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

B Endogenous Entry

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

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

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

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

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

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

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

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

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

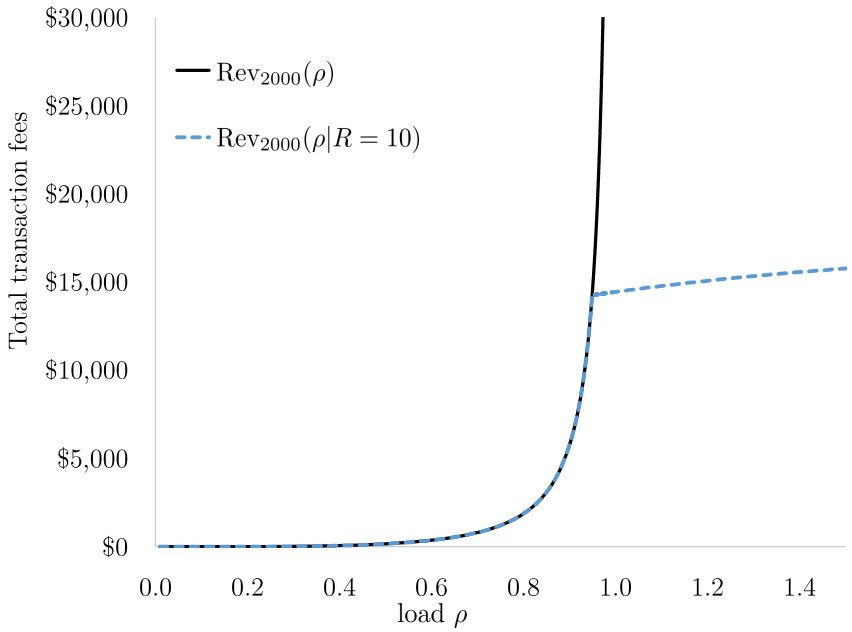

Figure 5: Total revenue per block as a function of ρ when $c \sim U[0,1]$. The curve $\text{Rev}_{2000}(\rho)$ shows total revenue from transaction fees when WTP is sufficiently high so that the participation constraint does not bind for any user, and it is only defined for $0 \leq \rho < 1$. The curve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

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

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

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

C Endogenous Willingness To Pay

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

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

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

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

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

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

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

where c^* is the unique solution to

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

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

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

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

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

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

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

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

The following example provides simplified expressions under additional assumptions.

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

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

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

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

and the implied revenue from transaction fees is

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

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

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

D Attributes of Transaction Fees

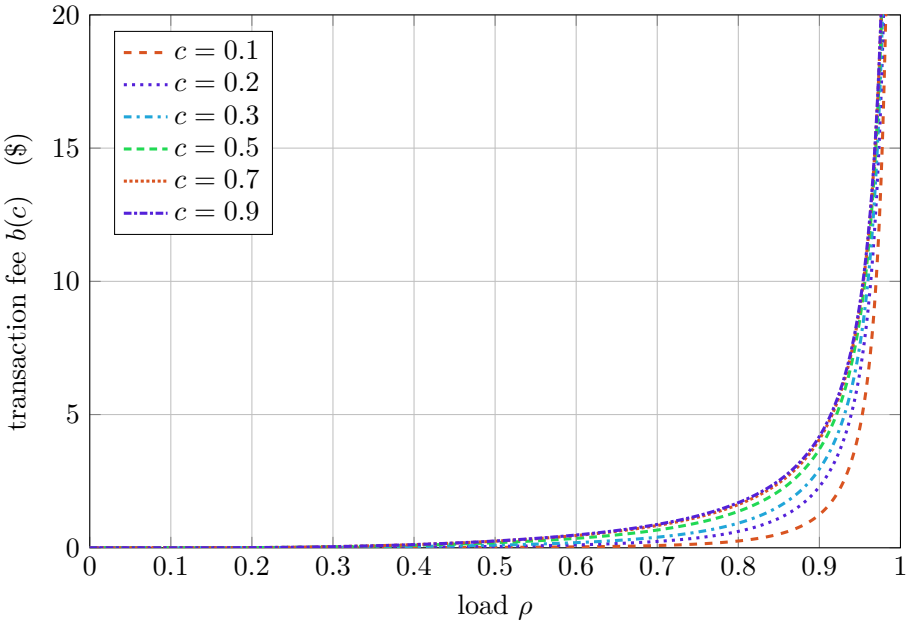

Figure 6: The dependence of equilibrium transaction fees on congestion ρ for fixed user's delay cost c. Block size is taken to be $K = 2,000$, block arrival rate $\mu = 1$, and delay costs are distributed according to $c \sim U[0,1]$.

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

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

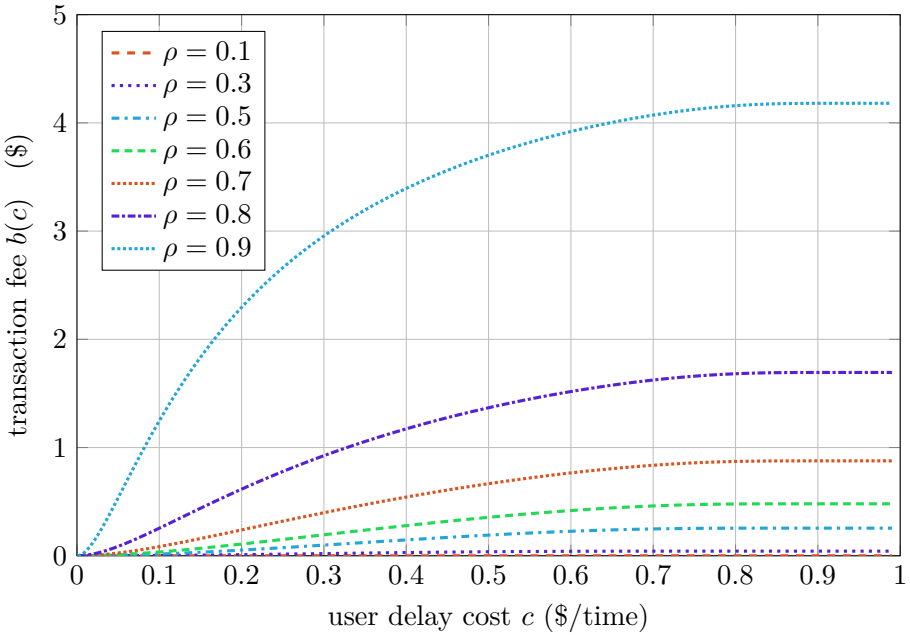

Figure 7: The dependence of equilibrium transaction fees on the user's delay cost c for fixed congestion ρ. Block size is taken to be $K = 2,000$, block arrival rate $\mu = 1$, and delay costs are distributed according to $c \sim U[0,1]$.

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

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

E Additional Figures

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

complement Figure 1.

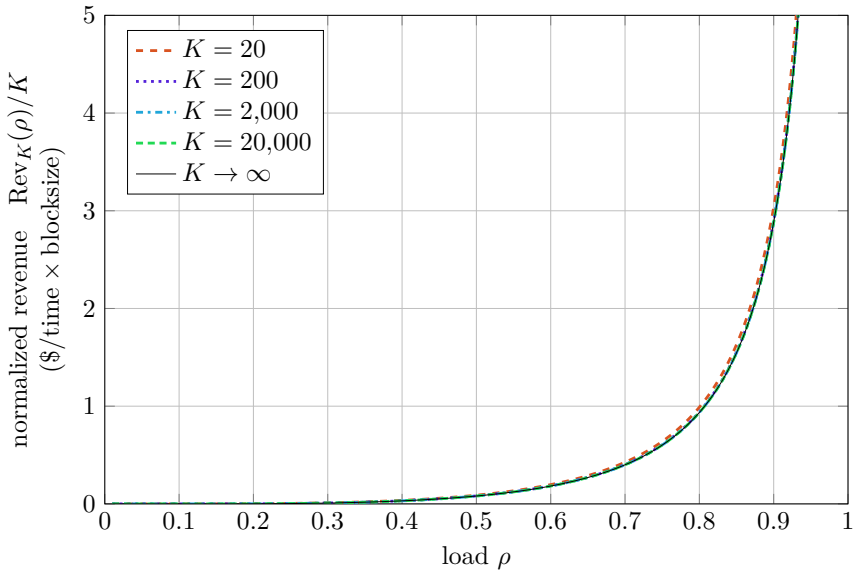

Figure 8: Normalized revenue $\mathrm{Rev}_K(\rho)/K$ *when* $c \sim U[0,1]$ *and* $K \in \{20, 200, 2000, 20000\}$*, compared to the limiting values obtained from the approximation using* $W_\infty(\cdot)$*. The plot may appear to have only one line because all lines overlap.*

OLS Regression Results

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

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

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

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

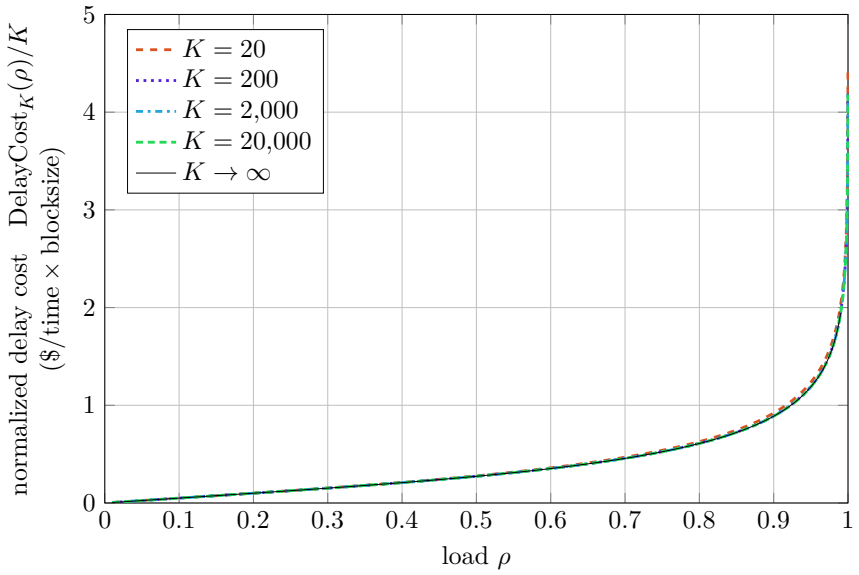

Figure 9: Normalized revenue $\text{Rev}_K(\rho)/K$ when $c \sim U[0,1]$ and $K \in \{20, 200, 2000, 20000\}$, compared to the limiting values obtained from the approximation using $W_\infty(\cdot)$. The plot may appear to have only one line because all lines overlap.

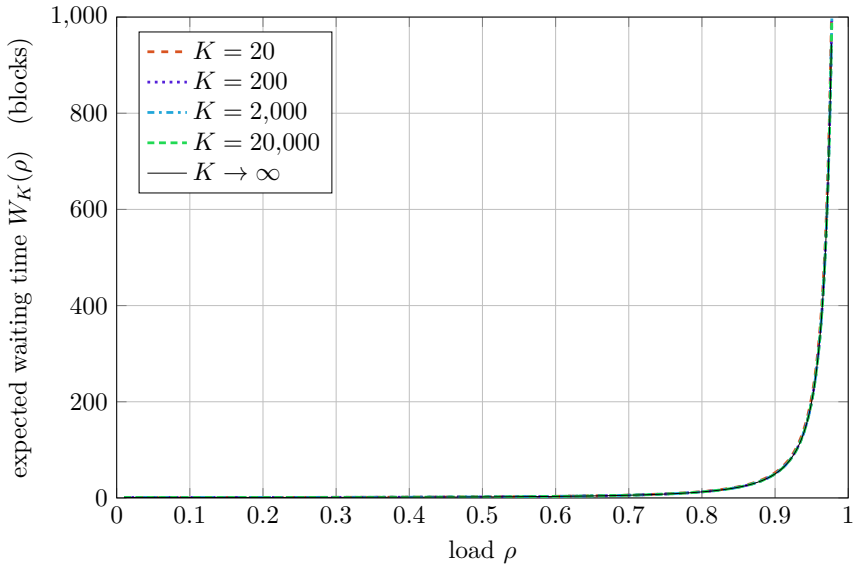

Figure 10: The expected delay in blocks $W_K(\rho)$ of the lowest priority transaction given $\rho = \lambda/\mu K$ and $K \in \{20, 200, 2000, 20000\}$.

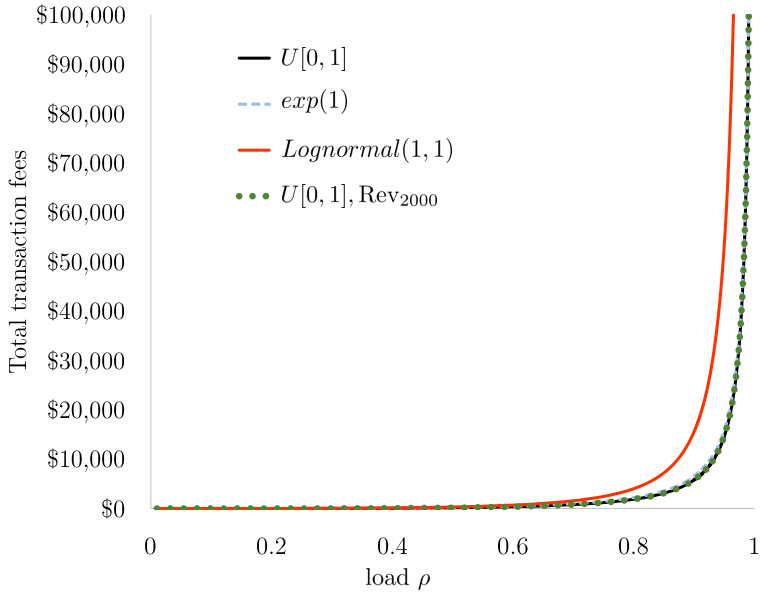

Figure 11: Revenue for $K = 2000$ and waiting costs c distributed (i) uniformly on $[0, 1]$, (ii) as an exponential with mean 1, (iii) as a Log-normal with mean and variance equal to 1. All were calculated using the asymptotic approximation. The plot also shows $\text{Rev}_{2000}(\rho)$ for the uniform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

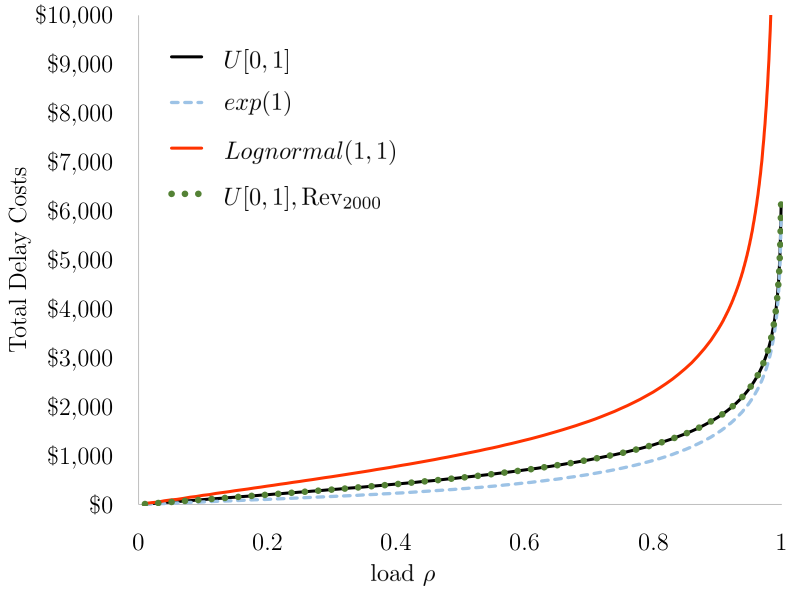

Figure 12: Delay costs for $K = 2000$ and waiting costs c distributed (i) uniformly on $[0, 1]$, (ii) as an exponential with mean 1 (iii) as a Log-normal with mean and variance equal to 1. All were calculated using the asymptotic approximation. The plot also shows $\mathrm{Rev}_{2000}(\rho)$ for the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

F Proofs

F.1 Queueing Analysis

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

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

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

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

solution of the polynomial equation

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

in the interval $[0, 1)$.

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

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

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

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

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

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

Finally, we have that

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

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

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

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

Applying Little's Law,

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

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

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

Substituting into (14), we have that

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

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

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

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

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

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

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

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

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

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

The same substitution into (18) yields that

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

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

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

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

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

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

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

Then,

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

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

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

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

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

Therefore, z_0 satisfies

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

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

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

□

F.2 Equilibrium

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

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

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

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

and

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

Each agent is bidding optimally if and only if

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

The first order condition implies

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

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

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

or

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

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

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

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

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

Finally,

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

Therefore,

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

□

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

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

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

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

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

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

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

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

F.3 Delay and Revenue

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

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

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

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

The total delay cost per unit time incurred by users is

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

The total overall cost per unit time borne by users is

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

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

transaction is

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

Therefore, the total expected revenue per unit time is

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

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

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

to obtain

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

as desired.

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

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

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

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

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

and (25). □

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

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

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

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

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

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

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

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

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

□

F.4 Large Block Asymptotics

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

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

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

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

The following lemma makes this intuition precise:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that

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

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

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

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

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

Then,

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

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

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

or equivalently,

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

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

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

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

Then,

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

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

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

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

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

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

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

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

Moreover, the asymptotic expected delay satisfies

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

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

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

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

Then,

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

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

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

Therefore,

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

as desired.

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

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

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

Simplifying, we obtain that

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

Then, differentiating (32), we have that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

The following Theorem implies Theorems 5–6:

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

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

where

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

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

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

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

Finally, as $\rho \to 0$,

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

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

Proof. Note that, from (24),

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

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

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

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

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

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

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

as desired.

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

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

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

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

where we have used the fact that

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

Then, applying Lemma A7

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

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

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

□

The following theorem implies Theorem 7:

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

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

where

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

for $r > 0$.

Then, as $K \to \infty$,

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

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

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

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

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

as $K \to \infty$.

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

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

Equivalently,

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

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

F.5 Profit-Maximizing Firm

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This allows us to replace IC-R_H with

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

and rewrite problem (38) as

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

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

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

Because the IC-R_H must bind, we have

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

Plugging this into the objective and rearranging we obtain

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

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

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

and

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

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