Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

On the Downs-Thomson Paradox in a Self-Financing Two-Tier Queuing System

Pengfei Guo, Robin Lindsey, Zhe George Zhang

To cite this article:
Pengfei Guo, Robin Lindsey, Zhe George Zhang (2014) On the Downs-Thomson Paradox in a Self-Financing Two-Tier Queuing System. Manufacturing & Service Operations Management 16(2):315-322. http://dx.doi.org/10.1287/msom.2014.0476

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 16, No. 2, Spring 2014, pp. 315–322
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2014.0476

On the Downs–Thomson Paradox in a Self-Financing Two-Tier Queuing System

Pengfei Guo
Faculty of Business, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong,
pengfei.guo@polyu.edu.hk

Robin Lindsey
Sauder School of Business, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada,
robin.lindsey@sauder.ubc.ca

Zhe George Zhang
Department of Decision Sciences, Western Washington University, Bellingham, Washington 98225; and
Beedie School of Business, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada,
george.zhang@wwu.edu

We model a two-tier queuing system with free and toll service options as two parallel $M/M/1$ servers. We solve for the welfare-maximizing toll service capacity and toll subject to the constraint that the toll service cover its costs. If the free and toll services are both used in equilibrium, a larger free-service capacity implies longer expected waiting time for the free service and lower welfare: an analogue to the Downs–Thomson paradox in transportation economics. The paradox is caused by the presence of scale economies in the toll service combined with the requirement that it be self-financing.

Keywords: queuing system; two-tier service system; equilibrium arrival rates; pricing and capacity decisions; Downs–Thomson paradox
History: Received: July 2, 2013; accepted: December 5, 2013. Published online in *Articles in Advance* March 28, 2014.

1. Introduction

Long waiting times and crowding at the point of service are common for public services such as healthcare, immigration service, and border-crossing inspection. Congestion is caused by a combination of limited capacity and lack of demand control. Both problems can be alleviated by charging user fees for service. Fees provide a revenue stream that can be used to expand capacity and upgrade services to increase throughput. Fees also serve to ration demand, and fee levels can be set to cover the direct cost of service rendered as well as to internalize the waiting time and crowding externalities that users impose on each other.

Thanks to their dual roles of revenue generation and demand management, user fees are now commonly used for many traditional public services. User fees provide an attractive alternative to income taxes, employment taxes, and other distortionary revenue sources for funding services. They are also consistent with the user-pay principle that is generally accepted in most privately operated sectors of the economy. Nevertheless, exclusive reliance on unregulated user fees may not be possible. Free-service options may be required by law. Concerns about excessive market power arise if paid services are supplied by private-sector firms. Exclusive reliance on user-pay systems may also be opposed on equity grounds. For these reasons, a mixed service delivery model with a free-service channel operating in parallel with a fee-for-service or toll channel is often considered an attractive alternative to a wholly free system or an all-pay system. When customers choose to use both channels, a two-tier service system results, in which the free channel is more congested than the toll channel.

Two-tier or multiple-tier service systems are commonplace for hotels, airlines, postal mail and package delivery companies, freight transport, telecommunications, and computer services. Two-tier systems are also being introduced for public and not-for-profit services such as roads, airport security, immigration services, and healthcare. For example, on high-occupancy toll (HOT) lane facilities in the United States, drivers can choose between paying a toll to use the HOT lanes and taking the toll-free but generally more congested lanes that run in parallel.

Unless a toll service is subsidized, it must at least cover its costs. Two questions then arise. First, is the self-financing constraint consistent with efficient pricing of service? Second, is a self-financing toll service viable when customers can obtain free service from a competing facility? Regarding the first question, Mohring and Harwitz (1962) showed that, under assumptions

consistent with constant scale economies, the revenues from optimal congestion tolls just cover the capacity costs of a facility built to optimal size.

Mohring and Harwitz's (1962) result might suggest that the answer to the first question is yes. But for two reasons, the self-financing result does not apply to a two-tier queuing system. First, if customer arrival or service times are random, undersaturated systems have scale economies: average waiting time falls when the mean arrival rate and service capacity increase by the same proportion. Long-run marginal costs are then less than average costs, and efficient (i.e., marginal-cost) pricing results in a deficit, as Dewan and Mendelson (1990) were the first to note. Second, the cost recovery theorem applies to a single, isolated facility. In a two-tier service system, congestion in the free service is not priced, and raising the fee for toll service exacerbates excessive usage of the free service. Second-best pricing therefore calls for setting the fee below the marginal cost of usage in the toll service.

Imposing a self-financing constraint on the toll service therefore introduces a second distortion on top of the failure to price congestion in the free service. In this paper, we solve for the optimal capacity and fee of the toll service, characterize the solution's properties, and examine the viability of the toll service under the self-financing constraint. We assume that capacity of the free service is fixed, and its costs are sunk so that no infrastructure costs are incurred to operate it. By contrast, the toll service capacity has to be built. We assume that customers are identical. They all incur the same cost per unit of time waiting in a queue, and they all demand one unit of service regardless of the price. Queues are unobservable. Hence, if a toll service exists, customers choose between it and the free service based on the toll service fee and the steady-state expected waiting times for each service. This assumption is reasonable for many services. For example, applicants to the U.S. immigration service can obtain information on expected response times for regular and premium processing, but not on the number of applications currently waiting in each queue. Healthcare plans are another example: customers choosing private service buy private medical insurance, whereas those choosing public service rely on the public healthcare plan, which is free or inexpensive. Since customers choose between healthcare plans once and for all, or at least for extended periods, their decision is mainly based on long-run statistics on expected waiting times and fees.

Given these assumptions, three types of systems are possible: a one-tier free service, a one-tier toll service, and a two-tier service. We identify the conditions under which each system is optimal. We then derive the comparative statics properties of a two-tier service. The most interesting property is that welfare is a decreasing function of free-service capacity. The reason is that a larger capacity attracts customers away from the toll service, reduces toll revenue, and forces toll capacity to shrink. Scale economies are lost, the toll has to increase to continue covering costs, and in the new equilibrium average waiting time is higher for both services. This result is analogous to the Downs–Thomson paradox in transportation.

The remainder of this paper is organized as follows. Section 2 reviews the literature. Section 3 describes the model. Section 4 derives optimal capacity and pricing decisions for the toll service and identifies the Downs–Thomson paradox. Section 5 concludes. All proofs are relegated to the appendix.

2. Related Literature

This paper is related to several branches of literature on queuing and congestion. Naor (1969) was the first to consider user fees or tolls to regulate the arrival stream of customers to a single service. Many variants of his model have since been studied in the operations research and economics literatures (see the survey by Hassin and Haviv 2003). Dewan and Mendelson (1990) were the first to study joint optimization of capacity and pricing for a service facility and demonstrate that welfare maximization results in a budget deficit for the service facility. Stidham (1992) examined the stability of equilibria, and Stidham (2009) provides a comprehensive review of pricing and capacity decisions in queuing systems.

The two-tier service system considered here is similar to the so-called Paris Metro Pricing (PMP) system in which identical subway cars are divided into first and second classes. A higher fare is assigned to first class, and riders who opt to travel first class enjoy less congested service. Printezis et al. (2009) study PMP in a service system with two $M/M/1$ queues and two types of customers with different sensitivities to delay. The goal of the operator is to maximize profits for the overall system. By contrast, our study concerns welfare maximization subject to a self-financing requirement for the toll service.

A final branch of literature examines the Downs–Thomson paradox, which is attributed to Downs (1962) and Thomson (1977); see Arnott and Small (1994) for an insightful review. The paradox occurs when expanding a road draws travelers away from public transit, which reduces transit scale economies and requires a fare increase to maintain cost recovery. In the new equilibrium, public transit users pay a higher fare for lower-quality service, and the road is also more congested so that all travelers are worse off. Calvert (1997) was the first to show that a similar paradox can occur in queuing systems. Afimeimounga et al. (2005) analyze the properties of Calvert's (1997) model in detail and examine the stability and uniqueness of

equilibrium. The model features two parallel routes or queues. One service operates as a $\cdot/M/1$ queue, and the other is a $\cdot/M(N)/\infty$ batch-service infinite server queue (e.g., an airport bus shuttle with an ample fleet of buses) that exhibits scale economies. Afimeimounga et al. (2005) show that an increase in the $\cdot/M/1$ queue capacity can increase total expected delay in the system. In a later study, Afimeimounga et al. (2010) generalize the model to a $\cdot/G(N)/\infty$ batch service and show that providing information to enable state-dependent routing between servers mitigates the Downs–Thomson paradox.

The general model used in these three studies differs from ours in three main respects. First, their model does not consider system optimization. Second, it does not feature user fees or a self-financing constraint. Third, service delay in their batch server occurs because early-arriving users have to wait for other users until the batch quota is met and service can begin. The service has scale economies because higher arrival rates reduce expected delay. Capacity is assumed to be sufficient that service can always begin immediately when the quota is reached. By contrast, in our model capacity is fixed in the short run, and scale economies exist because arrival and service times are random. Capacity is adjusted in response to changes in demand to reoptimize performance of the system subject to the self-financing constraint.

Two further studies have identified the Downs–Thomson paradox in other types of queuing systems. Ziedins (2007) considers a system of parallel, finite, tandem queues with two stages. Customers incur a loss if they fail to receive service, but waiting time is not costly. Ziedins (2007) shows that with state-dependent routing an increase in the service rate at the second stage can raise total expected costs. However, no such effect arises with probabilistic routing (i.e., when queue lengths are unobserved) as is assumed in our model. Chen et al. (2012) study two parallel batch-service queues and compare user equilibrium policies for probabilistic routing and state-dependent setting. They show numerically that the Downs–Thomson paradox can occur in this system.

3. The Model

Customers arrive for service according to a Poisson process with exogenous rate λ. All customers require service and do not balk or renege regardless of queuing times. They split into two flows that enter the free (f) and toll or charge (c) services at endogenous rates λ_f and λ_c, respectively, where $\lambda_f + \lambda_c = \lambda$, and λ_f and λ_c are to be determined. Service times for each service are independently and identically distributed exponential random variables with service rates given by the service capacities, $\mu_i, i = f, c$. Capacity μ_f is fixed, with $\mu_f > \lambda$ so that queuing time remains finite in the single-tier free system if no toll service is built. Customers cannot observe queue length for either service. Expected waiting time, including service time, for service i is $W(\mu_i, \lambda_i) = 1/(\mu_i - \lambda_i)$, $i = f, c$. All customers have the same unit cost of waiting time, θ, so that expected waiting time cost for service i is $\theta W(\mu_i, \lambda_i)$, $i = f, c$.

Given a charge p to use the toll service, a joining customer chooses the toll service if and only if $p + \theta W(\mu_c, \lambda_c) < \theta W(\mu_f, \lambda_f)$. The toll service earns revenue at rate $\lambda_c p$. Revenue is treated as deterministic because steady-state conditions are assumed to prevail, and fluctuations in revenue due to random arrival times and service times even out in the long run. The cost of the toll service is assumed to depend on capacity but not on the number of customers served. Given an annualized unit capacity cost of c for the toll service, the total cost is $c\mu_c$. Free-service capacity is already established, and its cost is sunk. Variable costs are assumed to be the same for the two services and are normalized without loss of generality to zero.

4. Optimal Toll Service Decisions

As a first step in deriving optimal toll service policies, we establish an upper bound on the toll so that the optimum is well defined. If $p > \theta W(\mu_f, \lambda)$, all customers prefer the free service. Such a toll can be ruled out because it yields the same outcome as $p = \theta W(\mu_f, \lambda)$. Hence, if a toll service is built, without loss of generality, $p \le \theta W(\mu_f, \lambda)$. If both services are used, the expected costs of using them are equal:

$$p + \theta W(\mu_c, \lambda_c) = \theta W(\mu_f, \lambda - \lambda_c). \tag{1}$$

The equilibrium arrival rate λ_c^e given $(p, \lambda, \mu_c, \mu_f)$ solves (1). The equilibrium is unique because the left-hand side of (1) is increasing in λ_c while the right-hand side is decreasing in λ_c. It can be shown that the equilibrium is also locally stable in the sense that a small and temporary deviation in λ_c away from λ_c^e will induce adjustments in usage that restore the equilibrium. This contrasts with the model in Afimeimounga et al. (2005), where customers' choice equilibrium solves a quadratic equation and can have one, two, or three equilibria with some equilibria that are unstable. The reason for this difference is that expected waiting time in the $\cdot/M(N)/\infty$ batch-service queue of their model decreases, rather than increases, with the arrival rate of customers to the service.

The planner's objective is to maximize welfare subject to the constraint that the toll service earn nonnegative profits. Profits of the toll service are

$$\begin{aligned}\Pi(\mu_c, \lambda_c) &= p\lambda_c - c\mu_c \\ &= \theta[W(\mu_f, \lambda - \lambda_c) - W(\mu_c, \lambda_c)]\lambda_c - c\mu_c, \end{aligned} \tag{2}$$

where the second inequality follows from (1). Given fixed demand, welfare is maximized by minimizing expected total costs. Expected total cost for the free service is $TC_f = \theta\lambda_f W(\mu_f, \lambda_f)$, and expected total cost for the toll service is $TC_c = \theta\lambda_c W(\mu_c, \lambda_c) + c\mu_c$. Average social costs for the two services are $AC_f = TC_f/\lambda_f = \theta W(\mu_f, \lambda_f)$ and $AC_c = TC_c/\lambda_c = \theta W(\mu_c, \lambda_c) + c\mu_c/\lambda_c$. By (2), the self-financing constraint can be written as $AC_f \geq AC_c$. The optimal choice (μ_c, λ_c) for the toll service therefore solves the following optimization problem:

$$\begin{aligned} &\min_{\mu_c, \lambda_c} \ \{TC = TC_f + TC_c\} \\ &\text{subject to} \ \ AC_f \geq AC_c. \end{aligned} \tag{3}$$

A difficulty in solving (3) is that the objective function is not jointly convex in (μ_c, λ_c). This problem was first identified by Stidham (1992), who showed that even for a simple $M/M/1$ system, the joint pricing and capacity decision problem has multiple local optima. The decision problem here is further complicated by the nonnegative profit constraint. Nevertheless, the solution can be characterized by solving for μ_c and λ_c sequentially. Optimal capacity conditional on usage is determined by the following lemma.

Lemma 1. *Given λ_c, the total cost function TC is convex, and the profit function Π is concave in μ_c. Optimal capacity, which maximizes profit as well as minimizing TC, is*

$$\mu_c(\lambda_c) = \lambda_c + \sqrt{\theta\lambda_c/c}. \tag{4}$$

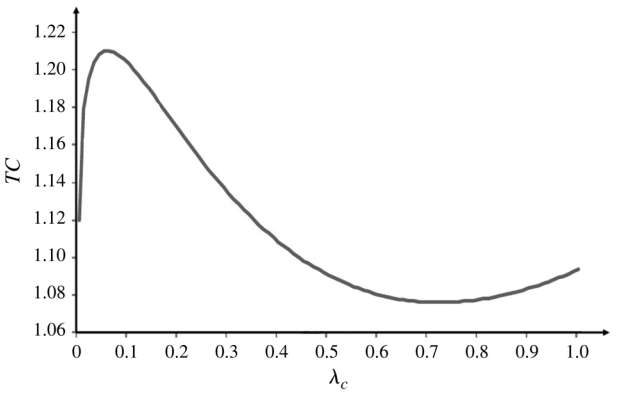
Equation (4) is derived by straightforward calculation. At the optimum, the marginal cost of extra capacity is just offset by the marginal benefit to customers from shorter delays. The solution therefore minimizes the sum of capacity and waiting time costs borne respectively by the operator and customers. The solution also maximizes profit because, for a given customer inflow λ_c, the operator can more than recoup the cost of a capacity expansion by charging a higher price if the expansion reduces customers' waiting time costs by more than the investment cost. The operator therefore maximizes profits by expanding capacity up to the point where the incremental cost exactly balances the waiting time cost reduction.

Substituting (4) into the formula for AC_c yields

$$AC_c(\lambda_c) = 2\sqrt{\frac{\theta c}{\lambda_c}} + c.$$

The average cost AC_c is decreasing in λ_c so that the facility has scale economies. An increase in the arrival rate to the toll service provides revenue to expand capacity and reduce the average cost. To show that the Downs–Thomson paradox arises, we need to determine the social planner's optimal choice of λ_c^*.

Figure 1 **Variation of Total Costs with Toll Service Usage, λ_c, Given $\lambda = 1$, $\mu_f = 1.9$, $c = 0.2$, and $\theta = 1$**



Using (4), TC_c and AC_c can be expressed as functions of λ_c alone: $TC_c = c\lambda_c + 2\sqrt{\theta c\lambda_c}$ and $AC_c = c + 2\sqrt{\theta c/\lambda_c}$. The optimization problem (3) can therefore be reduced to a function of λ_c. The marginal cost of an additional toll service user is $MC_c = dTC_c/d\lambda_c = c + \sqrt{\theta c/\lambda_c}$, which is decreasing in λ_c. In the limit $\lambda_c \to \infty$, it approaches c, and in the limit $\lambda_c \to 0$, it approaches infinity. For the free service, the marginal cost is $MC_f = dTC_f/d\lambda_f = \theta\mu_f/(\mu_f - \lambda_f)^2$, which is increasing in λ_f because capacity is fixed. The dissimilar cost properties of the two services underlie the irregular shape of the objective function (3) and the difficulty of identifying an optimal solution. To illustrate, consider parameter values $\lambda = 1$, $\mu_f = 1.9$, $c = 0.2$, and $\theta = 1$. As shown in Figure 1, the total cost curve has a local maximum at $\lambda_c \cong 0.06$ and a local minimum at $\lambda_c \cong 0.71$. The local minimum is also a global minimum so that a two-tier system is optimal for this example.

The following lemma describes some properties of the optimal solution.

Lemma 2. *With toll service capacity set according to Equation* (4):

(a) *For free service, the marginal cost exceeds the average cost (i.e., $MC_f > AC_f$); for toll service, the marginal cost is below the average cost (i.e., $MC_c < AC_c$).*

(b) *Total cost is strictly decreasing in λ_c on the domain where the self-financing constraint is satisfied; that is, when λ_c satisfies $AC_f \geq AC_c$, $dTC/d\lambda_c < 0$.*

(c) *Optimal usage of the toll service, λ_c^*, is the largest usage rate for which the self-financing constraint can be satisfied (i.e., $AC_f = AC_c$).*

Lemma 2 establishes that operating a toll service is desirable if it can cover its costs. Given part (c) of Lemma 2, λ_c^* can be solved using the condition $AC_f = AC_c$, which can be written as

$$\frac{\theta}{\mu_f - \lambda + \lambda_c} = c + 2\sqrt{\frac{\theta c}{\lambda_c}}. \tag{5}$$

To characterize the roots of Equation (5), define

$$\lambda_c^o \equiv \frac{\tau}{1-\tau}(\mu_f - \lambda),$$

where

$$\tau \equiv \sqrt[3]{\sqrt{\tfrac{1}{4}+\tfrac{8}{27}}+\tfrac{1}{2}} - \sqrt[3]{\sqrt{\tfrac{1}{4}+\tfrac{8}{27}}-\tfrac{1}{2}} \cong 0.4534.$$

The solution for λ_c^* is described in the following proposition.

Proposition 1. *Equation* (5) *has at most two positive roots for* λ_c *in* $(0, \lambda]$.

(a) *If* $c/\theta \le (\sqrt{1/\lambda + 1/\mu_f} - \sqrt{1/\lambda})^2$, *the toll service can be self-financing at the boundary solution* $\lambda_c^* = \lambda$, *and this is the optimal solution. All customers use the toll service, and the public service is not used.*

(b) *If* $(\sqrt{1/\lambda + 1/\mu_f} - \sqrt{1/\lambda})^2 < c/\theta < (\tau^3 - \tau^4)/(\mu_f - \lambda)$ *and* $\tau\mu_f \le \lambda$, *there exist two roots* λ_c^1 *and* λ_c^2, *satisfying* $0 < \lambda_c^1 < \lambda_c^o < \lambda_c^2 \le \lambda$. *The optimum is the interior solution* $\lambda_c^* = \lambda_c^2$ *in which both free and toll services are used.*

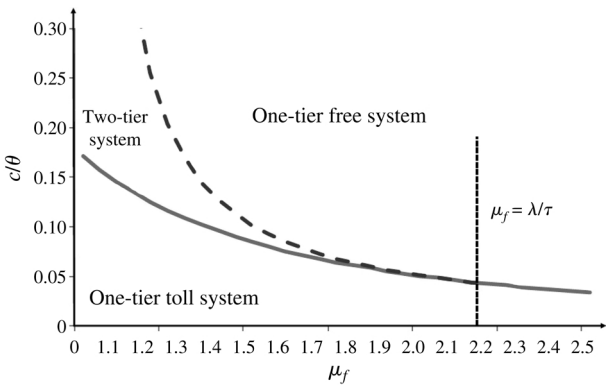
(c) *If* $c/\theta = (\tau^3 - \tau^4)/(\mu_f - \lambda)$ *and* $\tau\mu_f \le \lambda$, *Equation* (5) *has a unique root* λ_c^o. *At the optimum,* $\lambda_c^* = \lambda_c^o$, *and as in case* (*b*), *a two-tier system is operated.*

(d) *Otherwise, Equation* (5) *has no real root. The optimum is the boundary solution* $\lambda_c^* = 0$ *in which no toll service is built and all customers use the free service.*

Proposition 1 identifies parameter conditions for which each market structure is optimal: a one-tier free system, a one-tier toll system, and a two-tier system. The defining criteria are the level of demand, λ, the capacity of the free service, μ_f, and the relative costs of capacity and queuing delay, c/θ. An example with $\lambda = 1$ is shown in Figure 2.

If the free service has enough capacity to satisfy $\tau\mu_f > \lambda$, a two-tier system is never optimal. If c/θ is small, a one-tier toll system is preferred, and otherwise a one-tier free system. A one-tier system is superior because of the facility dominance property, familiar in the queuing literature (Stidham 2009), that derives from the scale economies of queuing systems. If capacity is costly, it is better not to build any toll service and instead rely exclusively on the free service that is relatively uncongested when its capacity is large. Contrarily, if capacity is inexpensive, it is better to build a large toll service. The free service should then not be used because it undermines scale economies of the toll service.

Figure 2 **Optimal Market Structure with $\lambda = 1$**



In the last case it might seem reasonable to allow the toll service to take over the unused capacity of the free service. However, as discussed in the introduction, availability of a free service may be required for legal, equity, or other reasons. Moreover, if free and paying customers were served in the same facility, it would be necessary to establish a priority rationing scheme, with free-channel customers getting bumped to the end of the queue when new paying customers arrive. Such a scheme would expose free customers to additional uncertainty about waiting time.

If a two-tier service is active, the usage rate of the toll service depends on all parameter values. As just noted, it decreases with the ratio c/θ. It also increases with the arrival rate of customers, λ, and it decreases with μ_f. When μ_f is large enough, it is not possible to build a toll service that covers its costs. This is consistent with the observation in the healthcare market that for-profit companies sometimes buy public hospitals to create a profitable market segment (Schmid et al. 2010).

The comparative statics properties of the two-tier service equilibrium, when it exists, are described in the following proposition.

Proposition 2. *For a self-financing two-tier system,* λ_c^* *and* μ_c^* *are decreasing in* c/θ *and in* μ_f, *and* p^* *is increasing in* c *and in* μ_f.

An increase in free-service capacity has opposing effects on total system costs. On one hand, it reduces queuing time in the free service for any given level of free-service usage, λ_f. On the other hand, it attracts customers from the toll service so that λ_c decreases, which undermines scale economies in the toll service. The second effect dominates, which follows immediately from the fact that the right-hand side of Equation (5) is a decreasing function of λ_c. This result is formalized in the following:

Proposition 3. *For a self-financing two-tier system, a larger free-service capacity results in longer waiting time for both free and toll service and thus reduces welfare.*

Proposition 3 is a counterpart to the Downs–Thomson paradox in transportation economics, which, as noted in the literature review, is also a characteristic of some queuing systems.

At the threshold point at which a self-financing toll service is just viable, a similar result applies:

Proposition 4. *Assume* $\tau\mu_f \le \lambda$. *If* μ_f *is increased above the threshold value* $\lambda + (\tau^3 - \tau^4)/(c/\theta)$, *welfare drops because the minimum total cost* TC^* *increases from* $\theta\lambda/(\mu_f - \lambda + \lambda_c^o)$ *to* $\theta\lambda/(\mu_f - \lambda)$.

Welfare drops because the self-financing constraint can no longer be met, and toll service is no longer viable. All customers are forced to use the free service, which becomes more heavily congested, leaving customers worse off.

As in the classical setting of the Downs–Thomson paradox, the paradox arises in the two-tier system we study because of the confluence of scale economies in the toll service and the requirement that it be self-financing. The scale economies arise because of randomness in arrival times and service times. It is worth emphasizing that scale economies are necessary, as well as sufficient, for the Downs–Thomson paradox to occur. If the cost of toll service dropped when customers switched from the toll service to the free service, an increase in free-service capacity would be beneficial rather than harmful.

5. Concluding Remarks

To illustrate the Downs–Thomson paradox in a two-tier service system, we have used a simple model with rudimentary queuing systems, a linear capacity cost function, and identical customers. Generalizing the model in any of these directions would complicate the analysis significantly. Nevertheless, as we now show, the Downs–Thomson can still emerge in each case.

First, consider a general queuing process for the toll system, where the general service time has a mean $1/\mu$. We assume only that the expected waiting time function $W(\mu_c, \lambda_c)$ is homogeneous of degree -1 in (λ_c, μ_c), i.e., $W(k\mu_c, k\lambda_c) = k^{-1}W(\mu_c, \lambda_c)$ for any positive scale parameter k. The optimal capacity for a given λ_c, denoted as $\mu_c(\lambda_c)$, solves

$$\theta\lambda_c \frac{\partial W(\mu_c, \lambda_c)}{\partial \mu_c} = -c. \tag{6}$$

Given homogeneity degree -1 of $W(\mu_c, \lambda_c)$,

$$\mu_c(\partial/\partial\mu_c)W(\mu_c, \lambda_c) + \lambda_c(\partial/\partial\lambda_c)W(\mu_c, \lambda_c) = -W(\mu_c, \lambda_c). \tag{7}$$

Now, the derivative of $AC_c(\lambda_c)$ with respect to the arrival rate λ_c, with endogenous capacity, can be expressed as

$$\begin{aligned}
\frac{d}{d\lambda_c}AC(\lambda_c) &= \theta\frac{d}{d\lambda_c}W(\mu_c(\lambda_c), \lambda_c) + \frac{d}{d\lambda_c}\left(\frac{c\mu_c(\lambda_c)}{\lambda_c}\right) \\
&= \theta\frac{\partial W(\mu_c(\lambda_c), \lambda_c)}{\partial \mu_c(\lambda_c)}\mu_c'(\lambda_c) + \theta\frac{\partial W(\mu_c(\lambda_c), \lambda_c)}{\partial \lambda_c} \\
&\quad + \frac{c\mu_c'(\lambda_c)}{\lambda_c} - \frac{c\mu_c(\lambda_c)}{\lambda_c^2} \\
&= \theta\frac{\partial W(\mu_c(\lambda_c), \lambda_c)}{\partial \mu_c(\lambda_c)}\mu_c'(\lambda_c) \\
&\quad - \theta\frac{\mu_c(\lambda_c)}{\lambda_c}\frac{\partial W(\mu_c(\lambda_c), \lambda_c)}{\partial \mu_c(\lambda_c)} \\
&\quad - \theta\frac{W(\mu_c(\lambda_c), \lambda_c)}{\lambda_c} + \frac{c\mu_c'(\lambda_c)}{\lambda_c} - \frac{c\mu_c(\lambda_c)}{\lambda_c^2} \\
&= -\frac{c\mu_c'(\lambda_c)}{\lambda_c} + \frac{c\mu_c(\lambda_c)}{\lambda_c^2} - \theta\frac{W(\mu_c(\lambda_c), \lambda_c)}{\lambda_c} \\
&\quad + \frac{c\mu_c'(\lambda_c)}{\lambda_c} - \frac{c\mu_c(\lambda_c)}{\lambda_c^2} \\
&= -\theta\frac{W(\mu_c(\lambda_c), \lambda_c)}{\lambda_c} < 0,
\end{aligned}$$

where the third equality follows from (7) and the fourth equality follows from (6). Therefore, scale economies exist for any queuing facility for which waiting time is homogeneous of degree -1 in (μ_c, λ_c). According to Lemma 2 of Dewan and Mendelson (1990), an $M/GI/1$ queue satisfies this requirement and some other queuing processes do so as well.

Second, suppose the linear capacity cost function is replaced with the affine function $K \cdot 1_{\{\mu_c > 0\}} + c\mu_c$, where $K > 0$ is a fixed capacity cost, and $1_{\{\}}$ is the indicator function. Optimal capacity given in (4) is unaffected by this change since it is determined by marginal rather than average capacity cost. The paradox still holds since the self-financing constraint is more difficult to satisfy with a positive fixed cost.

Finally, consider relaxing the assumption of homogeneous customers. It can be shown that the paradox still occurs if the reward for obtaining service is finite and differs across customers. The paradox also occurs if customers differ in their sensitivity to delay (i.e., parameter θ) if the distribution of θ is not too spread out. These results are derived in an earlier version of this paper, which is available from the authors upon request. The paradox is therefore robust to generalizations of the model in each of the three directions.

The Downs–Thomson paradox can be considered an unfortunate property of two-tier service systems because it makes them difficult to sustain. The problem can be circumvented by relaxing the strict self-financing requirement. One approach is to subsidize customers who choose the toll service. For example, many countries use public funds to subsidize patients who use private healthcare services (for a comparative analysis of subsidy schemes, see Guo et al. 2013). Another approach is to subsidize the toll service system directly.

Acknowledgments

The authors thank Stephen C. Graves, the associate editor, and two referees for many helpful comments and suggestions. Pengfei Guo acknowledges financial support from the Hong Kong Research Grants Council [General Research Fund Project PolyU543112]. Zhe George Zhang is thankful for support from the Natural Sciences and Engineering Research Council of Canada [Grant RGPIN197319].

Appendix. Proofs of Lemmas and Propositions

Proof of Lemma 2. Part (a) follows immediately from the formulas for marginal and average costs.

Part (b) follows from the chain of inequalities

$$\frac{dTC}{d\lambda_c} = MC_c - MC_f < AC_c - AC_f \le 0,$$

where the first inequality follows from part (a), and the second inequality from the self-financing constraint for the toll service.

Part (c) follows directly from monotonicity of the objective function established in part (b). □

Proof of Proposition 1. We first prove that Equation (5) has at most two positive roots in λ_c. Define $x = \sqrt{\lambda_c}$ and rewrite (5) as

$$x^3 + 2\sqrt{\frac{\theta}{c}}x^2 + (\mu_f - \lambda - \theta/c)x + 2\sqrt{\frac{\theta}{c}}(\mu_f - \lambda) = 0. \quad (8)$$

This is a cubic equation that has at most three roots. Since the left-hand side is positive at $x = -2\sqrt{\theta/c}$ and negative at $x = -3\sqrt{\theta/c}$, there exists a negative real root. This leaves at most two positive roots. There are two cases.

Case 1: $AC_f(\lambda) \ge AC_c(\lambda)$, i.e., $c/\theta \le (\sqrt{1/\lambda + 1/\mu_f} - \sqrt{1/\lambda})^2$. In this case, $AC_c(\lambda_c)$ crosses $AC_f(\lambda_c)$ exactly once from above as λ_c varies from 0 to λ. Hence, there exists a unique positive solution for λ_c between 0 and λ. Also, since $AC_f(\lambda) \ge AC_c(\lambda)$, the maximal profit is nonnegative when $\lambda_c = \lambda$. A self-financing system can be reached at λ by lowering the price. According to Lemma 2, welfare must be larger with $\lambda_c = \lambda$ than with $\lambda_c < \lambda$. Hence, the optimal arrival rate is $\lambda_c^* = \lambda$.

Case 2: $AC_f(\lambda) < AC_c(\lambda)$. In this case, there exists a critical value for c/θ such that $AC_f(\lambda)$ and $AC_c(\lambda)$ are tangent at a unique value λ_c^o, which we now solve. Since the two curves have the same slope at λ_c^o,

$$\frac{1}{\mu_f - \lambda + \lambda_c} = \sqrt[4]{\frac{c}{\theta}}\lambda_c^{-3/4}. \quad (9)$$

Substituting (9) into (5) yields

$$\lambda_c^{-3/4} = 2\sqrt[4]{\frac{c}{\theta}}\lambda_c^{-1/2} + \left(\frac{c}{\theta}\right)^{3/4}.$$

Define $y \equiv (c\lambda_c/\theta)^{1/4}$ and rewrite this equation as

$$y^3 + 2y = 1.$$

Since the left-hand side is increasing in y, there exists a unique real solution, which works out to

$$y^* = \tau = \sqrt[3]{\sqrt{\tfrac{1}{4} + \tfrac{8}{27}} + \tfrac{1}{2}} - \sqrt[3]{\sqrt{\tfrac{1}{4} + \tfrac{8}{27}} - \tfrac{1}{2}}.$$

Therefore,

$$\lambda_c^o = \tau^4\theta/c. \quad (10)$$

Substituting (10) into (9) gives

$$\frac{c}{\theta} = \frac{\tau^3 - \tau^4}{\mu_f - \lambda}, \quad (11)$$

and substituting (11) into (10) yields

$$\lambda_c^o = \frac{\tau}{1-\tau}(\mu_f - \lambda).$$

Tangent point λ_c^o is in the feasible range $[0, \lambda]$ if and only if $\tau\mu_f \le \lambda$. If $\tau\mu_f > \lambda$ and $AC_f(\lambda) < AC_c(\lambda)$, (5) has no feasible solution.

If $\tau\mu_f \le \lambda$, we need to consider several cases. If $(\sqrt{1/\lambda + 1/\mu_f} - \sqrt{1/\lambda})^2 \le c/\theta < (\tau^3 - \tau^4)/(\mu_f - \lambda)$, $AC_c(\lambda_c)$ crosses $AC_f(\lambda_c)$ exactly twice, from above in the range $\lambda_c \in [0, \lambda_c^o]$, and from below in the range $\lambda_c \in [\lambda_c^o, \lambda]$. Hence, if $c/\theta < (\tau^3 - \tau^4)/(\mu_f - \lambda)$, Equation (5) has two solutions, λ_c^1 and λ_c^2, satisfying $0 < \lambda_c^1 < \lambda_c^o < \lambda_c^2 \le \lambda$. The optimal solution is λ_c^2. If $c/\theta = (\tau^3 - \tau^4)/(\mu_f - \lambda)$, Equation (5) has a unique solution λ_c^o, and if $c/\theta > (\tau^3 - \tau^4)/(\mu_f - \lambda)$, Equation (5) has no solution.

The condition $AC_f(\lambda) < AC_c(\lambda)$ implies that the maximal profit at $\lambda_c = \lambda$ is negative. Hence, the self-financing constraint cannot be satisfied at λ. □

Proof of Proposition 2. According to Proposition 1, the optimal solution λ_c^* is the second crossing point, λ_c^2, of the curves $AC_f(\lambda_c)$ and $AC_c(\lambda_c)$ in the range $\lambda_c \in [0, \lambda]$. An increase in c/θ raises $AC_c(\lambda_c)$ without affecting $AC_f(\lambda_c)$, leading to a decrease in λ_c^2. This process continues until λ_c^2 falls to λ_c^o. If c/θ increases further, λ_c^* drops to 0. The comparative statics properties of λ_c^* with respect to μ_f are similar.

Since the optimal capacity $\mu_c^* = \lambda_c^* + \sqrt{\theta\lambda_c^*/c}$ is strictly increasing in λ_c^*, μ_c^* has the same monotonicity properties as λ_c^*. Given the self-financing constraint that $p^*\lambda_c^* = c\mu_c^*$, the optimal price $p^* = c(1 + \sqrt{\theta/(c\lambda_c^*)})$ is increasing in c and in μ_f. □

Proof of Proposition 4. From Proposition 1, we know that if $c/\theta = (\tau^3 - \tau^4)/(\mu_f - \lambda)$, an increase in μ_f will make self-financing infeasible, no toll service will be built, and λ_c will drop from a positive value λ_c^o to 0. All customers will use the public service, and their equilibrium cost rises from $AC_f(\lambda - \lambda_c^o) = \theta/(\mu_f - (\lambda - \lambda_c^o))$ to $AC_f(\lambda) = \theta/(\mu_f - \lambda)$. □

References

Afimeimounga H, Solomon W, Ziedins I (2005) The Downs-Thomson paradox: Existence, uniqueness and stability of user equilibria. *Queueing Systems* 49(3–4):321–334.

Afimeimounga H, Solomon W, Ziedins I (2010) User equilibrium for a parallel queueing system with state dependent routing. *Queueing Systems* 66(2):169–193.

Arnott R, Small KA (1994) The economics of traffic congestion. *Amer. Scientist* 82:446–455.

Calvert B (1997) The Downs-Thomson effect in a Markov process. *Probab. Engrg. Informational Sci.* 11(3):327–340.

Chen Y, Holmes M, Ziedins I (2012) Monotonicity properties of user equilibrium policies for parallel batch systems. *Queueing Systems* 70(1):81–103.

Dewan S, Mendelson H (1990) User delay costs and internal pricing for a service facility. *Management Sci.* 36(12):1502–1517.

Downs A (1962) The law of peak-hour expressway congestion. *Traffic Quart.* 16(3):393–409.

Guo P, Lindsey R, Qian Q (2013) Efficiency of subsidy schemes in reducing waiting times for public health-care services. Working paper, Sauder School of Business, University of British Columbia, Vancouver.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queuing Systems* (Kluwer, Boston).

Mohring H, Harwitz M (1962) *Highway Benefits: An Analytical Framework* (Northwestern University Press, Evanston, IL).

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Printezis A, Burnetas A, Mohan G (2009) Pricing and capacity allocation under asymmetric information using Paris Metro pricing. *Internat. J. Oper. Res.* 5(3):265–279.

Schmid A, Cacace M, Gotze R, Rothgang H (2010) Explaining health care system change: Problem pressure and the emergence of "hybrid" health care system. *J. Health Politics, Polices Law* 35(4):455–486.

Stidham S (1992) Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Sci.* 38(8): 1121–1139.

Stidham S (2009) *Optimal Design of Queueing Systems* (CRC Press, London).

Thomson JM (1977) *Great Cities and Their Traffic*, Peregrine ed. (Gollancz, London).

Ziedins I (2007) A paradox in a queueing network with state-dependent routing and loss. *J. Appl. Math. Decision Sci.* 2007(Article ID 68280). http://dx.doi.org/10.1155/2007/68280.