Online Appendix to

"Efficient Inaccuracy: User-Generated Information Sharing in a Queue"

OA1. Throughput Comparison

In this section, we compare the throughput under the shared, no, and full QL information (QLI) structures. Similar to the comparison of shared- and full-QLI structures in social welfare in Section 3.4 , we first generate insights by making the comparison in an asymptotic case $\Lambda = \infty$. Then, building on the intuition obtained from the asymptotic case, we verify the comparative statics of the throughput under shared-, no-, and full-QLI structures in the general case $\Lambda < \infty$.

Asymptotic case. Chen and Frank (2004) compare the throughput in observable and unobservable queues. They show that under a high arrival rate, providing the real-time QLI to customers improves the throughput. This is because, under no real-time QLI and a high arrival rate, customers will join with a relatively low probability. The service provider then prefers to reveal the real-time QLI to customers so that they will always join a short queue which they would not join under no QLI. The following proposition confirms that similar intuition holds in our model with two decision epochs.

Proposition 5. *In the asymptotic case $\Lambda = \infty$, the throughput under full QLI Λ^F is greater than that under no QLI Λ^N, i.e., $\Lambda^F > \Lambda^N$.*

The same intuition holds for the throughput under shared- and no-QLI structures. The somewhat inaccurate shared snapshot information encourages customers to enter the facility when the queue is more likely to be short, while the no-QLI structure does not provide any information regarding the real-time QL, so customers enter the facility with a low probability. The analytical comparison of throughput under shared- and no-QLI structures is difficult for two reasons. First, the derivation of ρ_N under no QLI involves solving a polynomial equation of degree $n+1$, which by Abel Ruffini Theorem has no algebraic solutions for $n \geq 4$; see, e.g., Corollary 2. Second, the derivation of the steady state probability distribution of online QL updates under shared QLI is cumbersome. We next investigate the difference between the throughput under shared- and no-QLI structures, i.e., $\Lambda^S - \Lambda^N$, numerically. Without loss of generality, we normalize the service rate and waiting cost to one; i.e., $\mu = 1$ and $c = 1$, and plot $\Lambda^S - \Lambda^N$ as functions of the hassle cost h under different service reward $R \in \{4, 5, \ldots, 10\}$ in Figure 6. We observe the throughput under shared QLI is higher than that under no QLI, i.e., $\Lambda^S > \Lambda^N$, for all values of R and h tested. (More numerical results are available upon request.)

We next compare the throughput under shared and full QLI in the asymptotic case. Recall from Corollary 3 that in the asymptotic case $\Lambda = \infty$ under full QLI, once a customer finishes service, another customer joins the queue immediately and brings the QL to m, so the server will be constantly busy, and the resulting throughput is identical to the service rate μ.

Under the shared-QLI structure, the online QLI ϕ is lagged and not as accurate as under full QLI. This leads to two effects on the throughput. On the one hand, even if the real-time QL drops

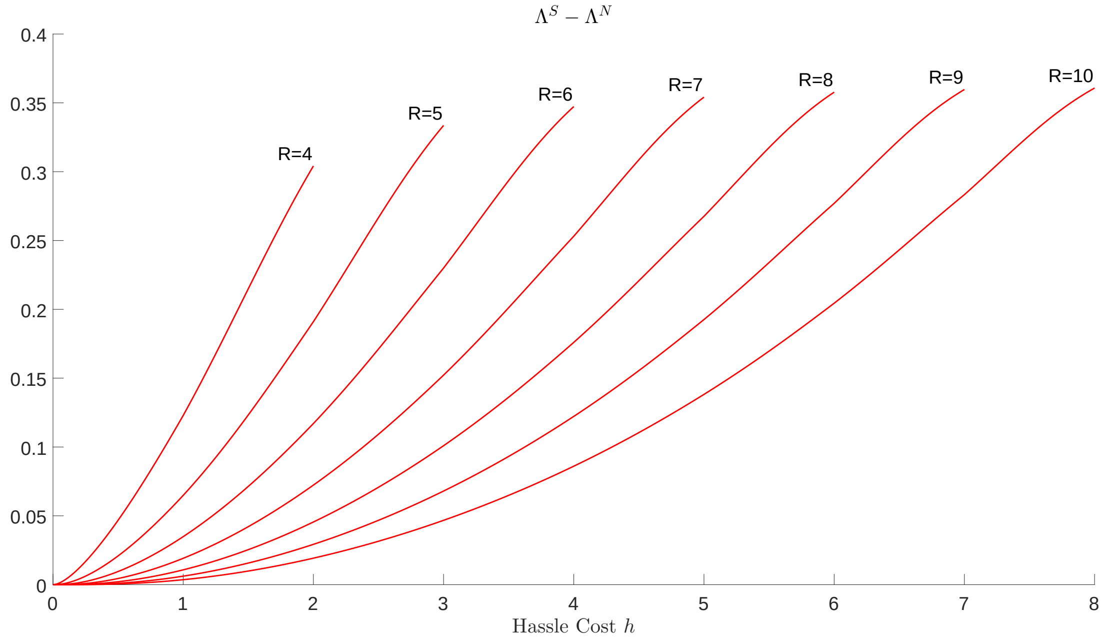

Figure 6 **Difference between the throughput under shared- and no-QLI structures, i.e., $\Lambda^S - \Lambda^N$ as a function of the hassle cost h under different service reward $R \in \{4, 5, \ldots, 10\}$, service rate $\mu = 1$, marginal waiting cost $c = 1$, and arrival rate $\Lambda = 10^8$.**

to $n-1$ during the arrival shutdown period, no customers will know about it, and neither do they enter the facility. Thus, under the shared-QLI structure, not all desirably short queues are filled by customers as soon as those under the full-QLI structure. This *input reduction* effect creates additional possibility for the server to become idle, hence resulting in less throughput. On the other hand, if the real-time QL stays long (i.e., greater than $m-1$) after the arrival shutdown period, no customers are aware of that, and they will enter the facility. Then, because of the sunk of the hassle cost, customers may join some queues that they would have not joined under full information. This *input boost* effect generates higher probability for the server to stay busy, so leads to higher throughput. When the arrival rate is high, the server's utilization is already close to one and the input boost effect is limited, so the throughput under full QLI dominates that under shared information.

When the service becomes more valuable, i.e., R increases, even though the inaccuracy in the shared QLI discourages some customers from entering the facility when the most recently shared QL ϕ reaches at least m, the probability all $\phi \geq m$ customers finish service during the arrival shutdown period and the server becomes idle decreases. Therefore, the throughput gap between the shared- and full-QLI structure diminishes. In the limiting case when the service becomes extremely valuable, i.e., $R \to \infty$, the throughput under shared QLI is almost identical to that under full QLI. The next theorem summarizes these results.

Theorem 4. *In the asymptotic case $\Lambda = \infty$, compare the throughput under the shared- and full-QLI structures:*
(i) The throughput under shared QLI is less than that under full QLI for any finite service reward;

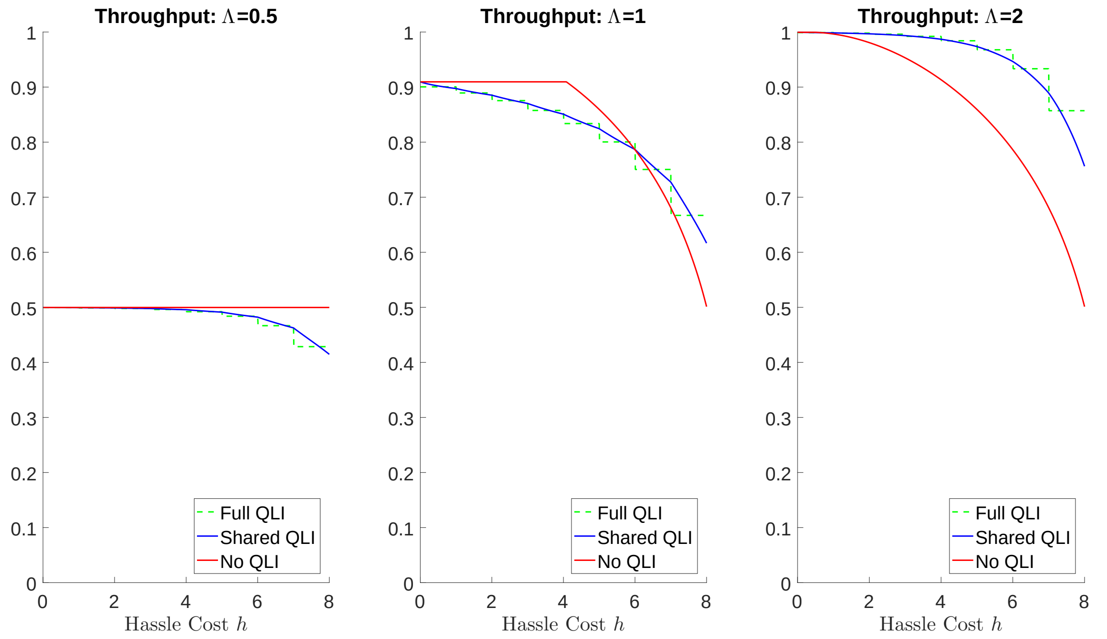

Figure 7 **Throughput under the shared, full, and no information structures as a function of hassle cost h, for service reward $R = 10$, service rate $\mu = 1$, marginal waiting cost $c = 1$, and arrival rate $\Lambda \in \{0.5, 1, 2\}$.**

i.e., $\Lambda^S < \Lambda^F$ for $R < \infty$.
(ii) When $R \to \infty$, the throughput under shared and full QLI is asymptotically identical; i.e., $\lim_{R\to\infty} \Lambda^S = \Lambda^F$.

Theorem 4 shows that a system with shared QLI has a lower throughput than one with full QLI. However, the service reward R mitigates the throughput gap between these two information structures. Moreover, when the service reward grows, systems are almost indifferent between the shared- and full-QLI structures with regard to the throughput measure. Thus, systems with high service reward can be confident in employing the shared-QLI structure by encouraging information sharing. Shared QLI generates almost identical throughput as the full QLI.

General case. Next, we adopt a numerical approach to investigate the comparative statics of throughput under three information structures in the general case. In Figure 7, we plot throughput under the three information structures, Λ^S, Λ^F, and Λ^N, as functions of hassle cost h, for $R = 10$ and $\mu = c = 1$ in the general cases $\Lambda \in \{0.5, 1, 2\}$.

First, we observe from Figure 7 that throughput under shared QLI remains at similar level as under full QLI for different arrival rates Λ. The throughput under full QLI is a decreasing staircase constant function of h, while that under shared QLI is a continuous decreasing function of h. When the hassle cost h approaches $R - \frac{c}{\mu} i$, for integer i, from below we have $\Lambda^S < \Lambda^F$; and when h approaches $R - \frac{c}{\mu} i$ from above we have $\Lambda^S > \Lambda^F$. There is not an interval longer than $\frac{c}{\mu}$ such that either shared- or full-QLI structure dominates the other. This observation improves our understanding from the asymptotic case. When the arrival rate declines, the input reduction and input boost effects have similar strength for different hassle costs, so the throughput under shared QLI is at a similar level to that under full QLI.

Under no QLI, recall from the expressions of Λ^N (Proposition 2) that throughput is a piecewise function. When the offered load is low, i.e., $\rho < \rho_N$, all customers enter the facility; and when the offered load is high, i.e., $\rho \geq \rho_N$, a fraction of customers enter. The value of the critical cutoff ρ_N decreases with the hassle cost h.

Furthermore, we observe from Figure 7 that when the arrival rate Λ is high, e.g., $\Lambda = 2$, or the hassle cost h is large, e.g., $\Lambda = 1$ and $h \geq 6$, the shared- and full-QLI structures continuously dominate the no-QLI structure on throughput for different hassle costs h. This observation is consistent with intuitions obtained from the asymptotic case. When Λ is small, e.g., $\Lambda = 0.5$, or the hassle cost h is small, e.g., $\Lambda = 1$ and $h < 4$, customers under no QLI all enter, but customers under shared or full QLI do not always join blindly and those who arrive when the queue is expected to be long will balk. Hence, in this case, the throughput under shared and full QLI is lower than that under no QLI.

OA2. Expected Queue Length Comparison

In this section, we compare the expected QL under shared-, no-, and full-QLI structures, L^S, L^N, and L^F, numerically.

We first compare the expected QL under shared- and full-QLI structures. Recall from Propositions 1 and 3 that we can apply a renewal theory based approach to derive various service level measures for the system under the shared- and full-QLI structures. The transition cycles under the shared and full QLI have a similar structure – in each transition cycle, there is an arrival shutdown period and an arrival open period in sequence. The expected QL at the end of the arrival shutdown period under shared QLI is $\omega - 1$, while that under full QLI is $m - 1$. Note that $m - 1$ and $\omega - 1$ are close to each other, and their difference $\omega - m$ is smaller than one. Then, after the exponentially distributed arrival open period with parameter Λ, the QL seen by the new arrival under shared and full information structures will be close on average. Based on the fact that the expected QLs at the beginning of the arrival shutdown and arrival open periods under both shared and full QLI are expected to be close, we anticipate that the expected QL under shared QLI is at a similar level to that under full QLI. Figure 8 in which we plot L^S, L^N, and L^F, as functions of hassle cost h, for $R = 10$ and $\mu = c = 1$ in the general cases $\Lambda \in \{0.5, 1, 2, 10^8\}$, confirms our intuition.

Next, we compare the expected QL under shared- and full-QLI structures with that under no-QLI structure. Comparing Figures 7 and 8, we see that higher throughput is associated with a greater expected QL when the offered load is sufficiently small or large. For example, when $\Lambda = 0.5$, the throughput and expected QL under no QLI are greater than those under shared or full QLI; and when $\Lambda = 10^8$, the throughput and expected QL under no QLI are less than those under shared or full QLI. However, when the offered load is in an intermediate range, higher throughput may not come at a cost of a higher waiting cost. For example, when $\Lambda = 2$, throughput under shared and full QLI is higher than that under no QLI in Figure 7, while the expected QL under shared and full QLI may not be higher than that under no QLI in Figure 8. This is because the QLI, even a little shared by previous customers, helps to better match the supply with demand, so customers enter the facility when the queue is expected to be relatively short and there is a relatively high probability for the server to become idle.

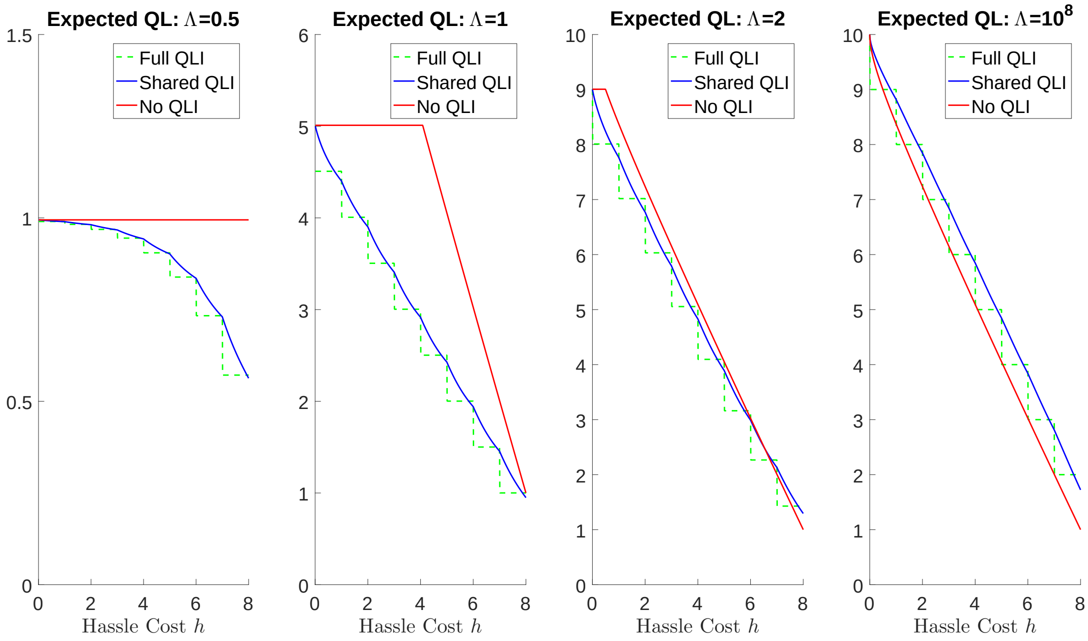

Figure 8 **Expected queue length under the shared, full, and no information structures as a function of hassle cost** h**, for service reward** $R = 10$**, service rate** $\mu = 1$**, marginal waiting cost** $c = 1$**, and arrival rate** $\Lambda \in \{0.5, 1, 2, 10^8\}$.

OA3. Individual Utility Comparison

We next look into the strategic behavior of connected and unconnected customers. We start with unconnected customers only. This is equivalent to our model under no QLI. Recall from the discussion of Proposition 2 that there are two critical cutoffs in the unconnected customers' offered load: ρ_L and ρ_N for any hassle cost h. If the unconnected customers' offered load is high (i.e., $\rho_U \geq \rho_L$), unconnected customers relentlessly enter the facility expecting non-negative utility, which causes excessive congestion and drives away connected customers. In this case, the expected QL is greater than $\omega - 1$. Then, if a long queue is updated online (i.e., of length n), the expected real-time QL will be greater than $\omega - 1$ thereafter; see, e.g., the curve of $\phi = 6$ in Figure 3(c). All future connected customers who see this QL update will choose to leave the facility, so connected customers' long-run average individual utility U_C is zero. Meanwhile, unconnected customers' utility U_U is zero if $\rho_U \geq \rho_N$, or positive if $\rho_L \leq \rho_U < \rho_N$. In both instances, unconnected customers have no less utility than connected customers. If the unconnected customers' offered load is low (i.e., $\rho_U < \rho_L$), some connected customers will enter the facility. In this case, the expected QL when all unconnected customers enter the facility is below $\omega - 1$. For any QL update on the platform, connected customers who arrive sufficiently long after the update will expect the real-time QL to below $\omega - 1$. They will choose to enter the facility from time to time expecting positive utility. In this case, connected customers may have greater utility than unconnected customers.

We confirm the above intuition in Figure 9(c), which illustrates the simulation result of the difference between the connected and unconnected customers' utility $U_C - U_U$ under arrival rate $\Lambda = 2$ and hassle cost $h = 1.5$, for which the two critical cutoffs in unconnected customers' offered

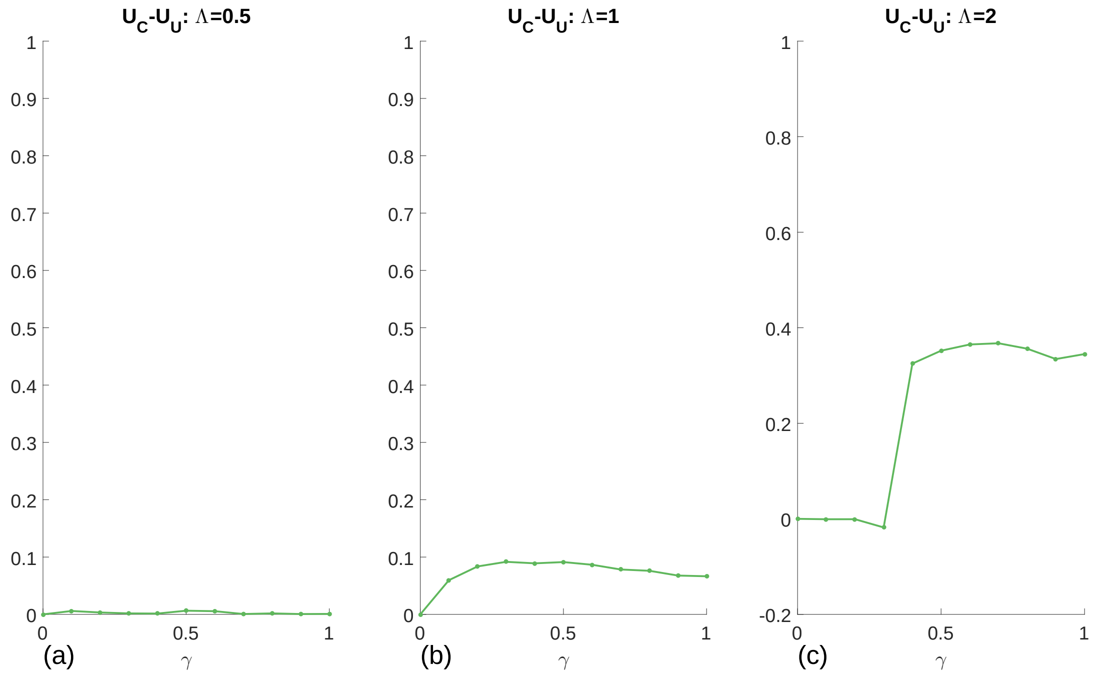

Figure 9 **Difference between connected and unconnected customers's individual utility $U_C - U_U$ as a function of social connectivity $\gamma \in \{0, 0.1, \ldots, 1\}$ under shared QLI with service reward $R = 10$, service rate $\mu = 1$, marginal waiting cost $c = 1$, arrival rate $\Lambda \in \{0.5, 1, 2\}$, and hassle cost $h = 1.5$.**

load ρ_U are $\rho_L = 1.336$ and $\rho_N = 1.404$. For $\gamma \le 0.2$, we have $\rho_U \ge 1.8 > \rho_N$, so both connected and unconnected customers obtain zero utility and $U_C - U_U = 0$. For $\gamma = 0.3$, we have ρ_U is in $(\rho_L, \rho_N]$. Here, unconnected customers have positive utility while connected customers have zero utility, which leads to $U_C - U_U < 0$. This is in line with our intuition above. For $\gamma \ge 0.4$, we have $\rho_U < \rho_L$, so some connected customers will enter the facility. In this case, we observe from Figure 9(c) that connected customers obtain greater utility than unconnected customers. Then, unconnected customers have incentive to become connected for any $\gamma \in [0.4, 1]$. This is also observed in Figures 9(a) and 9(b), where the total offered load ρ is smaller than ρ_L so $\rho_U = \gamma\rho$ is smaller than ρ_L for any $\gamma \in [0, 1]$.

We next endogenize customers' decision of becoming connected on this information-sharing platform to share and obtain the latest congestion update. We can view the degree of social connectivity γ as an outcome from customers' symmetric strategic behavior of deciding on whether to be connected on the platform. Imagine an information-sharing platform is established for a service facility originally under no QLI (i.e., $\gamma = 0$). If the total offered load is not high (see, e.g., Figures 9(a) and 9(b)), unconnected customers can obtain higher utility by becoming connected on the platform. Then, all customers would prefer to use the platform to share information. Otherwise, if the offered load is high (see, e.g., Figures 9(c)), customers may not join the platform spontaneously. Within the range of degrees of connectivity $\gamma \in [0, 1 - \rho_L/\rho]$, unconnected customers may relentlessly enter the facility expecting non-negative utility, which causes excessive congestion and drives away connected customers. In this case, the social planner needs to intervene to get sufficient customers connected with $\gamma > 1 - \rho_L/\rho$ so that the rest of the population will voluntarily follow, benefiting the social welfare as an outcome.

OA4. Proof of Lemma 1

During the time interval (T, t), no customers arrive but some customers may have their services completed and leave the system. (This will not be true if not all customers share information. See Section 4 for a relaxation of the all-sharing assumption.) Given ϕ customers in the queue at time T, the number of departures in (T, t), $D(\delta, \phi)$, has a probability mass function:

$$P\{D(\delta,\phi) = j\} = \begin{cases} \frac{e^{-\mu\delta}(\mu\delta)^j}{j!} & 0 \le j < \phi, \\ \sum_{k=\phi}^{\infty} \frac{e^{-\mu\delta}(\mu\delta)^k}{k!} & j = \phi, \end{cases}$$

and the expected number of departures in the time interval (T, t) is

$$\begin{aligned} E[D(\delta,\phi)] &= \sum_{k=0}^{\phi-1} k \frac{e^{-\mu\delta}(\mu\delta)^k}{k!} + \phi \sum_{k=\phi}^{\infty} \frac{e^{-\mu\delta}(\mu\delta)^k}{k!} \\ &= \sum_{k=0}^{\phi-1} k \frac{e^{-\mu\delta}(\mu\delta)^k}{k!} + \phi \left(1 - \sum_{k=0}^{\phi-1} \frac{e^{-\mu\delta}(\mu\delta)^k}{k!}\right) \\ &= \phi + \sum_{k=0}^{\phi-1} \frac{e^{-\mu\delta}(\mu\delta)^k}{k!}(k-\phi). \end{aligned}$$

Thus, the real-time QL at time t is the difference between ϕ and the number of departures during (T, t). Its distribution is

$$\begin{aligned} P\{\Psi(\delta,\phi) = k\} &= P\{D(\delta,\phi) = \phi - k\} \\ &= \begin{cases} \frac{e^{-\mu\delta}(\mu\delta)^{\phi-k}}{(\phi-k)!} & 0 < k \le \phi, \\ \sum_{k=\phi}^{\infty} \frac{e^{-\mu\delta}(\mu\delta)^k}{k!} & k = 0, \end{cases} \end{aligned}$$

and its expected value is

$$E[\Psi(\delta,\phi)] = \phi - E[D(\delta,\phi)] = \sum_{k=0}^{\phi-1} \frac{e^{-\mu\delta}(\mu\delta)^k}{k!}(\phi - k).$$

Note that when $t = T \Leftrightarrow \delta = 0$, we recover $E[\Psi(0,\phi)] = \phi$.

The first derivative of $E[\Psi(\delta,\phi)]$ with respect to δ is

$$\begin{aligned} \frac{\partial E[\Psi(\delta,\phi)]}{\partial\delta} &= -\sum_{k=0}^{\phi-1} \frac{(\phi-k)}{k!} \mu e^{-\mu\delta}(\mu\delta)^k + \sum_{k=1}^{\phi-1} \frac{(\phi-k)}{k!} \mu k e^{-\mu\delta}(\mu\delta)^{k-1} \\ &= -\mu \left(\sum_{k=0}^{\phi-1} (\phi-k) \frac{e^{-\mu\delta}(\mu\delta)^k}{k!} - \sum_{j=0}^{\phi-2} (\phi-j-1) \frac{e^{-\mu\delta}(\mu\delta)^j}{j!} \right) \\ &= -\mu \left(\frac{e^{-\mu\delta}(\mu\delta)^{\phi-1}}{(\phi-1)!} + \sum_{j=0}^{\phi-2} \frac{e^{-\mu\delta}(\mu\delta)^j}{j!} \right) \\ &= -\mu \sum_{j=0}^{\phi-1} \frac{e^{-\mu\delta}(\mu\delta)^j}{j!} \\ &= -\mu P\{D(\delta,\phi) < \phi\} < 0. \end{aligned}$$

Thus, $E[\Psi(\delta,\phi)]$ is strictly decreasing in δ. Clearly, $\delta = t - T$ is an increasing function of t, so we have that $E[\Psi(\delta,\phi)]$ is strictly decreasing in t. □

OA5. Proof of Lemma 2

(i) Recall from Lemma 1 that $E\left[\Psi\left(\delta,\phi\right)\right]$ is strictly decreasing in δ. Thus, when $\phi<n$, we have that $U_{enter}\left(\delta,\phi\right)$ is strictly increasing in δ. When $\phi=n$, by using the derivation of $\partial E\left[\Psi\left(\delta,\phi\right)\right]/\partial\delta$ in the proof of Lemma 1, we get the first derivative of $U_{enter}\left(\delta,\phi\right)$ with respect to δ

$$\frac{\partial U_{enter}\left(\delta,\phi\right)}{\partial\delta}=ce^{-\mu\delta}\left(\sum_{j=1}^{n-1}\frac{\left(\mu\delta\right)^{j}}{j!}+\nu-n\right),$$

which is positive because $\sum_{j=1}^{n-1}\left(\mu\delta\right)^{j}/j!\geq 0$ and $\nu\geq n$. Hence, $U_{enter}\left(\delta,\phi\right)$ increases with δ.

(ii) When $t=T$, we have $\delta=0$ and $E\left[\Psi\left(0,\phi\right)\right]=\phi$; substituting these into (2) yields

$$U_{enter}\left(0,\phi\right)=\begin{cases}\frac{c}{\mu}\left(\omega-1-\phi\right)\geq 0 & 1\leq\phi\leq m-1,\\ \frac{c}{\mu}\left(\omega-1-\phi\right)<0 & m\leq\phi<n,\\ -h \qquad\qquad\quad <0 & \phi=n.\end{cases}.$$

(iii) From (2), we have

$$U_{enter}\left(\delta,\phi\right)=\begin{cases}\frac{c}{\mu}\left(\omega-1-\sum_{k=0}^{\phi-1}\frac{e^{-\mu\delta}\left(\mu\delta\right)^{k}}{k!}\left(\phi-k\right)\right) & \phi<n,\\ \frac{c}{\mu}\left(\omega-1-\sum_{k=0}^{n-1}\frac{e^{-\mu\delta}\left(\mu\delta\right)^{k}}{k!}\left(n-k\right)-e^{-\mu\delta}\left(\nu-n-1\right)\right) & \phi=n.\end{cases}$$

Then, we can derive

$$\begin{aligned}&U_{enter}\left(\delta,\phi-1\right)-U_{enter}\left(\delta,\phi\right)\\ =&\begin{cases}\frac{c}{\mu}\left(\frac{e^{-\mu\delta}\left(\mu\delta\right)^{\left(\phi-1\right)}}{\left(\phi-1\right)!}+\sum_{k=0}^{\phi-2}\frac{e^{-\mu\delta}\left(\mu\delta\right)^{k}}{k!}\right) & 0<\phi<n,\\ \frac{c}{\mu}\left(\frac{e^{-\mu\delta}\left(\mu\delta\right)^{n-1}}{\left(n-1\right)!}+\sum_{k=1}^{n-2}\frac{e^{-\mu\delta}\left(\mu\delta\right)^{k}}{k!}+e^{-\mu\delta}\left(\nu-n\right)\right) & \phi=n,\end{cases}\end{aligned}$$

which is clearly positive. □

OA6. Proof of Lemma 3

(i) When $m\leq\phi<n$, we have $U_{enter}\left(\delta,\phi\right)=\frac{c}{\mu}\left(\omega-1-E\left[\Psi\left(\delta,\phi\right)\right]\right)$. From Lemma 1, we have that $E\left[\Psi\left(\delta,\phi\right)\right]$ is a decreasing function of δ. According to (3), τ_{ϕ} is the solution of the following equation of δ,

$$E\left[\Psi\left(\delta,\phi\right)\right]=\sum_{k=0}^{\phi-1}\frac{e^{-\mu\delta}\left(\mu\delta\right)^{k}}{k!}\left(\phi-k\right)=\omega-1.$$

From Lemma 1, we have $E\left[\Psi\left(\delta,\phi\right)\right]$ is strictly decreasing in δ. Due to the fact that the inverse of a decreasing function is still a decreasing function, $\tau_{\phi}=\Phi^{-1}\left(\omega-1\right)$ is also a decreasing function of ω.

When $\phi=n$, we have $U_{enter}\left(\delta,\phi\right)=\frac{c}{\mu}\left(\omega-1-E\left[\Psi\left(\delta,n\right)\right]-e^{-\mu\delta}\left(\nu-n-1\right)\right)$, which is increasing in δ from Lemma 2. Then, from (3), τ_{n} is the *unique* solution of the following equation,

$$E\left[\Psi\left(\delta,n\right)\right]+e^{-\mu\delta}\left(\nu-n-1\right)=\nu-\frac{h\mu}{c}-1.$$

Note that the left hand side

$$E\left[\Psi\left(\delta,n\right)\right]+e^{-\mu\delta}\left(\nu-n-1\right)=\omega-1-\frac{\mu}{c}U_{enter}\left(\delta,\phi\right)\text{ for }\phi=n$$

is decreasing in δ from Lemma 2(i); its first derivative

$$\frac{\partial}{\partial\delta}\left(E\left[\Psi\left(\delta,n\right)\right]+e^{-\mu\delta}\left(\nu-n-1\right)\right)=-\mu e^{-\mu\delta}\left(\sum_{j=1}^{n-1}\frac{\left(\mu\delta\right)^{j}}{j!}+\nu-n\right)$$

is negative and decreases with ν. The right hand side $\nu-\frac{h\mu}{c}-1$ increases with ν. Hence, τ_n is decreasing in ν.

(ii & iii) Using similar argument as in (i), we have τ_n increases in the hassle cost h, and τ_n decreases in ω.

(iv) Recall from Lemma 2(iii) that $U_{enter}\left(\delta,\phi\right)$ is strictly decreasing in ϕ. Thus, we have τ_ϕ, as the unique solution of $U_{enter}\left(\delta,\phi\right)=0$, increases in ϕ for $m\leq\phi\leq n$. □

OA7. Proof of Proposition 1

It is essential to derive the stationary distribution of the semi-Markov process with states $1\leq\phi\leq n$. Let $\mathcal{P}_{i,j}$ denote the probability that the next transition cycle will start with j customers given that the current one starts with i customers, for $1\leq i,j\leq n$; and let $\mathcal{P}$ denote the matrix of transition probabilities with entries $\mathcal{P}_{i,j}$. Let $\vec{\pi}=\left[\pi_1,\pi_2,\ldots,\pi_n\right]$ denote a row vector of the stationary probability that a transition cycle will start with i customers π_i. Clearly, $\vec{\pi}$ should be the unique solution of

$$\vec{\pi}\cdot\mathcal{P}=\vec{\pi},\qquad\text{and}\qquad\vec{\pi}\cdot\vec{1}=1,\tag{OA. 1}$$

where $\vec{1}=\left[1,1,\ldots,1\right]^{\top}$ is a column vector of the same size as $\vec{\pi}$. Once the probability matrix $\mathcal{P}$ is known, we can derive $\vec{\pi}$ by solving (OA. 1).

We next discuss these two time intervals in a transition cycle separately with more details in the next two sections. Let B_j^i denote the number of customers in the beginning of the time interval j, given i customers in the beginning of a transition cycle, for $1\leq i\leq n$ and $j=1,2$. Recall that I_j^i denotes the length of the time interval j given i customers in the beginning of a transition cycle, for $1\leq i\leq n$ and $j=1,2$. Let $P_{i,j}^{(1)}$ denote the probability that there are j customers at time $T_i+\tau_\phi$ given ϕ customers at time T_i for $1\leq i,j\leq n$, and $P_{j,k}^{(2)}$ denote the probability that there are k customers at time T_{i+1} given that there are j customers at time $T_i+\tau_\phi$ for $0\leq j\leq n$ and $1\leq j\leq n$. Let $P^{(1)}$ and $P^{(2)}$ denote the matrices of one-step transition probabilities $P_{i,j}^{(1)}$ and $P_{i,j}^{(2)}$, respectively.

Time Interval 1 $\left(T_i,T_i+\tau_\phi\right)$: The length of this time interval is a constant τ_ϕ, i.e.,

$$E\left[I_1^{\phi}\right]=\tau_\phi.\tag{OA. 2}$$

By the definition of the transition point T_i we have $B_1^{\phi}=\phi$. The number of service completions in this time interval determines the expected total customer waiting time. Let D_1 denote the number of service completions in this time interval. Due to exponential service times with rate μ, D_1 follows Poisson distribution with parameter $\mu\tau_\phi$:

$$P\left\{D_1=k\right\}=\frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^{k}}{k!}\text{ for }k=0,1,2,\ldots,\phi.$$

From Theorem 5.2 of Ross (2006), given $D_1=k$, the k service completion times have the same distribution as the order statistics corresponding to k independent random variables uniformly distributed on the interval $(0,\tau_\phi)$. Further, given $D_1=k$, the expected j-th service completion time is $\frac{j\tau}{k+1}$, for $j=1,\dots,k$.

Note that all quantities in this time interval are independent of customer arrival rate Λ. This is because the arrival process is effectively shut down in this time interval $(T_i,T_i+\tau_\phi)$.

Time Interval 2 $[T_i+\tau_\phi,T_{i+1}]$: The number of customers in the beginning of this time interval, B_2^ϕ, depends on D_1:

$$B_2^\phi=\max\left(\phi-D_1,0\right).$$

Thus, B_2^ϕ follows distribution

$$P\left\{B_2^\phi=k\right\}=\begin{cases}\sum_{j=\phi}^{\infty}\frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^j}{j!} & k=0,\\ \frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^{\phi-k}}{(\phi-k)!} & 0<k\le\phi.\end{cases}$$

The one-step transition matrix $\mathcal{P}^{(1)}$ is directly determined by the distribution of B_2^ϕ:

$$\mathcal{P}_{i,j}^{(1)}=\begin{cases}\sum_{l=i}^{\infty}\frac{e^{-\mu\tau_i}(\mu\tau_i)^l}{l!} & \text{if } m\le i\le n \text{ and } j=0\\ \frac{e^{-\mu\tau_i}(\mu\tau_i)^{i-j}}{(i-j)!} & \text{if } m\le i\le n \text{ and } 0<j\le i\\ 1 & \text{if } i=j\le m-1\\ 0 & \text{otherwise}\end{cases}. \qquad \text{(OA. 3)}$$

Let $N_{\mu\tau_i}^j=\frac{e^{-\mu\tau_i}(\mu\tau_i)^j}{j!}$, and we have

$$\mathcal{P}^{(1)}=\begin{bmatrix} i\diagdown j & j=0 & j=1 & j=2 & \cdots & j=m-1 & j=m & \cdots & j=n-1 & j=n\\ i=0 & 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0\\ i=1 & 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0\\ i=2 & 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots\\ i=m-1 & 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0\\ i=m & \sum_{l=m}^{\infty}N_{\mu\tau_m}^l & N_{\mu\tau_m}^{m-1} & N_{\mu\tau_m}^{m-2} & \cdots & N_{\mu\tau_m}^1 & N_{\mu\tau_m}^0 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots\\ i=n-1 & \sum_{l=n-1}^{\infty}N_{\mu\tau_{n-1}}^l & N_{\mu\tau_{n-1}}^{n-2} & N_{\mu\tau_{n-1}}^{n-3} & \cdots & N_{\mu\tau_{n-1}}^{n-m} & N_{\mu\tau_{n-1}}^{n-m-1} & \cdots & N_{\mu\tau_{n-1}}^0 & 0\\ i=n & \sum_{l=n}^{\infty}N_{\mu\tau_n}^l & N_{\mu\tau_n}^{n-1} & N_{\mu\tau_n}^{n-2} & \cdots & N_{\mu\tau_n}^{n-m+1} & N_{\mu\tau_n}^{n-m} & \cdots & N_{\mu\tau_n}^1 & N_{\mu\tau_n}^0 \end{bmatrix}.$$

The length of the second time interval is $\exp(\Lambda)$, so we have

$$E\left[I_2\right]=\frac{1}{\Lambda}.$$

Let D_2^ϕ denote the number of service completions in $[T_i+\tau_\phi,T_i+\tau_\phi+\exp(\Lambda)]$. Due to the exponential service time and there are B_2^ϕ customers at time $T_i+\tau_\phi$, we can determine the distribution of D_2^ϕ by considering a Poisson process with rate μ in an exponential time interval with mean $\frac{1}{\Lambda}$:

$$P\left\{D_2^\phi=k\right\}=\begin{cases}\frac{\rho}{\rho+1}\left(\frac{1}{\rho+1}\right)^k & 0\le k<B_2^\phi,\\ \left(\frac{1}{\rho+1}\right)^{B_2^\phi} & k=B_2^\phi.\end{cases} \qquad \text{(OA. 4)}$$

Further, the number of customers seen by the arrival at time $T_i+\tau_\phi+\exp(\Lambda)$ is $B_2^\phi-D_2^\phi$.

Further, the above discussion reveals the relation between the one-step transition matrix $\mathcal{P}^{(2)}$ and the distribution of D_2:

$$\mathcal{P}^{(2)}_{j,k} = \begin{cases} P\{D_2 = 1 \text{ or } 2\} & = \frac{\rho(\rho+2)}{(\rho+1)^2} & \text{if } j = k = n \\ \sum_{l=j}^{\infty} P\{D_2 = l\} & = \left(\frac{1}{\rho+1}\right)^j & \text{if } k = 1 \\ P\{D_2 = j-k+1\} & = \frac{\rho}{\rho+1}\left(\frac{1}{\rho+1}\right)^{j-k+1} & \text{if } k \le j+1 \\ 0 & 0 & \text{otherwise} \end{cases}. \tag{OA. 5}$$

Moreover, the one-step transition matrix of transition cycles $\mathcal{P}$ can be derives as

$$\mathcal{P} = \mathcal{P}^{(1)} \cdot \mathcal{P}^{(2)}, \tag{OA. 6}$$

where $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are from (OA. 3) and (OA. 5), respectively. Then, solving (OA. 1) gives the stationary probability of the semi-Markov process $\vec{\pi}$.

Then, by Lemma 5 in Section OA20 in the Online Appendix, we obtain

$$\bar{\Lambda}^S = \frac{1}{\sum_{i=1}^n \pi_i \tau_i + \frac{1}{\Lambda}}, \tag{OA. 7}$$

$$\Lambda^S = \frac{1 - \pi_n \frac{\rho}{\rho+1} e^{-\mu\tau_n}}{\sum_{i=1}^n \pi_i \tau_i + \frac{1}{\Lambda}}, \tag{OA. 8}$$

$$S^S = \frac{\sum_{i=1}^n \pi_i \left(R - c\frac{i}{\mu}\right) - \pi_n \frac{\rho}{\rho+1} e^{-\mu\tau_n}\left(R - c\frac{n}{\mu}\right) - h}{\sum_{i=1}^n \pi_i \tau_i + \frac{1}{\Lambda}}, \tag{OA. 9}$$

where π_i can be solved from (OA. 1) with $\mathcal{P}$ given by (OA. 6). □

OA8. Proof of Proposition 2

In the first decision epoch, customers do not have any QLI, it is an unobservable system for them. Assume they use a mixed strategy and enter the facility with a probability q, so the effective arrival rate to the service system is $\lambda = q\Lambda$. In the second decision epoch, customers have access to the real-time QLI, and they decide accordingly whether to join or balk. It is an observable queue for them, and they will only join the queue if it is shorter than n. Thus, a customer's utility of entering the facility is

$$U_{enter}(\delta, \phi) = \sum_{i=0}^{n-1} \frac{(1-q\rho)(q\rho)^i}{1-(q\rho)^{n+1}} \left(R - (i+1)\frac{c}{\mu}\right) - h = H(q\rho) - h$$

where $H(\rho) \equiv \frac{c}{\mu}\left(\nu\frac{(1-\rho^n)}{1-\rho^{n+1}} - \frac{1}{1-\rho} + \frac{(n+1)\rho^n}{1-\rho^{n+1}}\right)$. Then, if all going provides positive utility, all customers will go to the facility, i.e., $q = 1$. We first notice some properties of $H(\rho)$: (i) $H(0) = R - \frac{c}{\mu} > 0$; (ii) it $\lim_{\rho\to\infty} H(\rho) = 0$; and (iii) $H(\rho)$ is a decreasing function of ρ, if $\frac{1-\rho^i}{1-\rho^{n+1}}$ is a decreasing function of $\rho > 0$, for $\forall 1 \le i \le n$, which is true because its first derivative

$$\frac{\partial\left(\frac{1-\rho^i}{1-\rho^{n+1}}\right)}{\partial\rho} = \frac{i(n+1)(1-\rho)}{\rho^{(1-i)}(\rho^{n+1}-1)^2}\left(\frac{\sum_{j=n+1-i}^n \rho^j}{i} - \frac{\sum_{j=0}^n \rho^j}{n+1}\right)$$

is less than zero. Hence, for $0 < h \le R - \frac{c}{\mu}$, there is a unique root ρ_N of $H(\rho) = h$. Of course, when h decreases to 0, ρ_N increases to ∞; and when h increases to $R - \frac{c}{\mu}$, ρ_N decreases to 0.

When $\rho<\rho_N$, all Λ customers go to the facility, $\frac{1-\rho^n}{1-\rho^{n+1}}\Lambda$ will join the queue and $\frac{(1-\rho)\rho^n}{1-\rho^{n+1}}\Lambda$ will balk at the second decision epoch. In this case, the social welfare is $\mu\rho\left(H\left(\rho\right)-h\right)$. When $\rho\geq\rho_N$, not all but ρ_N customers will go to the facility, and social welfare is zero.

The expected QL for $\rho<\rho_N$ under no QLI L_n can be derived as the expected QL of an M/M/1 queue with finish waiting room n:

$$L_n=\frac{\rho\left(1-\rho^n-n\rho^n+n\rho^{n+1}\right)}{\left(1-\rho\right)\left(1-\rho^{n+1}\right)},$$

which is clearly an increasing function of ρ. Simplifying $L_n=\omega-1$ gives $\frac{c}{\mu}\left(\nu-1-L_n\right)=h$, where $\frac{c}{\mu}\left(\nu-1-L_n\right)=h$ is a decreasing function of ρ. Hence, ρ_L is unique.

Moreover, we have

$$\frac{c}{\mu}\left(\nu-1-L_n\right)-H\left(\rho\right)=-\frac{c}{\mu}\frac{\left(n+1-\nu\right)\rho^n\left(1-\rho\right)}{1-\rho^{n+1}}<0.$$

Combining with the fact that both $\frac{c}{\mu}\left(\nu-1-L_n\right)$ and $H\left(\rho\right)$ are decreasing functions of ρ, we have $\rho_L<\rho_N$. □

OA9. Proof of Proposition 3

Recall from the discussion of Lemma 3 that queues under the full-QLI structure have the joining shutdowns that are different from the arrival shutdowns under the shared-QLI structure. However, we can use a similar approach to Proposition 1 here to derive the entry rate, throughput, and social welfare under full QLI, and the result is identical to Proposition 3.

We consider each arrival of customer as transition point and apply the same method as in the proof of Proposition 1 to the two time intervals: (1) $\left(T_i,T_i+\exp\left(\mu\right)\right)$, and (2) $\left[T_i+\exp\left(\mu\right),T_{i+1}\right]$, where $T_{i+1}=T_i+\exp\left(\mu\right)+\exp\left(\Lambda\right)$. Let $B_i\leq m$ denote the number of customers in the beginning of the time interval i and D_i denote the number of service completions in time interval i, for $i=1,2$.

For the first time interval $\left(T_i,T_i+\exp\left(\mu\right)\right)$, we have

- $B_1=\phi$.
- The length of this time interval is $\tau_\phi=\begin{cases}0 & \phi<m,\\ \exp\left(\mu\right) & \phi=m.\end{cases}$
- The number of service completion in this time interval is $D_1=\begin{cases}0 & \phi<m,\\ 1 & \phi=m.\end{cases}$
- The one step transition matrix of this time interval is

$$\mathcal{P}^{(1)}_{i,j}=\begin{cases}1 & \text{if } i<m \text{ and } i=j\\ 1 & \text{if } i=m \text{ and } j=m-1\\ 0 & \text{otherwise}\end{cases}$$

Further, for the second time interval $\left[T_i+\tau_\phi,T_{i+1}\right]$, we have

- $B_2=\begin{cases}\phi & \phi<m,\\ m-1 & \phi=m.\end{cases}$
- The expected length of this time interval is $1/\Lambda$.
- We can determine the distribution of the number of service completions in this time interval

D_2 by considering a Poisson process with rate μ in an exponential time interval with mean $\frac{1}{\Lambda}$:

$$P\left\{D_2=k\right\}=\begin{cases}\frac{\rho}{\rho+1}\left(\frac{1}{\rho+1}\right)^k & 0\leq k<B_2,\\ \left(\frac{1}{\rho+1}\right)^{B_2} & k=B_2.\end{cases}\qquad\text{(OA. 10)}$$

- Let $P_{j,k}$ denote the probability that there are k customers at time T_{i+1} given that there are j customers at time $T_i + \tau_\phi$ for $1 \le j, k \le m$. Further, we have the relation between the one-step transition matrix $\mathcal{P}^{(2)}$ and the distribution of D_2:

$$\mathcal{P}^{(2)}_{j,k} = \begin{cases} \sum_{l=j}^{\infty} P\{D_2 = l\} & = \left(\frac{1}{\rho+1}\right)^j & \text{if } k = 1 \\ P\{D_2 = j-k+1\} & = \frac{\rho}{\rho+1}\left(\frac{1}{\rho+1}\right)^{j-k+1} & \text{if } 1 < k \le j \\ 0 & = 0 & \text{otherwise} \end{cases}.$$

$$\begin{aligned} \mathcal{P} &= \mathcal{P}^{(1)} \cdot \mathcal{P}^{(2)} \\ &= \left[\begin{array}{ccccc} i \diagdown k & k=1 & k=2 & \cdots & k=m \\ i=1 & \frac{1}{\rho+1} & \frac{\rho}{\rho+1} & \cdots & 0 \\ i=2 & \left(\frac{1}{\rho+1}\right)^2 & \frac{\rho}{\rho+1}\left(\frac{1}{\rho+1}\right) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ i=m-1 & \left(\frac{1}{\rho+1}\right)^{m-1} & \frac{\rho}{\rho+1}\left(\frac{1}{\rho+1}\right)^{m-2} & \cdots & \frac{\rho}{\rho+1} \\ i=m & \left(\frac{1}{\rho+1}\right)^{m-1} & \frac{\rho}{\rho+1}\left(\frac{1}{\rho+1}\right)^{m-2} & \cdots & \frac{\rho}{\rho+1} \end{array}\right] \end{aligned}$$

Let π_i denote the stationary probability that a transition cycle will start with i customers, and $\vec{\pi} = [\pi_1, \pi_2, \ldots, \pi_m]$ is a row vector. Clearly, $\vec{\pi}$ should be the unique solution of

$$\vec{\pi} \cdot \mathcal{P} = \vec{\pi}, \quad \text{and} \quad \vec{\pi} \cdot \vec{1} = 1, \tag{OA. 11}$$

where $\vec{1} = [1, 1, \ldots, 1]^\top$ is a column vector of the same size as $\vec{\pi}$. Once the probability matrix $\mathcal{P}$ is known, we can derive $\vec{\pi}$ by solving the above equation:

$$\pi_i = \frac{\rho^{i-1}(1-\rho)}{1-\rho^m} \text{ for } 1 \le i \le m.$$

Then, by Lemma 5 in Section OA20 in the Online Appendix, the throughput rate is one divided by the expected transition cycle length

$$\Lambda^F = \frac{1}{\sum_{i=1}^m \pi_i (\tau_i + 1/\Lambda)} = \frac{\Lambda}{\left(1 - \frac{\rho^{m-1}(1-\rho)}{1-\rho^m}\right) + \frac{\rho^{m-1}(1-\rho)}{1-\rho^m}(\rho+1)} = \mu\frac{\rho(1-\rho^m)}{1-\rho^{m+1}}.$$

The social welfare is

$$S^F = \frac{\sum_{i=1}^m \pi_i \left(R - h - c\frac{i}{\mu}\right)}{\sum_{i=1}^m \pi_i (\tau_i + 1/\Lambda)} = c\rho\left(\frac{1-\rho^m}{1-\rho^{m+1}}\omega - \left(\frac{1}{1-\rho} - \frac{(m+1)\rho^m}{1-\rho^{m+1}}\right)\right).$$

At last, using Naor's formula $S^F = (R-h)\Lambda^F - cL^F$, we have

$$L^F = \frac{\rho}{1-\rho} - \frac{(m+1)\rho^{m+1}}{1-\rho^{m+1}}. \quad \square$$

OA10. Proof of Corollary 1

When $\Lambda = \infty$, we have the one-step transition matrix $\mathcal{P}^{(1)}$ as

$$\mathcal{P}^{(1)}_{i,j} = \begin{cases} \sum_{l=i}^{\infty} \frac{e^{-\mu\tau_i}(\mu\tau_i)^l}{l!} & \text{if } m \leq i \leq n \text{ and } j = 0 \\ \frac{e^{-\mu\tau_i}(\mu\tau_i)^{i-j}}{(i-j)!} & \text{if } m \leq i \leq n \text{ and } 0 < j \leq i \\ 1 & \text{if } i = j \leq m-1 \\ 0 & \text{otherwise} \end{cases}$$

The length of the second time interval is zero, and there are no service completions. Further, the one-step transition matrix $\mathcal{P}^{(2)}$ is

$$\mathcal{P}^{(2)}_{j,k} = \begin{cases} 1 & \text{if } j = k = n \\ 1 & \text{if } k = j+1 \\ 0 & \text{otherwise} \end{cases}.$$

Moreover, the one-step transition matrix of transition cycles $\mathcal{P}$ can be derives as

$$\mathcal{P} = \mathcal{P}^{(1)} \cdot \mathcal{P}^{(2)}$$
$$= \left[\begin{array}{cccccccccc} i \diagdown j & j=0 & j=1 & j=2 & \cdots & j=m-1 & j=m & \cdots & j=n-1 & j=n \\ i=0 & 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ i=1 & 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ i=m-1 & 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 & 0 \\ i=m & 0 & \sum_{l=m}^{\infty} N^l_{\mu\tau_m} & N^{m-1}_{\mu\tau_m} & \cdots & N^2_{\mu\tau_m} & N^1_{\mu\tau_m} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ i=n-2 & 0 & \sum_{l=n-2}^{\infty} N^l_{\mu\tau_{n-2}} & N^{n-3}_{\mu\tau_{n-2}} & \cdots & N^{n-m}_{\mu\tau_{n-2}} & N^{n-m-1}_{\mu\tau_{n-2}} & \cdots & N^0_{\mu\tau_{n-2}} & 0 \\ i=n-1 & 0 & \sum_{l=n-1}^{\infty} N^l_{\mu\tau_{n-1}} & N^{n-2}_{\mu\tau_{n-1}} & \cdots & N^{n-m+1}_{\mu\tau_{n-1}} & N^{n-m}_{\mu\tau_{n-1}} & \cdots & N^1_{\mu\tau_{n-1}} & N^0_{\mu\tau_{n-1}} \\ i=n & 0 & \sum_{l=n}^{\infty} N^l_{\mu\tau_n} & N^{n-1}_{\mu\tau_n} & \cdots & N^{n-m+2}_{\mu\tau_n} & N^{n-m+1}_{\mu\tau_n} & \cdots & N^2_{\mu\tau_n} & N^1_{\mu\tau_n} + N^0_{\mu\tau_n} \end{array}\right],$$

where $N^j_{\mu\tau_i} = \frac{e^{-\mu\tau_i}(\mu\tau_i)^j}{j!}$. Then, time both sides of (OA. 1) with $\left[0\ 1 \cdots m-1\ m \cdots n-1\ n\right]^\top$ gives

$$\vec{\pi}\mathcal{P}\left[0\ 1 \cdots m-1\ m \cdots n-1\ n\right]^\top = \vec{\pi}\left[0\ 1 \cdots m-1\ m \cdots n-1\ n\right]^\top$$
$$\sum_{i=0}^{n} i\pi_i + 1 - \sum_{i=m}^{n} \mu\tau_i\pi_i - e^{-\mu\tau_n}\pi_n + \sum_{\phi=m}^{n} \pi_\phi \sum_{k=\phi}^{\infty} \frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^k}{k!}(k-\phi) = \sum_{i=0}^{n} i\pi_i$$

from which we obtain the expected cycle length

$$\sum_{i=m}^{n} \pi_i\tau_i = \frac{1}{\mu}\left(1 - \pi_n e^{-\mu\tau_n} + \sum_{\phi=m}^{n} \pi_\phi \sum_{k=\phi}^{\infty} \frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^k}{k!}(k-\phi)\right).$$

Substituting $\sum_{i=m}^{n} \pi_i\tau_i$, $\Lambda = \infty$, and $\frac{\rho}{\rho+1} = 1$ into (OA. 7), (OA. 8), and (OA. 9) gives

$$\bar{\Lambda}^S = \mu/\Xi, \tag{OA. 12}$$
$$\Lambda^S = \mu\left(1 - \pi_n e^{-\mu\tau_n}\right)/\Xi, \tag{OA. 13}$$
$$S^S = c\left(\omega - \sum_{i=1}^{n} i\pi_i - \pi_n e^{-\mu\tau_n}(\nu - n)\right)/\Xi \tag{OA. 14}$$

where $\Xi = 1 - \pi_n e^{-\mu\tau_n} + \sum_{\phi=m}^{n} \pi_\phi \sum_{k=\phi}^{\infty} \frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^k}{k!}(k-\phi)$ and π_i can be solved from (OA. 1) with $\mathcal{P}$ given by (OA. 6).

When $R \to \infty$, we have ω and m approach ∞, so $\sum_{i=m}^{n} \pi_i \tau_i = \frac{1}{\mu}\left(1 - \pi_n e^{-\mu \tau_n}\right)$. Furthermore, when $n - m$ is also large, we have $\sum_{i=m}^{n} \pi_i \tau_i = \frac{1}{\mu}$.

For any transition cycles starting with more than $m-1$ customers, there are arrival shutdown periods, and the customer who joins the queue at the end of the transition cycle expects to obtain zero utility. Of course, for other transition cycles starting with no more than $m-1$ customers, the customers joining at the end of the cycle will obtain non-negative utility. By contradiction, we can show that there are positive probabilities for any transition cycle to start with no more than $m-1$ customers. Thus, the social welfare S^S in this situation is positive.

Moreover, when the hassle cost is $h = R - 2c/\mu$, the customer joining a transition cycle starting with $m - 1 = 1$ customer obtains zero utility. Recall that there is at least one customer in the beginning of any transition cycle. In this case, the social welfare is zero. □

OA11. Proof of Corollary 2

When $\rho = \Lambda/\mu = \infty$, we should always have $\rho_N < \rho$. Using Proposition 2, we complete the proof. □

OA12. Proof of Corollary 3

The proof is immediate by letting ρ approach ∞ in Proposition 3. □

OA13. Proof of Proposition 4

When the hassle cost is zero, i.e., $h = 0$, we have $\omega = \nu$ and $m = n$.

(i) Under the shared-QLI structure, when $m = n$, we have $\tau_i = 0$ for $i < n$. Further, $\tau_n = 0$ is the unique solution of

$$U_{enter}\left(\delta, \phi\right) = 0$$
$$\sum_{k=1}^{n-1} \frac{e^{-\mu\delta}\left(\mu\delta\right)^k}{k!}\left(n-k\right) + e^{-\mu\delta}\left(\nu - n - 1\right) = \nu - 1.$$

Then, there are no arrival shutdown periods, and all customers enter the facility to access the real-time QL. This completes the proof.

(ii) Under the no-QLI structure, when $h = 0$, from the proof of Proposition 2 we have $\rho_N = \infty$. Then, we must have $\rho < \rho_N$ in Proposition 2. This completes the proof.

(iii) For the full QLI setting, by letting $\omega = \nu$ and $m = n$ in Proposition 3, we complete the proof.

$$\begin{aligned} \Lambda &= \Lambda \frac{1-\rho^n}{1-\rho^{n+1}}, \\ S &= c\rho\left(\frac{1-\rho^n}{1-\rho^{n+1}}\nu - \left(\frac{1}{1-\rho} - \frac{(n+1)\rho^n}{1-\rho^{n+1}}\right)\right). \quad \square \end{aligned}$$

OA14. Proof of Theorem 1

When $S^N = 0$, we only need to prove $S^F > 0$:

$$\begin{aligned} S^F &= c\rho\left(\frac{1-\rho^m}{1-\rho^{m+1}}\omega - \left(\frac{1}{1-\rho} - \frac{(m+1)\rho^m}{1-\rho^{m+1}}\right)\right) \\ &= c\rho\left(\frac{(1-\rho)\sum_{i=0}^{m-1}\left(1-\rho^i\right)}{\left(1-\rho^{m+1}\right)(1-\rho)} + \langle\omega\rangle\frac{1-\rho^m}{1-\rho^{m+1}}\right) \\ &> 0. \end{aligned}$$

When $S^N > 0$, we can derive

$$S^F - S^N = c\rho\left(\left(\nu\frac{\rho^n(1-\rho)}{1-\rho^{n+1}} - \frac{(n+1)\rho^n}{1-\rho^{n+1}}\right) - \left(\omega\frac{\rho^m(1-\rho)}{1-\rho^{m+1}} - \frac{(m+1)\rho^m}{1-\rho^{m+1}}\right)\right).$$

Then, due to the fact that $\nu \geq \omega$, to prove $S^F > S^N$ is equivalent to prove that $\Phi(\nu) = \nu\frac{\rho^n(1-\rho)}{1-\rho^{n+1}} - \frac{(n+1)\rho^n}{1-\rho^{n+1}}$ is an increasing function of ν. First, it is clear that for $\nu \in [n, n+1)$, it is an increasing function of ν, because $\frac{\rho^n(1-\rho)}{1-\rho^{n+1}} > 0$. Then, we only need to show that $\Phi(\nu)$ has upward jumps at integer values; i.e.,

$$\begin{aligned}
&\Phi(n+1) - \lim_{\nu\nearrow n+1}\Phi(\nu)\\
&= \left((n+1)\frac{\rho^{n+1}(1-\rho)}{1-\rho^{n+1+1}} - \frac{(n+1+1)\rho^{n+1}}{1-\rho^{n+1+1}}\right) - \left((n+1)\frac{\rho^n(1-\rho)}{1-\rho^{n+1}} - \frac{(n+1)\rho^n}{1-\rho^{n+1}}\right)\\
&= \frac{\rho^{n+1}(1-\rho)\sum_{i=1}^{n}(1-\rho^i)}{(1-\rho^{n+2})(1-\rho^{n+1})} > 0,
\end{aligned}$$

which completes the proof. □

OA15. Proof of Theorem 2

Under no information, the expected social welfare is zero if customers enter the facility with a probability less than one. In this case, we must have the social welfare under shared QLI, which is positive, is greater than that under no information. Next, we compare the social welfare under shared and no information structures when *all* customers enter the facility under no information.

We will establish the fact that the shared-QLI structure dominates the no information structure in social welfare using a sample path discussion. Specifically, we introduce the shared QLI to any sample path under no information, and show that the social welfare will be improved as a result. Note that under no information, the system operates like an M/M/1 queue with a finite waiting room.

Note that the steady state operations of an M/M/1 queue is composed of a sequence of busy periods. In one busy period of an M/M/1 queue, if we take away an arrival from the arrival process while keeping everything else the same, all the following customers that arrive in this busy period will not see a longer queue upon arrival compared to the previous sample path with this arrival.

We formally describe this result. Consider a set of customers $\{1, 2, 3, \ldots\}$ ordered by their arrival times. Let a_i denote the arrival time of customer i, and s_i denote the i^{th} service completion time. Since we consider a busy period of the M/M/1 queue, the server is busy all the time, each service completion takes away one customer. Thus, s_i is also the departure time of customer i. Clearly, a customer departs after her arrival time; i.e., $s_i \geq a_i$. We assume that the system starts with no customers, then the first busy period of this M/M/1 queue ends at s_B, s.t., $B = \min\{i | s_i < a_{i+1}\}$. Let l_i denote the number of customers seen by customer i.

Then, we remove a random customer j from the previous process while letting all other customers $\{1, 2, 3, \ldots, B\} \setminus j$ behave the same as before. Let $l_i^{'}$ denote the number of customers seen by customer i. Clearly, for those customers who arrive before customer j, their l_i stay the same; i.e., $l_i^{'} = l_i$ if $i < j$. For those customers who arrive after customer j, they will not see a longer queue; i.e., $l_i^{'} \leq l_i$ if $i > j$, and they expect identical or higher utility by joining the queue.

We shall obtain the same conclusion when the M/M/1 queue has a finite waiting room and we remove several consecutive customers from the arrival process at the same time. Note that removing customers from a busy period does not change operations of the following busy periods.

We then focus on a busy period of the system under no information. Clearly, if one can show that the shared-QLI structure dominates the no information structure in one busy period, one can show that the same conclusion holds for an M/M/1 queue in steady state, which is composed of infinitely many busy periods.

There are infinite potential sample paths of the first busy period. In some of these sample paths, the QL never reaches m. Then, even if the shared QLI is introduced to the system, it will not stop any customers from entering the system, and the service facility operates the same as under no information. The social welfare under these two information structures is identical in this case.

We next discuss other sample paths where the QL reaches m. We categorize these sample paths into subsets according to the path before the QL first reaches m. For sample paths in one subset, the system operates exactly the same under both information structures until the QL first reaches m. Then, under the shared QLI, an arrival shutdown period is initiated and some customers may choose not to enter the facility, while under no information, all customers enter the facility. Introducing shared QLI to the system is equivalent to removing arrivals during the arrival shutdown period. If there are no arrivals during the arrival shutdown period, the shared QLI does not remove any arrivals from the system. If there are some arrivals during the arrival shutdown period, removing them will improve the utility of all customers that arrive after the arrival shutdown period in any sample paths in this subset. Moreover, for those arrivals during the arrival shutdown period, on average they obtain negative utility by entering the facility, so removing these arrivals will also improve the social welfare. Hence, at the incidence when the QL first reaches m, it would be socially desirable to provide customers with the shared QLI. Then, we can apply the same discussion to all the following incidences during the busy period when the QL becomes no less than m and show that the shared-QLI structure improves the social welfare of one busy period under no information. □

OA16. Proof of Theorem 3

In the asymptotic case $\Lambda = \infty$, from Corollary 3, we have $S^F = c\left(\frac{(R-h)\mu}{c} - m\right)$; and from Corollary 1, we have $S^S > 0$ when $h < R - 2c/\mu$. Hence, there exists a range $h \in \left(R - \frac{c}{\mu}i - \epsilon, R - \frac{c}{\mu}i\right]$ for $i = 3, \ldots, n-1$, for an $\epsilon > 0$, in which $S^S > S^F$. □

OA17. Proof of Proposition 5

In the asymptotic case $\Lambda = \infty$, from Corollary 3, we have $\Lambda^F = \mu$; and from Corollary 2, we have $\Lambda^N = \mu \frac{\rho_N\left(1-\rho_N^n\right)}{1-\rho_N^{n+1}}$. Note from the fact that

$$\frac{\rho\left(1-\rho^n\right)}{1-\rho^{n+1}} = \frac{\sum_{i=1}^n \rho^i}{\sum_{i=0}^n \rho^i} < 1,$$

we have $\Lambda^F > \Lambda^N$. □

OA18. Proof of Theorem 4

(i) In the asymptotic case $\Lambda=\infty$, from Corollary 1, we have

$$\mu\frac{1-\pi_n e^{-\mu\tau_n}}{1-\pi_n e^{-\mu\tau_n}+\sum_{\phi=m}^{n}\pi_\phi\sum_{k=\phi}^{\infty}\frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^k}{k!}(k-\phi)};$$

and from Corollary 3, we have $\Lambda^F=\mu$. From the fact that $\sum_{\phi=m}^{n}\pi_\phi\sum_{k=\phi}^{\infty}\frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^k}{k!}(k-\phi)>0$, we have $\frac{1-\pi_n e^{-\mu\tau_n}}{1-\pi_n e^{-\mu\tau_n}+\sum_{\phi=m}^{n}\pi_\phi\sum_{k=\phi}^{\infty}\frac{e^{-\mu\tau_\phi}\left(\mu\tau_\phi\right)^k}{k!}(k-\phi)}<1$. Hence, we get $\Lambda^S<\Lambda^F$.

(ii) From Corollary 1(ii), we have $\lim_{R\to\infty}\Lambda^S=\mu=\Lambda^F$. □

OA19. Lemma 4

Lemma 4. *In a finite capacity M/M/1 queue, the expected real-time QL at time t given the initial state ϕ, $E\left[\Psi(\delta,\phi)\right]$ can be derived as $E\left[\Psi(\delta,\phi)\right]=\sum_{i=0}^{\infty}P\{N=i\}E\left[\Psi(\delta,\phi)|N=i\right]$, where*

$$P\{N=i\}=\frac{e^{-(\lambda+\mu)\delta}\left((\lambda+\mu)\delta\right)^i}{i!}\qquad \textit{for } i=0,1,\ldots. \tag{OA. 15}$$

and $E\left[\Psi(\delta,\phi)|N=i\right]$ can be derived iteratively using

$$\begin{aligned}&E\left[\Psi(\delta,\phi)|N=i+1\right]\\ &=\begin{cases}\frac{\lambda}{\lambda+\mu}E\left[\Psi(\delta,1)|N=i\right]+\frac{\mu}{\lambda+\mu}E\left[\Psi(\delta,0)|N=i\right] & \textit{if } \phi=0,\\ \frac{\lambda}{\lambda+\mu}E\left[\Psi(\delta,\phi+1)|N=i\right]+\frac{\mu}{\lambda+\mu}E\left[\Psi(\delta,\phi-1)|N=i\right] & \textit{if } 0<\phi<n,\\ \frac{\lambda}{\lambda+\mu}E\left[\Psi(\delta,n)|N=i\right]+\frac{\mu}{\lambda+\mu}E\left[\Psi(\delta,n-1)|N=i\right] & \textit{if } \phi=n.\end{cases}.\end{aligned} \tag{OA. 16}$$

Proof of Lemma 4 The transient behavior of an M/M/1 queue with a finite waiting room is governed by two stochastic processes: the Poisson arrival process with rate λ and the Poisson service process with rate μ. Thus, at any moment, the next event will be either an arrival with probability $\frac{\lambda}{\lambda+\mu}$ or a service completion with probability $\frac{\mu}{\lambda+\mu}$. However, due to the finite waiting room, the arrivals or service completions do not necessarily increase or decrease the QL by one. Let $e\in(T,t)$ denote the time of an event . If the QL right before e is zero, an arrival at time e increases the QL by one, but a service completion at time e does not change the QL, which will stay at zero. If the QL right before e is in the set $\{1,2,\ldots,n-1\}$, an arrival at time e increases the QL by one and a service completion at time e decreases the QL by one. And if the QL right before e is n, an arrival at time e does not change the QL and a service completion at time e decreases the QL by one.

Let N denote the total number of events in (T,t), including all arrivals and service completions. Clearly, N is a Poisson process and follows the distribution in (OA. 15).Let $E\left[\Psi(\delta,\phi)|N\right]$ denote the conditional expected real-time QL given there are ϕ customers at time T and N events occur in (T,t). Clearly, if no events happen during (T,t), the expected real-time QL at time t will stay the same as the latest QL update ϕ; i.e., $E\left[\Psi(\delta,\phi)|N=0\right]=\phi$. Furthermore, $E\left[\Psi(\delta,\phi)|N=i+1\right]$ can be calculated iteratively from $E\left[\Psi(\delta,\phi)|N=i\right]$ for $i\geq 0$ using (OA. 16). At last, using the total probability theorem, we can derive the expected real-time QL at time t as in Lemma 4. □

OA20. Renewal Theory in Semi-Markov Processes

In this section, we consider a *semi-Markov process* with states 1, 2, ..., n. Each time the process enters state i, it earns a reward r_i and remains there for a random amount of time with mean t_i before the next transition. Let π_i denote the steady state probability distribution of this semi-Markov process. Then, using a similar proof as the one for (7.24) in Ross (2006) we can obtain the average reward rate earned in this semi-Markov process. For the purpose of completeness, we include it as a Lemma.

LEMMA 5. *The average reward rate earned in this semi-Markov process is* $\frac{\sum_{i=1}^{M}\pi_i r_i}{\sum_{i=1}^{M}\pi_i t_i}$.

Proof of Lemma 5 Let $N_i\left(m\right)$ denote the number of visits to state i in the first m transitions, X_i^j denote the reward earned during the j^{th} visit to state i, and T_i^j denote the amount of time during the j^{th} visit to state i. Then, the total reward earned in state i during the first m transitions is $\sum_{j=1}^{N_i(m)}X_i^j$, and the amount of time during the first m transitions that the process is in state i is $\sum_{j=1}^{N_i(m)}T_i^j$.

Furthermore, we have the total reward rate earned during the first m transitions is

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i(m)}X_i^j}{\sum_{i=1}^{n}\sum_{j=1}^{N_i(m)}T_i^j}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i(m)}\frac{X_i^j}{N_i(m)}\frac{N_i(m)}{m}}{\sum_{i=1}^{n}\sum_{j=1}^{N_i(m)}\frac{T_i^j}{N_i(m)}\frac{N_i(m)}{m}}. \tag{OA. 17}$$

Since the T_i^j are independent and identically distributed and have mean t_i, from the strong law of large numbers, we have

$$\lim_{m\to\infty}\sum_{j=1}^{N_i(m)}\frac{T_i^j}{N_i\left(m\right)}=t_i. \tag{OA. 18}$$

Similarly, we have

$$\lim_{m\to\infty}\sum_{j=1}^{N_i(m)}\frac{X_i^j}{N_i\left(m\right)}=r_i. \tag{OA. 19}$$

Next, to derive $\lim_{m\to\infty}\frac{N_i(m)}{m}$ we temporarily consider the model under the assumption that each transition takes one unit of time. Then $\frac{N_i(m)}{m}$ is the rate at which visits to state i occur, which, as such visits can be viewed as renewals, converges to $\left(E\left[\text{number of transitions between visits}\right]\right)^{-1}$ by Proposition 7.1 in Ross (2006). But by Markov-chain theory, this must equal π_i. As $\lim_{m\to\infty}\frac{N_i(m)}{m}$ is clearly unaffected by the actual times between transitions, we have

$$\lim_{m\to\infty}\frac{N_i\left(m\right)}{m}=\pi_i. \tag{OA. 20}$$

At last, by substuting (OA. 18), (OA. 19), and (OA. 20) into (OA. 17), we have

$$\lim_{m\to\infty}\frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i(m)}X_i^j}{\sum_{i=1}^{n}\sum_{j=1}^{N_i(m)}T_i^j}=\lim_{m\to\infty}\frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i(m)}\frac{X_i^j}{N_i(m)}\frac{N_i(m)}{m}}{\sum_{i=1}^{n}\sum_{j=1}^{N_i(m)}\frac{T_i^j}{N_i(m)}\frac{N_i(m)}{m}}=\frac{\sum_{i=1}^{n}\pi_i r_i}{\sum_{i=1}^{n}\pi_i t_i}. \quad \square$$

OA21. Useful Equations

Some relevant derivations with respect to τ:

$$\frac{\partial \sum_{j=0}^{i-1} \frac{e^{-\mu\tau_i}(\mu\tau_i)^j}{j!}(i-j)}{\partial \tau_i} = -\mu \sum_{j=0}^{i-1} \frac{e^{-\mu\tau_i}(\mu\tau_i)^j}{j!} \tag{OA. 21}$$

$$\frac{\partial \sum_{j=0}^{i-1} \frac{e^{-\mu\tau_i}(\mu\tau_i)^j}{j!}\left((\rho+1)^i-(\rho+1)^j\right)}{\partial \tau_i} = -\mu\rho \sum_{j=0}^{i-1} \frac{e^{-\mu\tau_i}(\mu\tau_i)^j}{j!}(\rho+1)^j \tag{OA. 22}$$

$$\frac{\partial \sum_{j=0}^{i-1} \frac{e^{-\mu\tau_i}(\mu\tau_i)^j}{j!}}{\partial \tau_i} = -\mu \frac{e^{-\mu\tau_i}(\mu\tau_i)^{i-1}}{(i-1)!} \tag{OA. 23}$$

Further, some relevant derivations with respect to ρ:

$$\frac{\partial \sum_{j=0}^{n-1} \frac{e^{-\mu\tau}(\mu\tau)^j}{j!}(\rho^n-\rho^j)}{\partial \rho} = n\rho^{n-1} \sum_{j=0}^{n-1} \frac{e^{-\mu\tau}(\mu\tau)^j}{j!} - \mu\tau \sum_{j=0}^{n-2} \frac{e^{-\mu\tau}(\mu\tau)^j}{j!}\rho^j \tag{OA. 24}$$

$$\frac{\partial \sum_{j=0}^{n-1} \frac{e^{-\mu\tau}(\mu\tau)^j}{j!}\rho^j}{\partial \rho} = \mu\tau \sum_{j=0}^{n-2} \frac{e^{-\mu\tau}(\mu\tau)^j}{j!}\rho^j \tag{OA. 25}$$

$$\frac{\partial \sum_{j=0}^{n-1} \frac{e^{-\mu\tau}(\mu\tau)^j}{j!}\rho^n}{\partial \rho} = n\rho^{n-1} \sum_{j=0}^{n-1} \frac{e^{-\mu\tau}(\mu\tau)^j}{j!} \tag{OA. 26}$$