK}{\eta}\right)\right]^+\right) \\
&\leq \lambda^*\left(z\middle|\frac{\eta}{K_m}, K_m, \left[R - c\left(W_q(\Lambda|\mu, K) + \frac{z}{\mu}\right)\right]^+\right)
\end{aligned} \tag{32}$$

where the last two inequalities follow as $W_q(\lambda|\mu,K)$ is decreasing in μ for a given (λ,K), and, by Lemma 7 below, $W_q\left(\lambda\middle|\frac{\eta}{K},K\right)$ is increasing in K for a given λ. Note that

$$\frac{\partial\Phi^*(z|\mu,K,\underline{b})}{\partial\underline{b}}=\begin{cases}\Lambda>0, & \text{if } \underline{b}<\left[R-c\left(W_q(\Lambda|\mu,K)+\frac{z}{\mu}\right)\right]^+,\\ \underline{b}\frac{\partial\lambda^*(z|\mu,K,\underline{b})}{\partial\underline{b}}<0, & \text{if } \underline{b}>\left[R-c\left(W_q(\Lambda|\mu,K)+\frac{z}{\mu}\right)\right]^+\end{cases}$$

and

$$\Phi^*\left(z\middle|\mu,K,\left[R-c\left(W_q(\Lambda|\mu,K)+\frac{z}{\mu}\right)\right]^+\right)=\int_0^{\lambda^*\left(z\middle|\mu,K,\left[R-c\left(W_q(\Lambda|\mu,K)+\frac{z}{\mu}\right)\right]^+\right)}\left[R-c\left(W_q(\tilde{\lambda}|\mu,K)\right)\right]d\tilde{\lambda}$$

increases in μ as both the upper limit and the integrand are non-negative and increasing in μ. Thus,

$$\begin{aligned}\Phi^*(z|\mu,K,\underline{b}) &\le \Phi^*\left(z\middle|\mu,K,\left[R-c\left(W_q(\Lambda|\mu,K)+\frac{z}{\mu}\right)\right]^+\right)\\ &\le \Phi^*\left(z\middle|\frac{\eta}{K},K,\left[R-c\left(W_q\left(\Lambda\middle|\frac{\eta}{K},K\right)+\frac{zK}{\eta}\right)\right]^+\right)\\ &\le \Phi^*\left(z\middle|\frac{\eta}{K_m},K_m,\left[R-c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]^+\right).\end{aligned}$$

The last inequality follows a similar argument as the previous one and Lemma 7.

It remains to show that $\Phi^*(z|\mu,K,\underline{b})\ge\Phi^*(z|\frac{\eta}{K_m},K_m,0)$. By the optimality of $(z,\mu,K,\underline{b})$, equality holds for (32), implying that either $(\mu,K,\underline{b})=\left(\frac{\eta}{K_m},K_m,\left[R-c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]^+\right)$ or $\lambda^*(z|\mu,K,\underline{b})=\Lambda$. For the former, the result holds trivially. Otherwise, it follows from (9) that $(\tilde{\mu},\tilde{K},\tilde{\underline{b}})=(\frac{\eta}{K_m},K_m,0)$ minimizes $\Phi^*(z|\tilde{\mu},\tilde{K},\tilde{\underline{b}})$ when $\lambda^*(z|\tilde{\mu},\tilde{K},\tilde{\underline{b}})=\Lambda$. Hence $\Phi^*(z|\frac{\eta}{K_m},K_m,0)\le\Phi^*(z|\mu,K,\underline{b})$. □

Lemma 7. *For a given λ, $W_q\left(\lambda\middle|\frac{\eta}{K},K\right)$ is increasing in K.*

Proof. By (27) and letting $\theta\in(0,1)$ be the unique solution to $\frac{\theta-\theta^{K+1}}{1-\theta}=K\frac{\lambda}{\eta}$, we have

$$W_q\left(\lambda\middle|\frac{\eta}{K},K\right)=\frac{K}{\eta[1-(1+K)\theta^K+K\theta^{K+1}]}=\frac{\theta}{(1-\theta)\left[\lambda(1+K)-(\eta+\lambda K)\,\theta\right]}.\tag{33}$$

It is obvious that $W_q(\lambda|\eta,1)<W_q\left(\lambda\middle|\frac{\eta}{2},2\right)$. Thus, it suffices to show that $\frac{d}{dK}W_q\left(\lambda\middle|\frac{\eta}{K},K\right)>0$ for $K\in[2,\infty)$, which is equivalent to

$$\frac{(\eta+2\lambda K+\lambda)\,\theta}{\lambda(1+K)-(\eta+\lambda K)\,\theta^2}>\frac{-\eta\theta^{K+1}\ln\theta}{\lambda K(1-\theta)}+1-\frac{1}{K}\tag{34}$$

by applying

$$\frac{d\theta}{dK}=\frac{\theta\left[\lambda(1-\theta)+\eta\theta^{K+1}\ln\theta\right]}{K\left[\lambda(1+K)-(\eta+\lambda K)\,\theta\right]}.\tag{35}$$

Since $1+\frac{\lambda}{\eta}K-\frac{\lambda K}{\theta\eta}=\theta^K\in\left(0,\frac{\lambda}{\eta}\right)$, $\theta\in\left(\frac{\lambda K}{\eta+\lambda K},\frac{\lambda K}{\eta-\lambda+\lambda K}\right)$. Therefore,

$$\frac{(\eta+2\lambda K+\lambda)\,\theta}{\lambda(1+K)-(\eta+\lambda K)\,\theta^2}>\frac{K\,(\eta+2\lambda K+\lambda)}{(1+K)\eta+\lambda K}. \tag{36}$$

Furthermore,

$$\theta^K<\left(\frac{\lambda K}{\eta-\lambda+\lambda K}\right)^K\leq\frac{2\lambda^2 K}{\left[\lambda^2+\eta^2\right]K-(\eta-\lambda)^2}$$

and $-\frac{\ln\theta}{1-\theta}\leq\frac{1}{\theta}$ imply that

$$\frac{-\eta\theta^{K+1}\ln\theta}{\lambda K(1-\theta)}+1-\frac{1}{K}<\frac{2\lambda}{\left[\lambda^2+\eta^2\right]K-(\eta-\lambda)^2}-\frac{1}{K}+1\leq\frac{K\,(\eta+2\lambda K+\lambda)}{(1+K)\eta+\lambda K}$$

for $K\geq 2$. Thus (34) and hence, the lemma hold. □

A.12 Proof for Lemma 4

Proof. The second statement follows by Proposition 6 with $\underline{b}=0$. For ease of notation, we denote $b_1(z)=\left[R-c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]^+$.

If $R-c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{K_m}{\eta}\right)<0$, then $b_1(z)=0$ for all $z\geq 1$, and $\underline{\gamma}\left(\frac{A}{\Phi^*\left(z\middle|\frac{\eta}{K_m},K_m,b_1(z)\right)},z\right)$ is quasi-convex.

If $R-c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{K_m}{\eta}\right)\geq 0$, let z_0 be the smallest such that $R-c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{z_0K_m}{\eta}\right)\leq 0$,

$$\Phi^*\left(z\middle|\frac{\eta}{K_m},K_m,b_1(z)\right)=\begin{cases}\int_0^{\Lambda}\left[R-c\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]d\lambda, & \text{if } z\leq z_0,\\ \int_0^{W_q^{-1}\left(c^{-1}(R)-\frac{zK_m}{\eta}\middle|\frac{\eta}{K_m},K_m\right)}\left[R-c\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]d\lambda, & \text{if } z>z_0.\end{cases}$$

Since $c\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)$ is increasing convex in z, $\ln\left(\int_0^{\Lambda}\left[R-c\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]d\lambda\right)$ is decreasing concave in z. Following a similar argument as in the proof of Proposition 6, we have that $\underline{\gamma}\left(\frac{A}{\int_0^{\Lambda}\left[R-c\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]d\lambda},z\right)$ and $\underline{\gamma}\left(\frac{A}{\int_0^{W_q^{-1}\left(c^{-1}(R)-\frac{zK_m}{\eta}\middle|\frac{\eta}{K_m},K_m\right)}\left[R-c\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]d\lambda},z\right)$ are quasi-convex for $z\geq 0$.

When $z\leq z_0$, $W_q^{-1}\left(c^{-1}(R)-\frac{zK_m}{\eta}\middle|\frac{\eta}{K_m},K_m\right)\geq\Lambda$ and hence,

$$\begin{aligned}&\underline{\gamma}\left(\frac{A}{\int_0^{\Lambda}\left[R-c\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]d\lambda},z\right)\\ &\geq\underline{\gamma}\left(\frac{A}{\int_0^{W_q^{-1}\left(c^{-1}(R)-\frac{zK_m}{\eta}\middle|\frac{\eta}{K_m},K_m\right)}\left[R-c\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]d\lambda},z\right).\end{aligned}$$

Therefore, $\underline{\gamma}\left(\frac{A}{\Phi^*\left(z\middle|\frac{\eta}{K_m},K_m,b_1(z)\right)},z\right)$ is quasi-convex. □

A.13 Proof of Lemma 5

Proof. The feasibility of $(z,\frac{\eta}{K_m},K_m,\underline{b}(z))$ follows from Lemma 3 if $(z,\mu,K,\underline{b})$ is feasible. Here, the users' utility can be reformulated as

$$U^*(z|\mu,K,\underline{b})=\Lambda R-\Phi^*(z|\mu,K,\underline{b})-\int_0^\Lambda c\left(W_q(\lambda|\mu,K)+\frac{z}{\mu}\right)d\lambda.$$

Since $\Phi^*(z|\mu,K,\underline{b})=\Phi^*\left(z\middle|\frac{\eta}{K_m},K_m,\underline{b}(z)\right)$ and $W_q(\lambda|\mu,K)+\frac{z}{\mu}\geq W_q\left(\lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}$ with equality holds only when $(\mu,K)=\left(\frac{\eta}{K_m},K_m\right)$, $U^*(z|\mu,K,\underline{b})\leq U^*\left(z\middle|\frac{\eta}{K_m},K_m,\underline{b}(z)\right)$. □

A.14 Proof of Lemma 6

Proof. Here, we prove the lemma for differentiable $c(\cdot)$. Since any convex function can be approximated by a sequence of differentiable convex functions, the result follows for general $c(\cdot)$. It suffices to show that

$$\frac{dU^*(\Phi^*)}{d\Phi^*}=-1+\frac{\ln(2\alpha)}{A\frac{\eta}{K_m}}\frac{\left(\frac{A}{\Phi^*}\right)^2-\frac{A}{\Phi^*}}{\left[1-\frac{A}{\Phi^*}+\ln\left(\frac{A}{\Phi^*}\right)\right]^2}\cdot\int_0^\Lambda c'\left(W_q\left(\lambda\middle|\frac{\eta}{K_m},1\right)+\frac{\ln(2\alpha)}{\frac{\eta}{K_m}}\frac{1}{1-\frac{A}{\Phi^*}+\ln(\frac{A}{\Phi^*})}\right)d\lambda=0$$

has a unique solution. Since $\frac{\left(\frac{A}{\Phi^*}\right)^2-\frac{A}{\Phi^*}}{\left[1-\frac{A}{\Phi^*}+\ln\left(\frac{A}{\Phi^*}\right)\right]^2}$ is increasing in $\Phi^*>A$ with $\lim_{\frac{A}{\Phi^*}\downarrow 0}\frac{\left(\frac{A}{\Phi^*}\right)^2-\frac{A}{\Phi^*}}{\left[1-\frac{A}{\Phi^*}+\ln\left(\frac{A}{\Phi^*}\right)\right]^2}=0$ and $\lim_{\frac{A}{\Phi^*}\uparrow 1}\frac{\left(\frac{A}{\Phi^*}\right)^2-\frac{A}{\Phi^*}}{\left[1-\frac{A}{\Phi^*}+\ln\left(\frac{A}{\Phi^*}\right)\right]^2}=-\infty$, $\frac{\ln(2\alpha)}{A\frac{\eta}{K_m}}\frac{\left(\frac{A}{\Phi^*}\right)^2-\frac{A}{\Phi^*}}{\left[1-\frac{A}{\Phi^*}+\ln\left(\frac{A}{\Phi^*}\right)\right]^2}\geq 0$ and decreases in Φ^*. Noting that the integration is also non-negative and decreasing in Φ^* by the convexity of $c(\cdot)$, we have $\frac{dU^*(\Phi^*)}{d\Phi^*}$ is decreasing in Φ^*. Thus, $\lim_{\Phi^*\downarrow A}\frac{dU^*(\Phi^*)}{d\Phi^*}=+\infty$ and $\lim_{\Phi^*\uparrow+\infty}\frac{dU^*(\Phi^*)}{d\Phi^*}=-1$ guarantee a unique solution $\hat{\Phi}^*$. Since $U^*\left(\Phi^*\left(z_3\middle|\frac{\eta}{K_m},1,0\right)\right)=0$, $\frac{dU^*(\Phi^*)}{d\Phi^*}|_{\Phi^*=\Phi^*\left(z_3\middle|\frac{\eta}{K_m},1,0\right)}\leq 0$ and $\hat{\Phi}^*=\Phi^*\left(z^*(\hat{\Phi}^*)\middle|\frac{\eta}{K_m},1,0\right)\leq\Phi^*\left(z_3\middle|\frac{\eta}{K_m},1,0\right)$, which implies $z^*(\hat{\Phi}^*)\geq z_3$. □

A.15 Proof of Theorem 3

Proof. Since $\underline{\gamma}\left(\frac{A}{\Phi^*(z^*)},z^*\right)<\underline{\gamma}\left(\frac{A}{\Phi^*(z^*)},z^*-1\right)$ and both functions are quasi-convex, $\underline{\gamma}\left(\frac{A}{\Phi^*(z^*)},z^*-1\right)=\alpha$ has at most two roots $z_1'\leq z_2'$ such that $z_1^*<z_1'\leq z_2'\leq z_2^*$ and all the integers in $[z_1^*,z_1')$ and $(z_2',z_2^*]$ are equilibria. Furthermore, $\underline{\gamma}\left(\frac{A}{\Phi^*(\lceil z_1^*\rceil)},\lceil z_1^*\rceil-1\right)\geq\underline{\gamma}\left(\frac{A}{\Phi^*(\lceil z_1^*\rceil-1)},\lceil z_1^*\rceil-1\right)>\alpha$ implies $z_1'>\lceil z_1^*\rceil$. □

A.16 Proof of Proposition 11

Proof. The constraints of the reduced optimization problem analogous to (18)-(20) can be written as:

$$\underline{\gamma}\left(\frac{A}{B_0+\Phi^*\left(z\middle|\frac{\eta}{K_m},K_m,\left[R-c\left(W_q\left(\Lambda\middle|\frac{\eta}{K_m},K_m\right)+\frac{zK_m}{\eta}\right)\right]^+\right)}\right) \leq \alpha \tag{37}$$

$$\underline{\gamma}\left(B_0+\frac{A}{\Phi^*\left(z\middle|\frac{\eta}{K_m},K_m,0\right)},z-1\right) > \alpha \tag{38}$$

Due to the monotonicity of the objective function, the first statement follows from the same reasoning as that of Proposition 8. Note that when $\lambda^*\left(z\middle|\frac{\eta}{K_m},K_m,0\right)=\Lambda$, the left side of (37) is quasi-convex while the left side of (38) is decreasing in z. Thus, the second statement follows. □

A.17 Proof of Proposition 13

Proof. Following a similar argument as that provided in the proof of Theorem 1, we can show that G^* is continuous for any given p^*. Next, we show that b increases in C. Suppose that $c_1 < c_2$ but $b_1 > b_2$. As a result, it is more costly for users with c_1 to bid at b_2 than at b_1, i.e.,

$$b_1 + c_1 W_q\left((1-G^*(b_1)p^*\Lambda)\right) \leq b_2 + c_1 W_q((1-G^*(b_2)p^*\Lambda))$$

or

$$W_q\left((1-G^*(b_2))p^*\Lambda\right) - W_q((1-G^*(b_2))p^*\Lambda) \geq \frac{b_1-b_2}{c_1}.$$

Hence, it is more costly for users with c_2 to bid at b_2 than at b_1 as

$$[b_2 + c_2 W_q((1-G^*(b_2))p^*\Lambda)] - [b_1 + c_2 W_q((1-G^*(b_1))p^*\Lambda)] \geq b_2 - b_1 + \frac{c_2}{c_1}(b_1-b_2) > 0,$$

which contradicts with the definition of an equilibrium.

The monotonicity of b depending on C implies that users with a waiting cost $C(qp^*)$ bid at $b(q)$ in equilibrium, i.e., $b(q)$ is a minimizer of the total cost $b + C\left(qp^*\right) W_q((1-G^*(b))p^*\Lambda)$. By the first-order optimality condition, we have:

$$\frac{db(q)}{dq} = C\left(qp^*\right) W_q'((1-q)p^*\Lambda).$$

Solving the above differential equation with the boundary condition $b(0)=\underline{b}$, we obtain the

desired result for $b(q)$. The total cost of bidding at $b(q)$ is given by

$$\begin{aligned}
&b(q) + C\left(qp^*\right)\left(W_q((1-q)p^*\Lambda) + \frac{z}{\mu}\right) \\
&= \underline{b} + \int_{(1-q)p^*}^{p^*} C\left(p^* - p\right) dW_q(p\Lambda) + C\left(qp^*\right)\left(W_q((1-q)p^*\Lambda) + \frac{z}{\mu}\right) \\
&= b(0) + C(0)W_q(p^*\Lambda) + \int_{(1-q)p^*}^{p^*} W_q(p\Lambda) dC\left(p^* - p\right) + C\left(qp^*\right)\frac{z}{\mu},
\end{aligned}$$

and is increasing in q. Thus, p^* is the largest such that $b(1) + C\left(p^*\right)\left(\frac{1+z}{\mu}\right) \leq R$. The expression of Φ^* follows immediately. □