CONGESTION TOLLING
WITH
AGGLOMERATION ECONOMIES

Richard Arnott*

January 30, 2006

JEL codes:

Keywords: congestion, toll, agglomeration, externalities

Preliminary draft: Please do not cite or quote without the permission of the author.

This paper was prepared for the Conference in Honor of Kenneth A. Small to be held at the University of California, Irvine on February 3 and 4, 2006.

*Department of Economics
Boston College
Chestnut Hill, MA 02467
richard.arnott@bc.edu
617-552-3674

Congestion Tolling with Agglomeration Economies

1. Introduction

This paper explores an idea that has been in the air for some time, but has not, to my knowledge, been formally examined. Many believe that non-market interaction is a dominant, and perhaps the principal, cause of urban agglomeration. Non-market interaction, it is argued, generates positive externalities on both the consumption and production sides, which results in its being undersupplied in an otherwise efficient economy. One feature of "an otherwise efficient economy" is Pigouvian tolling of urban traffic congestion. If the congestion toll is reduced below this first-best level, there will be more travel. If more travel generates more interaction, then a small reduction in the congestion toll below its first-best level generates a first-order welfare gain with respect to non-market interaction and only a second-order welfare loss with respect to congestion. The second-best optimal toll is therefore lower than the first-best optimal toll.

What does the validity of this intuitive argument depend on? And if the argument is valid, what determines how much lower the second-best toll is than the first-best toll? Under reasonable parameter values, might this effect be so strong as to warrant excluding urban travel from congestion tolling? If current empirical knowledge is insufficient to estimate the second-best congestion toll, what additional empirical information is needed to do so?

This paper addresses these questions in the context of a sequence of models. Section 2 considers the basic monocentric city model, extended to include traffic congestion à la Solow (1972) and external economies of scale à la Henderson (1972). Section 3 extends the model to include labor-leisure choice and a fixed workday à la Parry-Bento (2001) and external economies of scale à la Henderson (1981), and discusses how the model could be extended to treat the land market and polycentricity. Section 4 provides a thorough analysis of a model with intra-day dynamics. Section 5 concludes.

2. *The Monocentric City with Traffic Congestion and External Economies in Production*

The monocentric city with traffic congestion was first explored by Solow (1972) and further developed by Kanemoto (1977) and Arnott (1979). The treatment of traffic congestion assumes that traffic is uniform over the course of the day, with each resident taking a single return trip from home to the point CBD (central business district), that traffic congestion uses up physical resources rather than time, and that the cost to a driver of traversing a section of road between distances y and y + dy away from the CBD is c(Q(y),w(y)), where Q(y) is the number of travelers on the road at y per day, which equals the number of residents living beyond y, and w(y) is the capacity of the road at y.

The treatment of external economies of scale in production follows Henderson (1974). He assumed that the aggregate urban production function is F(N), with F', F" > 0, where N is the number of city residents. In the treatment of decentralization, the scale economies are external to the individual, atomistic firm. Each of the identical firms views itself as facing a horizontal marginal cost curve, but firms' marginal cost curves shift down as the number of city residents increase. In the urban economics context, these Marshallian production externalities have come to be termed agglomeration economies[1].

[1] The term agglomeration economies is sometimes used in this specific sense and sometimes in a more general sense, to include all sources of increasing returns to scale, on both the consumption and production sides. Unless otherwise noted, I shall use the term in the specific sense.

There is a large and relatively recent literature on agglomeration economies. Fujita and Thisse (2002) and Duranton and Puga (2004) review the theoretical literatures. Rosenthal and Strange (2004) reviews the empirical literature. And Moretti (2004) reviews the theoretical and empirical literatures on human-capital-based agglomeration economies.

The essential feature of the model is that urban production depends on the *number* of urban residents. In a closed, urban economy, in which the number of urban residents is fixed, congestion tolling has no effect on the number of residents and hence on urban output. Thus, there is no interaction between the agglomeration and congestion externalities, so that the second-best congestion toll coincides with the first-best congestion toll[2].

The analysis for an open urban economy is considerably more complex. It is necessary to be precise about the rest of the economy and about the conceptual exercise being performed. What is the pattern of land ownership? How are new cities formed? Is migration atomistic or cooperative? Are we considering the effects of introducing congestion tolling in a city in which agglomeration externalities are uninternalized, when all other cities' congestion and agglomeration externalities are internalized, when all other cities' congestion externalities are internalized but not the agglomeration externalities, or when all other cities' congestion externalities and agglomeration economies are uninternalized? To further complicate matters, even though everyone is identical and technology is everywhere the same, the nonconvexities associated with increasing returns to scale in production may give rise to asymmetric optima and equlibria, with cities of different sizes. Since modern cities are much closer to being open than closed, the issues raised by openness are obviously important. I have however chosen to abstract from them in order to highlight some other considerations, and for the rest of the paper shall consider only a single, closed city.

---

[2] It is surprising that the interaction between unpriced congestion and uninternalized agglomeration economies has been so little explored. One reason is that we are so used to thinking in terms of the monocentric model, in which this interaction is absent.

3. Two-Island City: Productivity Depends on Aggregate Work Hours

Assuming that a city's output depends only on its population is an evident simplification. It would be a good assumption if everyone works and has the same workday, since then whatever underlies the benefits from interaction would be captured by population. But employees now have more flexibility with respect to the number of hours they work per day, the number of days they work, and when they work. As well, labor force participation rates show considerable variation. The next section will present a dynamic model that solves for the pattern of employment and congestion over the day. This section considers a simpler, steady-state model in which workers decide how many hours to work per day, and productivity depends on aggregate work hours. Thus, there will be interaction between productivity, congestion, and labor-leisure choice.

The interaction between congestion and labor-leisure choice has been considered in a series of papers in the literature (Parry and Bento, 2001; Calthrop, Proost, and van Dender, 2000). Income taxation distorts the labor-leisure choice, the substitution effect encouraging leisure. If all travel is for commuting purposes, unpriced congestion distorts the labor-leisure choice in the opposite direction, encouraging labor, since individuals do not face the full social cost of their journeys to work. Thus, with distortionary income taxation, the second-best congestion toll falls short of the first-best toll, and at high rates of income taxation could be negative, if this were institutionally feasible. Here, income taxation will not be treated, but the basic idea is the same. External economies of production result in workers being paid their average product of labor, while efficiency calls for them to be paid their marginal product, which exceeds the average product. Thus, the wage rate is inefficiently low, encouraging leisure, and labor can be stimulated

by making the journey to work cheaper.

The model employed is about the simplest possible. The principal difference between the main model of this section and the closed variant of the model of the previous section is that here productivity is an increasing function not of the number of workers but of the aggregate number of work hours. Henderson (1981) was the first to employ this assumption in the context of a model of staggered work hours. Subsequent works that employ the assumption include Arnott (2002), Arnott, Rave, and Schöb (2005), Mun and Yokekawa (2000), and Yoshimura and Okumura (2001). Empirical support for the assumption is provided in Wilson (1988). Using the results of a 1975 survey for Singapore, he finds that, after controlling for measured differences between workers, the daily wage is on average almost twice as high for workers with a peak work start time as for those with an off-peak work start time. Intuition suggests that this difference is too large to be explained by intra-day productivity effects alone and that sorting of workers across work start times on the basis of unobservables must play an important role, but no empirical work has been done that attempts to distinguish between the two effects.

3.1 Model description

The economy is in steady state, and the unit of time is a day.

Residents: The city is closed, with N identical residents. Individuals have a well-behaved utility function, with utility depending on the consumption of other goods (the numéraire) c and leisure l : $u = u(c, l)$.

Congestion: Production takes place on one island, residence on another. Between the two islands is a causeway of fixed capacity, and travel time is increasing in the daily flow on the causeway, f; specifically, commuting time to and from work is $t(f) = t_0 + t_1 f^{\beta}$,

where t_0, t_1, and β are positive constants.

Production: Aggregate daily output is $F(H) = kH^{\alpha}$, where H is aggregate work hours and k and α positive constants. Since there are economies to scale, $\alpha > 1$, with $\alpha - 1$ being the degree of increasing returns to scale. And since these scale economies are external to the individual firm, the individual firms views itself as facing the production function Kh, with $K = kH^{\alpha-1}$, and h is the firm's work hours. Thus, the average product of labor is $kH^{\alpha-1}$ and the marginal product $\alpha kH^{\alpha-1}$. The workday is a fixed fraction of the day, L, and x is the fraction of workdays that a worker works, which is chosen by the worker[3].

3.2 Competitive equilibrium

Because the returns to scale are external to the individual competitive firm, the workday wage, w, equals the average product of labor over the workday. With a congestion toll of τ per trip and a lump-sum redistribution of toll revenue of T, a resident's daily budget constraint is

$$(w - \tau)x + T - c = 0, \qquad (1)$$

and his daily time constraint is

$$1 - (L + t)x - l = 0. \qquad (2)$$

After substituting these constraints into the utility function, the individual's maximization

[3] If one were to assume that a worker's productivity is independent of the length of the workday, and that the worker chooses the length of the workday, he would choose to work one long shift, since he would then have to make only one trip.

This treatment of the relationship between the number of trips a worker makes and the number of hours he works is due to Parry and Bento (2000).

The assumption that the length of the workday is standard and exogenous is distastefully *ad hoc*. I know of no models of the length of the workday. A structural but rather implausible assumption is that workers because discontinuously less productive after L hours of work. A more satisfactory model would take into account intra-day productivity effects and workers' scheduling constraints. If a workday different from the norm disrupted workers' scheduling, a firm that deviated from the norm would have to pay its workers a wage premium.

problem is

$$\max_{x} u((w - \tau)x + T, 1 - (L + t)x). \quad (3)$$

An interior maximum is assumed. The corresponding first-order condition is

$$(w - \tau)u_c - (L+ t)u_l = 0. \quad (4)$$

A resident views his opportunity cost of leisure as $(w - \tau)/(L + t)$. Letting AP denote the average product of a workday and PT the private time cost of a workday, this may be written as $(AP - \tau)/PT$.

3.3 Social optimum

Since daily output is $F(H) = F(NLx)$, the resource constraint for the economy is

$$F(NLx) - Nc = 0. \quad (5)$$

The economy-wide time constraint from the planner's perspective is

$$N(1 - x(L + t(Nx)) - l) = 0, \quad (6)$$

since the flow rate on the causeway is $f = Nx$. After substituting these constraints into the utility function, the planner's maximization problem is

$$\max_{x} u(F(NLx)/N, 1 - x(L + t(Nx)). \quad (7)$$

It is assumed that the global maximum is interior. The corresponding first-order condition is

$$LF'u_c - (L + t + Nxt')u_l = 0. \quad (8)$$

Letting MP denote the marginal product of a workday and ST the social time cost of a workday, the social opportunity cost of leisure may be written as MP/ST.

3.4 Competitive decentralization of the social optimum

It is assumed to be impossible to internalize the agglomeration externality by subsidizing work. The government can however impose a congestion toll τ. Since there is only one margin of choice, x, with toll revenues being distributed in lump-sum fashion, it should be possible to set the toll at such a level that the social optimum is decentralized. This is shown in Figure 1. It plots the transformation frontier and indifference curves in l-c space. Since it has been assumed that the global maximum is interior, the social optimum occurs at a point of tangency between an indifference curve and the transformation frontier, the point Ω in the Figure. The social optimum can be decentralized by having the individual maximize utility subject to a budget constraint that is tangent to the indifference curve at the social optimum.

The toll should be set to equalize the private and social opportunity cost of leisure. From (4) and (8) one obtains

$$\tau^{**} = AP\,(1 - [(PT)(MP)]\div[(ST)(AP)]). \qquad (9)$$

Using (4), this can be rewritten in a more familiar form:

$$\tau^{**} = -\,(MP - AP) + (u_l/u_c)(ST - PT); \qquad (10)$$

thus, as expected the optimal toll equals minus the agglomeration externality plus the congestion externality in terms of time multiplied by the private value of time. We refer to τ^{**} as the second-best congestion toll since a wage subsidy to internalize the agglomeration externality cannot be imposed, even though, in the model, application of the second-best toll results in attainment of the social optimum. Now, AP equals w, the wage for a workday, and, with the form of the production function assumed, the ratio of MP to AP is α. Thus, with the specific functional forms assumed substituted in, (9) can be rewritten as

$$\tau^{**}/w = 1 - \alpha(L + t_0 + t_1(f^*)^{\beta}) \div (L + t_o + (\beta+1)t_1(f^*)^{\beta}). \qquad (11)$$

The assumed congestion technology has the neat property that the ratio of the congestion externality to private congestion (the time lost by a driver due to congestion) is β.

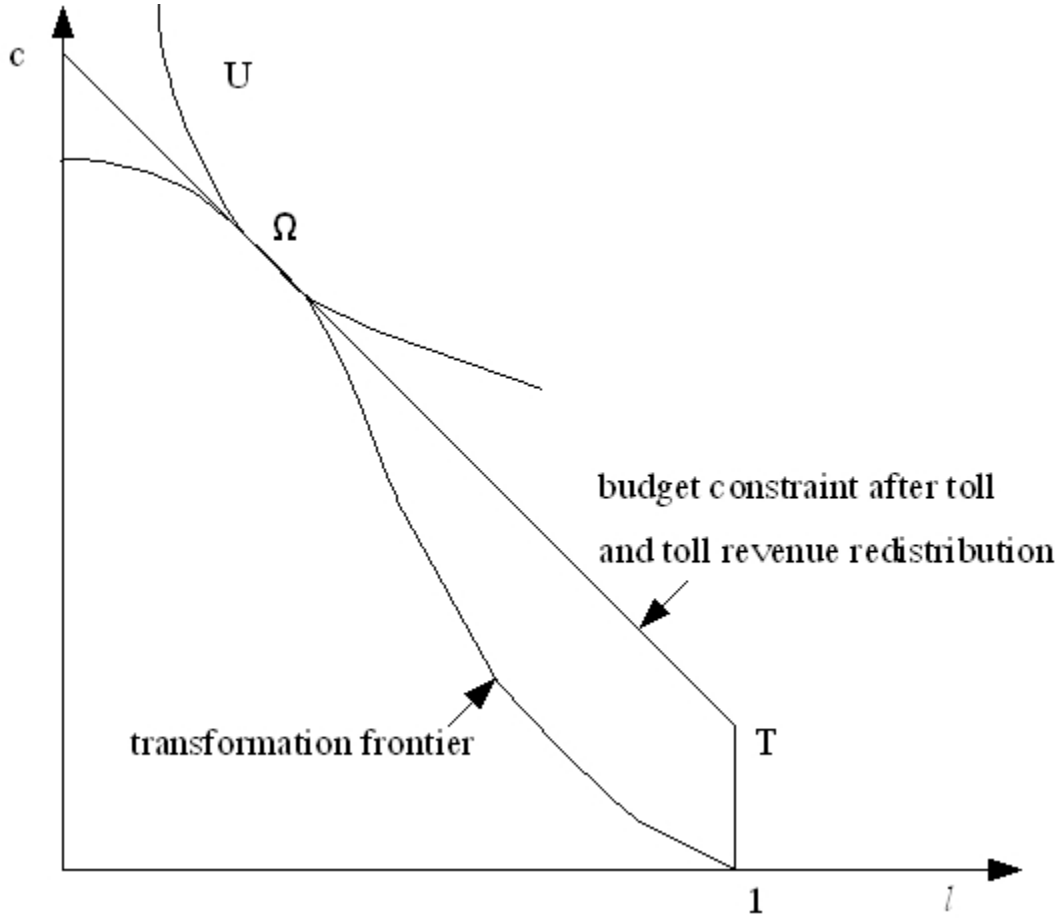

Figure 1: Decentralization of the optimum

Let us experiment with some numbers. Empirically, α is measured as one plus the elasticity of the wage rate with respect to city size. This varies across industries, but estimates of the order of 1.05 are typical. There is more disagreement concerning the magnitude[4] of β. Based on observing the relationship between flow and velocity at a point on a road, estimates of 3.0 or higher used to the norm. But more recent studies that

[4] See the discussion in Chapter 5 of Arnott, Rave, and Schöb (2005).

look at long sections of a road find estimates closer to 1.0. Both values will be considered here. Consider a typical city in which the daily commute in each direction takes 20 minutes under free-flow conditions and 30 minutes under congested conditions, and assume the workday to be 8 hours. If $\beta = 1.0$, $t^{**}/w = -0.0125$, while if $\beta = 3.0$, $t^{**}/w = 0.055$. Thus, when the ratio of the congestion externality to private congestion is 1.0 , the second-best optimal toll is negative 1/80th of the daily wage; when this ratio increases to 3.0, the second-best optimal toll is positive 1/20th of the daily wage. The magnitudes of the corresponding first-best congestion tolls, evaluated at the optimum, are 0.0375 and 0.105 of the daily wage.

The above example does not take into account that measured values of time are typically only a fraction of the social opportunity cost of leisure, as measured above. We can adjust for this in a somewhat *ad hoc* way by assuming that a fraction of travel time is treated as leisure. Suppose, for the sake of argument, that the fraction is one-half, which accords fairly well with the results of the most recent studies (Small, Winston, and Yan (2005)). Then the corresponding figures are $t^{**}/w = -0.0298$ for $\beta = 1.0$ and 0.0083 for $\beta = 3.0$. The magnitudes of the corresponding first-best congestion tolls, evaluated at the optimum, are 0.0201 for $\beta = 1.0$ and 0.0583 for $\beta = 2.0$, as a proportion of the daily wage.

The above model is very simple. The only margin is the labor-leisure tradeoff; there is only the single employment location; at the level of the economy, work and commuting are uniformly distributed over the course of the day; there is no land; the assumption of a fixed-length workday, while empirically reasonable, is *ad hoc*; and the agglomeration externalities are a black box.

3.4 Incorporating land

This section sketches how land could be incorporated into the model. Assume, as above, that there are two islands. But now to treat the distortion introduced into the land market by not imposing a first-best congestion toll, assume that the central city, which is on one island of area A_0 contains both firms and residences, and that the suburbs, which are located on the other island of area A_1, contain only residences. Central city residents incur no commuting costs. All suburban residents have to traverse the causeway connecting the central city and the suburbs, which is subject to flow congestion. Individuals demand land for residential purposes[5], while firms demand land in production. Equilibrium in the land market establishes the rents on the two islands, r_0 and r_1. The lower rent in the suburbs compensates suburbanites for the costly commute they undertake, and central city land is allocated between firms and residents such that their bid rents are equalized. Residents have a utility function of the form $u(c, l, s)$, where s is lot size, and a firm's production function is given by $[(N_o x_o + N_1 x_1)L]^{\alpha-1} f(\theta)$ per unit area of land, where θ indicates the labor-land ratio.

There are now two groups of individuals. Central city residents face the following maximization problem:

$$\max_{x,c,l,s} \; u(c, l, s) \qquad \text{s.t.} \qquad wx + T - r_o s - c = 0 \qquad (12)$$
$$1 - l - Lx = 0$$

where T is the equal lump sum redistributed from toll revenues and land rents. From the solution one obtains the demands for x, c, l, and s as functions of w, r_0, and T. The

[5] The modeling of residential land and congestion follows that in Pines and Sadka (1985), except that in their model the suburb is the "mainland", with an arbitrarily large amount of land available at an exogenous opportunity cost, rather than an island of fixed area.

analogous problem for suburban residents is

$$\max_{x,c,l,s} u(c, l, s) \quad \text{s.t.} \quad (w - \tau)s + T - r_1 s - c = 0 \qquad (13)$$
$$1 - l - (L + t)x = 0,$$

from which the demands for x, c, l, and s as functions of w, r_1, T, and τ are obtained. An individual firm chooses the land-labor ratio to employ so as to maximize profits per unit area of land:

$$\max_{\theta} [(N_0 x_0 + N_1 x_1)L]^{\alpha-1} f(\theta) - r_0 - w\theta/L; \qquad (14)$$

w is the wage paid for a workday and so w/L the wage paid for an entire day. From this maximization problem, θ can be solved for as a function of w, r_0, N_o, N_1, x_0, and x_1. There are market-clearing conditions for land on both islands and the goods market; the wage rate equals the average product of labor; the government's budget balances; and if there are residents in the central city, their utility must equal that of suburban residents.

As a result of introducing land, setting the congestion toll below the first-best level will generate three distortions, one on the labor-leisure margin, and two on land margins. The second-best toll balances the distortion on these three margins against the distortion on the labor-leisure margin deriving from the agglomeration externality. Since there is only one instrument, the congestion toll, and three margins, the social optimum is not in general attainable, and the second-best optimal toll is obtained as an exercise in the theory of the second best. Intuition suggests that the introduction of land will result in the second-best congestion toll being closer to the first-best congestion toll.

It would be interesting to extend the above model to allow for polycentricity, with the possibility of employment centers on both islands. The simplest way to do this is to

assume that a resident works at either one employment center or the other, and that the productivity at each employment center is a function of the number of workdays per day at that employment center. But this specification of agglomeration economies may miss some important effects. There have been a number of recent papers (Rosenthal and Strange (2002), Fu (forthcoming), and Henderson (xxxx)) that estimate empirically the rate at which the productivity benefits from proximity to other workers attenuates with distance. They find that the rate of attenuation varies across occupations and industries. Two modifications would need to be made to the specification to incorporate the possibility of interaction between workers at different locations (and hence at the two employment centers). The most obvious is to make the productivity at one employment center a function of the number of workdays at both that employment center and the other. Since the cost of communicating with someone by telecommunication or by computer is insensitive to distance, the interaction that underlies agglomeration presumably is face to face. Consequently, it is important to take into account the costs and time involved in face-to-face meetings. One *ad hoc* way to do this in the context of the model would be to have the individual choose the proportion of days x' to visit the other employment center, with the firm deciding on the wage function w(x').

4. Intra-day Dynamics

4.1 Introduction

The previous sections assume that the urban economy is in a steady state, with uniform traffic flow, and an equal number of workers at work at all times. In fact, while the distribution of work start times is not as concentrated as it used to be, it remains highly concentrated. Since workers traveling at peak hours experience considerably more congestion than workers traveling in off-peak hours, they must be compensated in some way for the higher congestion they experience. One form of compensation is reduced schedule delay costs; working normal hours allows them to better synchronize their non-work activities with others. Another is higher wages. If wages are indeed higher for a normal workday, it must be because an individual who works a normal work day is more productive, presumably because of expanded scope for interaction. These observations suggest that it is important to examine the intra-day dynamics of work productivity and congestion in a persuasive analysis of uninternalized agglomeration externalities[6].

Henderson (1981) is the seminal paper along these lines. He assumes that the rate at which a worker produces the composite good at a point in time is an increasing function of the number of workers then at work. Since each worker is uncompensated for the

[6] The literature has evolved in the context of studying the positive and normative aspects of flextime and staggered work hours. A central question has been: Should the government encourage the staggering of work start times? It can do this either by regulating the distribution of private firms' work start times or by staggering the work start times of government workers. Works that examine issues related to flextime and staggered work hours include Henderson (1981), Giuliano and Golob (1990), Mun and Yonekawa (2000), Yoshimura and Okumura (2001), and Arnott, Rave, and Schöb(2005), Ch.4.

benefit his presence at work provides other workers then at work, there is an positive uninternalized externality, that Yoshimura and Okumura (2001) refer to as a temporal agglomeration externality. Having specified the congestion technology, one can solve for the equilibrium distribution of work start times and the equilibrium evolution of congestion over the day. Henderson's model can be criticized in two respects. First, the interaction technology is unduly restrictive. Generally, a worker's productivity over his workday should depend on function relating the number of workers at the employment center as a function of time throughout his workday. Second, and more seriously, the congestion technology Henderson assumes is unsound. He assumes that a worker's commute time depends on the number of workers who leave home at the same time as he does[7]. This is inconsistent with the physics of traffic flow, and leads to the out-of-equilibrium possibility that a worker who leaves home later arrives at worker earlier. At first glance, the bottleneck model (Vickrey (1969), Arnott, de Palma, and Lindsey (1993)) would appear to be an attractive alternative since it is tractable and its physics sound. Unfortunately, the bottleneck's capacity determines the distribution of work start times; furthermore, if one assumes a workday of fixed length, there is no congestion during the evening rush hour[8] (Arnott (2002)). Another possibility is to incorporate non-stationary flow congestion, which entails combining the equation of continuity with some function relating travel speed and density (Lighthill and Whitman (1955)), but determining equilibrium then entails solving a partial differential equation whose

[7] The Henderson treatment of congestion does however have the virtue of tractability, and for this reason has been employed in a number of studies including Mun and Yokenama (2000), Yoshimura and Okumura (2001), Lindsey (2002), and Arnott, Rave, and Schöb (2005, Ch.4).

[8] Suppose the bottleneck's capacity is b. Then during the morning rush hour individuals arrive at the employment center at the rate b per unit time. If all of them start work as soon as they arrive, then they start work at the rate b per unit time. And if they all have the same length of workday, they leave work at the rate b per unit time. But leaving work at this rate results in no queue behind the bottleneck, and therefore no congestion.

properties are not well understood.

This paper provides a general formulation of the intra-day dynamics of both productivity and congestion due to Marvin Kraus (1998). The only restriction it imposes on either the dynamics of traffic congestion and of intra-day productivity is that time be discretized. The city is closed with population N. All individuals are *ex ante* identical, with the common utility function $u(c, l)$. The day is divided up into an arbitrarily large, but finite, number of discrete increments of equal length, indexed by i. The distribution of work start times is given by the vector **n**, where n_i is the number of individuals who start work during time interval i. It is assumed that the length of the workday is the same for all workers, and that firms are competitive. It is also assumed that there is a one-to-one mapping between a distribution of work start times and a distribution of travel times, so that one may write **t**(**n**), where t_i(**n**) is the travel time of an individual who starts work at time i, conditional on the distribution of start times **n**. **t**(**n**) is referred to simply as the travel time function. Given a distribution of work start times and a productivity-determining technology, it should be possible to determine the productivity of a worker over his entire workday as a function of the distribution of start times. Refer to **z**(**n**) as the private product function, where z_i(**n**) is the private product, over his entire workday, of an individual who starts work at time i, conditional on the distribution of work start times.

4.2 A simple variant of the model

Since the model is rather abstract, a simple variant of the model in which each worker works every day (x = 1) is presented to begin. On the assumption that individuals are paid their private products the utility of an individual who starts work at time i is then

$u(z_i(\mathbf{n}) - \tau_i + T, 1 - L - t_i(\mathbf{n}))$, where τ_i is the toll payable by a worker who starts work at time i and T is the lump sum amount received by each individual from the redistribution of toll revenue. It is therefore assumed that the toll can be conditioned on work start time but not the lump-sum that is redistributed from toll revenue. A competitive equilibrium conditional on $\boldsymbol{\tau}$ is defined to be a pair ($\mathbf{n}$, T) satisfying the following three conditions:

1. $\mathbf{n}$ is determined as the outcome of decentralized choice. Letting W denote the set of times at which workers start work, this condition can be written as

$$u(z_i(\mathbf{n}) - \tau_i + T, 1 - L - t_i(\mathbf{n})) = \underline{u} \qquad \text{for all } i \in W \tag{15}$$

$$u(z_i(\mathbf{n}) - \tau_i + T, 1 - L - t_i(\mathbf{n})) \text{ LTE } \underline{u} \qquad \text{for all } i \notin W,$$

which implies that no individual can improve his utility by shifting to a different work start time.

2. All the toll revenue must be redistributed.

$$\sum_i n_i\tau_i - NT = 0 \tag{16}$$

3. The entire population must work, and no one more.

$$\sum_i n_i - N = 0 \tag{17}$$

The first exercise to be considered will be the maximization problem of a government that can control the vector of tolls but not the distribution of work start times. The most natural specification of this problem is to maximize the common utility level with respect to $\underline{u}$, $\mathbf{n}$, $\boldsymbol{\tau}$, and T, subject to (15), (16), and (17). Since, however, the solution will be more intuitive if the Lagrange multipliers are expressed in units of money, an alternative and less intuitive specification will be analyzed, in which the government is assumed to maximize its surplus, defined as toll revenue less lump-sum transfers, subject to the utility and population constraints. The solution to this maximization problem

characterizes the set of constrained Pareto optimal allocations, where the constraints are that **n** is determined as the outcome of decentralized choice and that each individual is paid his private product.

The maximization problem in Lagrangean form is

$$\max_{\mathbf{n},\boldsymbol{\tau},T} \quad \sum_i n_i\tau_i - NT + \sum_i\lambda_i[u(z_i(\mathbf{n}) - \tau_i + T, 1 - L - t_i(\mathbf{n})) - \underline{u}] + \phi[N - \sum_i n_i] \qquad (18)$$

where λ_i is the Lagrange multiplier on (15) for time i, and ϕ that on (17). Before proceeding with the analysis, it will be useful to ponder the problem. In this economy, the individual's only margin of choice is his work start time. The government is therefore choosing tolls in order to vary the distribution of work start times, which will alter the time profile of productivity and congestion. Where I is the number of time intervals, the government has I + 1 instruments, the vector $\boldsymbol{\tau}$, and T, and I objectives, to achieve the Pareto optimal **n**. Thus, there is a degree of freedom in the choice of tolls.

The first-order conditions are:

$$n_j: \qquad \tau_j - \phi + \sum_i\lambda_i[(\partial u_i/\partial c_i)(\partial z_i/\partial n_j) - (\partial u_i/\partial l_i)(\partial t_i/\partial n_j)] = 0 \qquad (19)$$

$$\tau_i: \qquad n_i - \lambda_i(\partial u_i/\partial c_i) = 0 \qquad (20)$$

$$T: \qquad - N + \sum_i\lambda_i(\partial u_i/\partial c_i) = 0 \qquad (21)$$

Define

$$\Gamma_{ij} = - \partial z_i/\partial n_j + [(\partial u_i/\partial l_i)\div(\partial u_i/\partial c_i)]\partial t_i/\partial n_j. \qquad (22)$$

Substituting (20) and (22) into (19) yields

$$\tau_j = \phi + \sum_i n_i\Gamma_{ij}. \qquad (23)$$

Now, Γ_{ij} is the net negative externality an individual at time j imposes on an individual who arrives at work at time i; this equals the negative congestion externality minus the

positive agglomeration externality. Thus, $\sum_i n_i \Gamma_{ij}$ is the net negative externality an individual in time j imposes on all other individuals, and hence (21) states that the second-best optimal congestion toll to apply to a person at time j equals the net negative externality he imposes, plus a constant. What is the role of the constant? It can be seen from the expression for utility that the individual's decision of work start time depends on the vector with elements τ_i - T. He decides on work start time on the basis of the net transfer from the government. If all tolls are raised by \$1, the lump-sum transfer is raised by \$1 too, and the net transfer is unchanged; thus, the indeterminacy in the absolute level of the toll that was alluded to earlier.

The above specification assumed the utility function to have the form $u(c, l)$. It may be, however, that individuals have preferences over the work time *per se* for scheduling reasons. This can easily be incorporated by assuming instead that the utility function has the form $u_i(c, l)$. It is easily checked that the introduction of work start times preferences, which loom large in the literature on the bottleneck model in the form of schedule delay costs, do not alter the above result.

The stage is now set to tackle the central model of the paper.

4.2 The central model

In the model of the previous section, the effect of tolls was to redistribute individuals over work start times. It was shown that tolls can decentralize the equal-utilities Pareto optimal allocation. This is achieved when tolls are set equal to an indeterminate constant plus the net negative externality at each work start time. The constant is inessential.

What matters is that the set of tolls internalizes the difference in the externalities an individual imposes when switching from one start time to another. The optimum is attainable since there are at least as many instruments as targets.

The model developed in sections 3.1 to 3.3 looked at optimal tolls from a different perspective. Since the model is steady state, ensuring the efficient allocation of individuals over work start times does not arise. Instead, the role of tolling is to induce individuals to make an efficient labor-leisure tradeoff, or equivalently to choose the efficient proportion of days to go into work. Leisure is the only margin of choice, and the single toll is sufficient to decentralize the optimum. The optimal toll equals the net negative externality. Thus, in both models the optimum is decentralized by setting the toll(s) equal to the corresponding net negative externality(ies).

The model in this section combines the two models by extending the model of the previous section to include a labor-leisure tradeoff. Tolls now play a dual role. They reallocate individuals over work start times and affect individuals' leisure decisions. If the toll redistribution can be made work-start-time-specific, then there are 2I goals and 2I instruments, which suggests that the equal utilities optimum can be decentralized.. It is however hard to conceive of a situation in which work-start-time-specific toll redistribution is possible, but not work-start-time-specific wage subsidies. Thus, the more relevant policy exercise has the amount of toll revenue redistributed to each individual independent of his work start time. In that case, there are 2I goals but only I + 1 instruments, which raises doubts about decentralizability of the optimum.. Imposing constraints on the time variation of the toll would further restrict the number of policy instruments. When the number of policy instruments falls short of the number of policy

goals, determination of optimal policy is generally an exercise in the theory of the second

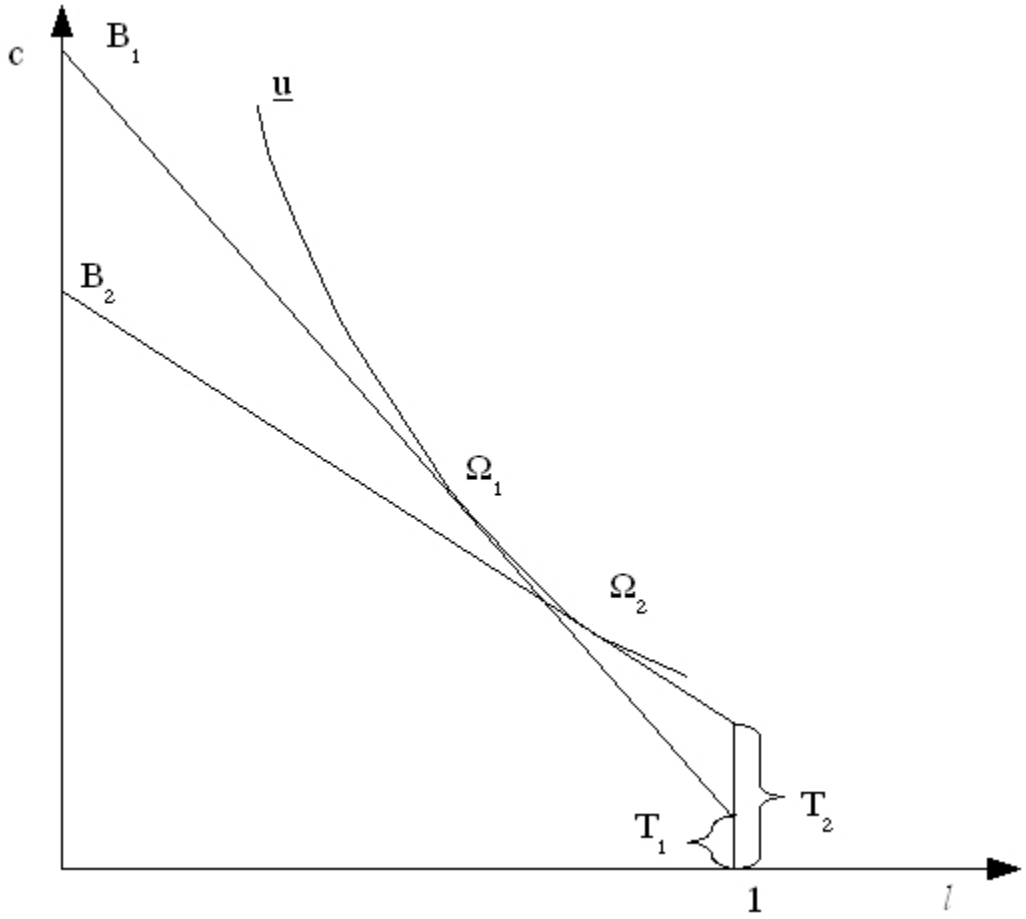

Figure 2: A second-best problem

best[9].

An intuition for why setting tolls equal to the corresponding net negative externalities and then redistributing toll revenue equally across individuals and therefore independent of work start time may not permit attainment of the full optimum can be given in terms of Figure 2, for the case in which individuals have no taste for work start time *per se.* Suppose, for the sake of argument, that there are only two work start times, 1 and 2.

[9] There is a large literature that applies the theory of the second best to problems in urban transportation. Notable contributions include Lévy-Lambert (1968), Wheaton (1978), Wilson (1988), Mayeres and Proost (1997), Arnott and Yan (2000), and Verhoef (2002).

Points Ω_1 and Ω_2 indicate the equal-utilities optimal allocations for the two work start times, with the optimal allocation of individuals across the work start times. At each work start time, the toll is set equal to the corresponding net negative externality. The optimal allocations can be decentralized by providing individuals with work start time 1 with the budget line B_1, and those with work start time 2 with the budget line B_2. But these budget lines entail different lump-sum transfers, shown as T_1 and T_2. Suppose now that the constraint is imposed that the lump-sum transfers be equalized across the work start times, with no reallocation of individuals across work start times. This causes B_1 to shift up and B_2 to shift down. Individuals with work start time 1 can now achieve a level of utility higher than $\underline{u}$, while the highest utility individuals with work start time 2 can achieve is less than $\underline{u}$. This provides an incentive for migration from work start time 2 to work start time 1. This migration will upset the optimality of the allocation. The choice of tolls would then be an exercise in the theory of the second best.

One interesting result that follows from the diagram is that, in any equilibrium in which the revenue redistributed from toll revenues is independent of work start time, the private opportunity cost of leisure, as well as the amount of leisure chosen, is independent of work start time. The diagram assumes that individuals do not have a preference *per se* for work start time, and the result requires this assumption.

It shall be shown, however, contrary to the above intuition, that the optimum can be decentralized with tolls that completely internalize the externalities and a lump-sum transfer of toll revenue that is uniform across work start times. The above argument for the non-optimality of Pigouvian pricing combined with identical lump-sum transfers of toll revenue was predicated on the assumption that the optimum (with no preference for

work start time *per se*) entails different levels of leisure being optimal at different work start times. But in fact the optimum for this case entails a level of leisure that is independent of work start time. The reason is that, even though the *ex ante* identical individuals differ in terms of their productivities and travel times, at the optimum they remain identical in terms of the marginal rate of substitution between consumption and leisure.

Let $q_i = n_i x_i$ be the number of individuals who start work at time i on a given day. This equals the number of individuals who choose this work start time when they work, n_i, times the proportion of days on which each goes into work, x_i. Both the congestion and the private product functions depend on **q**. The government maximizes its net surplus, equal to toll revenues less the cost of the lump-sum redistribution required to meet the exogenous utility level, subject to three constraints. The first constraint is

$$n_i x_i(\mathbf{q},T,\underline{u}) = q_i; \tag{24}$$

This constraint combines the equation $q_i = n_i x_i$ with the demand function for x_i, written as

$$x_i(\mathbf{q},T,\underline{u}) = (1 - l_i(T,\underline{u})) \div (L + t_i(\mathbf{q})). \tag{25}$$

The second constraint is

$$v_i(\mathbf{q},\tau_i,T) \geq \underline{u}\ ; \tag{26}$$

the partial derivatives of v_i are obtained from the following equation:

$$v_i(\mathbf{q},\tau_i,T) = u_i((z_i(\mathbf{q}) - \tau_i)x_i + T,\ 1 - (L + t_i(\mathbf{q}))x_i). \tag{27}$$

The third constraint specifies that the city's population equals the sum over work start times of the number of individuals choosing each work start time. This specification of the optimization problem automatically allows for preferences *per se* with respect to work start times.

Written in Lagrangean form, the optimization problem is

$$\max \sum_i q_i\tau_i - T\sum_i n_i + \sum_i \gamma_i(n_i x_i(\mathbf{q},T,\underline{u}) - q_i) + \sum_i \lambda_i n_i(v_i(\mathbf{q},\tau_i,T) - \underline{u}) + \phi(\sum_i n_i - N). \quad (28)$$

The corresponding first-order conditions are

$$n_i: \quad -T + \gamma_i x_i + \phi = 0 \quad (29)$$

$$q_j: \quad \tau_j - \gamma_j + \sum_i \gamma_i \partial x_i/\partial q_j + \sum_i \lambda_i n_i \partial v_i/\partial q_j = 0 \quad (30)$$

$$\tau_i: \quad q_i + \lambda_i n_i \partial v_i/\partial \tau_i = 0 \quad (31)$$

$$T: \quad - N + \sum_i \gamma_i n_i \partial x_i/\partial T + \sum_i \lambda_i n_i \partial v_i/\partial T = 0 \quad (32)$$

From (25)

$$\partial v_i/\partial \tau_i = - x_i \partial v_i/\partial T \quad (33)$$

$$\partial v_i/\partial q_j = - x_i \Gamma_{ij}(\partial v_i/\partial T), \quad (34)$$

where Γ_{ij} is defined as before, but is now per trip rather than per person. After substituting (33) and (34) into (30) - (32), it can be seen by inspection that the first-order conditions are solved by $\gamma_i = 0$. Then from the equation corresponding to (30), with (33) and (34) substituted in, it follows that

$$\tau_j = \sum_i q_i \Gamma_{ij} ; \quad (35)$$

the second-best congestion toll is the Pigouvian tax that fully internalizes the congestion and agglomeration externalities.

A further set of results can be obtained depending on whether the city's population is larger or smaller than second-best optimal. The derivative of the Lagrangean with respect to N is - ϕ. Thus, if the city's population is (locally) suboptimal, ϕ is negative; if it is (locally) superoptimal, ϕ is positive; and if is (locally) optimal, ϕ is zero. From (29), if ϕ is zero, T = 0. In the institutional context in which the government balances its bugdet,

this implies that revenue from the optimal toll equals zero. Multiplying both sides of (35) by q_j, and summing over j:

$$\sum_j q_j \tau_j = \sum_j q_j \sum_i q_i \Gamma_{ij} \quad (36)$$

Thus, in a city of optimal population size, $\sum_j q_j \sum_i q_i \Gamma_{ij} = 0$. This is the Henry George Theorem in this context.

5. Conclusion

This paper has explored second-best congestion tolling in the presence of unpriced agglomeration externalities in the context of four simple, closed-city models with identical individuals. In the first, the monocentric city model with congestion externalities and Marshallian externalities, the second-best congestion toll equals the first-best congestion toll, since the Marshallian externalities depend on population, and are therefore unaffected by tolling. In the second, individuals decide how many trips to take to the CBD and Marshallian externalities depend on the number of trips. As a result of the Marshallian externalities, with first-best congestion tolling individuals would take too few trips, and consequently the second-best congestion toll should be set below the first-best level and equal to the net negative externality cost, the congestion externality cost minus the positive agglomeration externality benefit; the second-best congestion toll might be negative, if institutionally feasible. Furthermore, the second-best toll can be set at a level that achieves the equal-utilities optimum. In the third, individual trip frequency is fixed but intra-day dynamics are introduced, so that individuals decide on work start time. The modeling of intra-day dynamics is very general, allowing an individual's productivity and his travel time to depend on the distribution of work start times. In this model, congestion tolling affects the distribution of individuals across work start times, and implementation of the time-varying, second-best congestion tolls, which

again equal the net negative externality costs, again results in decentralized attainment of the equal-utilities optimum. The fourth model incorporates both trip frequency and intra-day dynamics. Setting tolls equal to the corresponding Pigouvian taxes and redistributing the revenue as equal lump-sum transfers decentralizes the optimum. .

Some back-of-the-envelope calculations for the second model indicate that, with reasonable parameter values, consideration of uninternalized agglomeration externalities can result in second-best tolls being substantially below first-best congestion tolls.

How seriously should the results of this paper be taken? On one hand, I am uncomfortable basing policy advice on unmeasured and perhaps unmeasurable "agglomeration economies" -- Marshallian external economies of scale. On the other, it is hard (impossible?) to reconcile the twin empirical regularities of small average firm size and a significant positive elasticity of the wage with respect to city population size without Marshallian external economies of scale. If one accepts that Marshallian external economies of scale are important, and that their magnitude is positively related to the amount of travel that individuals undertake, the conclusion that congestion tolls are optimally set below their first-best levels seems unavoidable.

The paper gave scant attention to the effects of congestion tolling on land use, though it did sketch a couple of models that could be developed to explore this.

REFERENCES

Arnottr, R. 1979. Unpriced transport congestion. *Journal of Economic Theory* 21: 294-316.

Arnott, R. 2002. Staggered work hours with a dominant employer. Unpublished manuscript. Boston College, Chestnut Hill, MA.

Arnott, R., A. de Palma, and R. Lindsey. 1993. A structural model of peak period congestion: A traffic bottleneck with elastic demand. *American Economic Review* 83: 161-179.

Arnott, R., T. Rave, and R. Schöb. 2005. *Alleviating urban traffic congestion.* Cambridge, MA: MIT Press.

Arnott, R. and A. Yan. 2000. The two-mode problem: Second-best pricing and capacity. *Review of Urban and Regional Development Studies* 12:170 -199.

Calthrop, E., S. Proost, and K. Van Dender. 2000. Optimal road tolls in the presence of a labor tax. Leuven, Center for Economic Studies.

Duranton, G. and D. Puga. 2004. Micro-foundations of urban agglomeration economies. In J.V. Henderson and J.-F.Thisse, eds. *Handbook of urban and regional economics*, vol.4: Amsterdam: Elsevier.

Fu, S. Forthcoming. Smart café cities. *Journal of Urban Economics.*

Fujita, M. and J.-F. Thisse. 2002. *Economics of agglomeration: Cities, industrial location and regional economic growth.* Cambridge: Cambridge University Press.

Giuliano, G. and T. Golob. 1990. Staggered work hours for traffic management: A case study. *Transportation Research Record* 1280:46-58.

Henderson, J.V. 1974. The sizes and types of cities. *American Economic Review* 64:640-656.

Henderson, J.V. 1981. The economics of staggered work hours. *Journal of Urban Economics* 9:349-364.

Henderson, J.V.

Kanemoto, Y. 1977. Cost-benefit analysis and the second-best land use for transportation. *Journal of Urban Economics* 4:483-503.

Kraus, M. 1998. Discrete-time modeling of congestion on a network. Unpublished notes.

Lévy-Lambert, H. 1968. Tarification des services à qualité variable: Applications aux péages de circulation. *Econometrica* 36:564-574.

Lighthill, M. and G. Whitman. 1955. On kinematic waves, II: A theory of traffic flow on long, crowded roads. *Proceedings of the Royal Society A*:229:317-345.

Lindsey, R. 2002. Staggered work hours and flextime. Unpublished manuscript. University of Edmonton, Alberta.

Mayeres, I. and S. Proost. 1997. Optimal tax and investment rules for congestion types of externalities. *Scandinavian Journal of Economics* 99:261-279.

Moretti, E. 2004. Human capital externalities in cities. In J.V. Henderson and J.-F. Thisse, eds. *Handbook of regional and urban economics*, vol.4. Amsterdam: Elsevier.

Mun, S. and M. Yonekawa. 2000. Flextime, traffic congestion, and urban productivity. Paper presented at the Second International Symposium on "Structural Change in Transportation and Communications in Knowledge Society: Implications for Theory, Modeling and Data," Kyoto, Japan.

Parry, I. and A. Bento. 2001. Revenue recycling and the welfare effects of road pricing. *Scandinavian Journal of Economics* 103:645-671.

Pines, D. and E. Sadka. 1985. Zoning, first-best, second-best, and third-best criteria for allocating land to roads. *Journal of Urban Economics* 17:167-183.

Rosenthal, S. and W. Strange. 2002. Geography, industrial organization, and agglomeration *Review of Economics and Statistics* 85:377-393.

Rosenthal, S. and W. Strange. 2004. Evidence on the nature and source of agglomeration economies. In J.V. Henderson and J.-F. Thisse, eds. *Handbook of regional and urban economics*, vol.4. Amsterdam: Elsevier.

Small, K., C.Winston, and J.Yan.2005. Uncovering the distribution of motorists' preferences for travel time and reliability: Implications for road pricing. *Econometrica* 73:1367-1382.

Solow, R. 1972. Congestion, density, and the use of land in transportation. *Swedish Jouirnal of Economics* 74:161-173.

Verhoef, E. 2002. Second-best congestion pricing in general networks: Heuristic algorithms for finding second-best optimal toll levels and toll points: *Transportation Research B* 36:707-729.

Vickrey, W. 1969. Congestion theory and transport investment. *American Economic Review Proceedings* 59:251-260.

Wheaton, W. 1978. Price-induced distortions in urban highway investment. *Bell Journal of Economics* 9:622-632.

Wilson, J. 1983. Optimal road capacity in the presence of unpriced congestion. *Journal of Urban Economics* 13:337-357.

Wilson, P. 1988. Wage variation resulting from staggered work hours. *Journal of Urban Economics* 24:9-26.

Yoshimura, M. and M.Okumura. 2001. Optimal commuting and work-start-time distribution under flexible work hours system on motor commuting. *Proceedings of the Eastern Asian Society for Transportation Studies* 2:455-469.

Appendix

(not intended for publication)

This appendix solves the optimization problem for the last model in the case that the lump-sum transfer can be made conditional on i. Written in Lagrangean form, the maximization problem is:

$$\max \sum_i (q_i\tau_i - n_iT_i) + \sum_i\gamma_i(n_ix_i(\mathbf{q},T_i,\underline{u}) - q_i) + \sum_i\lambda_in_i(v_i(\mathbf{q},\tau_i,T_i) - \underline{u}) + \phi(\sum_i n_i - N) \qquad (A1)$$

Note that the only difference between this specification and that of the last model is that T is now indexed by i. The corresponding first-order conditions are

$$n_i: \quad -T_i + \gamma_ix_i + \phi = 0 \qquad (A2)$$

$$q_j: \quad \tau_j - \gamma_j + \sum_i\gamma_in_i\partial x_i/\partial q_j + \sum_i\lambda_in_i\partial v_i/\partial q_j = 0 \qquad (A3)$$

$$\tau_i: \quad q_i + \lambda_in_i\partial v_i/\partial\tau_i = 0 \qquad (A4)$$

$$T_i: \quad - n_i + \gamma_in_i\partial x_i/\partial T_i + \lambda_in_i\partial v_i/\partial T_i = 0 \qquad (A5)$$

Substituting (33) into (A4) yields $\lambda_i\partial v_i/\partial T_i = 1$. Substituting this result into (A5) gives $\gamma_in_i\partial x_i/\partial T_i = 0$. Now, from (25), $\partial x_i/\partial T_i = - (\partial l/\partial T_i)/(L + t_i(\mathbf{q})) > 0$ if there is any substitutability between l and c. Thus, $\gamma_i = 0$. But this implies from (A2) that $T_i = \phi$ for all i, so that the the optimal lump-sum redistribution is independent of work start time. It also implies from (A3) that the optimal toll vector completely internalizes the externalities.

The most interesting feature of the solution is that, in the absence of preferences for work start time *per se*, the first-best optimum entails all households consuming the same amount of leisure, independent of work start time.