DELAYS AND UNCERTAINTY IN FREIGHT MOVEMENTS AT CANADA-US BORDER CROSSINGS

William P. Anderson and Andrew Coates, University of Windsor[i]

Introduction

Canada and the United States maintain the largest bilateral trade relationship in world. Although Canada has recently been eclipsed by China as the number one source of US imports, Canada is still the largest destination country for US exports. Most of the trade flows across the long Canada-US land border, with over 50% of total trade going by truck (Transport Canada, 2008, Table EC6). Despite the fact that the border is over 6400 km long (excluding the Alaska border), over 60% of the trade carried by trucks crosses at the Ambassador Bridge across the Detroit River, the Blue Water Bridge across the St Clair River and the Peace and the Lewiston-Queenston Bridges across the Niagara River (Transport Canada, 2008, Table EC10.).

For a number of reasons, these bridges are subject to substantial and often unpredictable congestion. Despite the rapid growth in Canada-US trade since the signing of NAFTA in 1993, only the Blue Water Bridge has been expanded in recent years. In fact, the Peace and Ambassador Bridges still have the same capacities they had when they came into service in the 1920s. US bound trucks approach the Ambassador Bridge through a long series of signalized intersections over roads maintained by the City of Windsor. There is insufficient space for needed expansion of the US inspection plazas at the Peace Bridge. While there are plans to improve and expand the existing infrastructure – including a plan to build a new bridge across the Detroit River and to twin the Ambassador and Peace Bridges – it is still uncertain whether any or all of these plans will come to fruition.

Delays are not just due to inadequate infrastructure, however. The more rigorous border security regime that has emerged in the aftermath of the terrorist attacks of September 11, 2001 is responsible for substantial delay for both trucks and passenger cars (Taylor *et al*, 2004). Despite increased staffing by US Customs and Border Protection and the Canadian Border Services Agency, expanded inspection plazas at some bridges, improved technology and the implementation of trusted traveler and trusted shipper programs[ii], the "thickening" of the border due to security is still viewed as a major problem. (For discussions of this issue see Sands, 2009; Kergin and Matthiesen, 2008; Canadian and US Chambers of Commerce, 2009.)

While delays are a problem for everyone, they are especially onerous for shippers and carriers involved in cross-border supply chains. A large proportion of Canada-US trade comprises intermediate goods moving from one link in a manufacturing supply chain to another. This is especially true in the automotive industry, where tariff-free movements of vehicles and parts dating back to the Auto Pact of 1965 have made it possible to achieve greater scale economies by integrating US and Canadian production facilities (Anastakis, 2005.) Since supply chain managers seek to minimize inventory carrying costs by receiving shipments of components on a just-in-time basis, a truck being delayed at the border could lead to the shutdown of a production line (Andrea and Smith, 2002.) In this context, it is not so much the speed of border crossing as the reliability of crossing times (or inversely, the variance of crossing times) that presents problems as deviations from expected delivery times make it difficult to schedule shipments.

This paper examines a new GPS-based data set of truck border crossing times at the Ambassador, Blue Water and Peace Bridges over the period from July 2008 through June 2009. While the data reveal some interesting trends in average crossing times, the main focus here is on the variability in crossing times and its implications for buffer times in cross-border supply chains.

Delay and Uncertainty in Supply Chains

Consider a simple supply chain where a component is manufactured by a firm (the shipper) on one side of a border crossing and shipped to a firm (the receiver) on the other side. Uncertainty as to crossing times will result in unreliable delivery times, with the potential to shut down the receiver's production if the components do not arrive at a specified time. The receiver may cope with this possibility in two ways: 1) it may find an alternative shipper that does not have to cross the border and is therefore more reliable or 2) it may write penalties for late deliveries into its contract with the shipper that are high enough to compensate for production disruptions. For the sake of this discussion, assume that it adopts the second strategy.

There are two ways that the shipper can protect itself against having to pay the penalty for late deliveries: 1) it can stockpile a supply of the component on the receiver's side of the border or 2) it can build a buffer time into its shipping schedule to cover unexpected border delays. Again, for the sake of this discussion assume the shipper adopts the second strategy. The question is how great should the buffer time be?

Figure 1 shows a schedule of incremental costs due to uncertainty that the shipper faces. The planned arrival time (PAT) is the time at which the receiver has stipulated that the components should be delivered. The shipper incurs a cost whether the goods arrive before or after the PAT. The cost per minute of being early (the early penalty rate) is due to the labour and capital that is idle at the receiver's location between the arrival time and the PAT. This assumes that the receiver will not accept the components early, but even if it will it may not be possible to reassign the truck and driver to other shipments so their time is still wasted. The cost of being late is the penalty for being late imposed by the receiver. Here the penalty is shown as a rate per minute late (the late penalty rate), but a discrete lateness penalty might also be included.

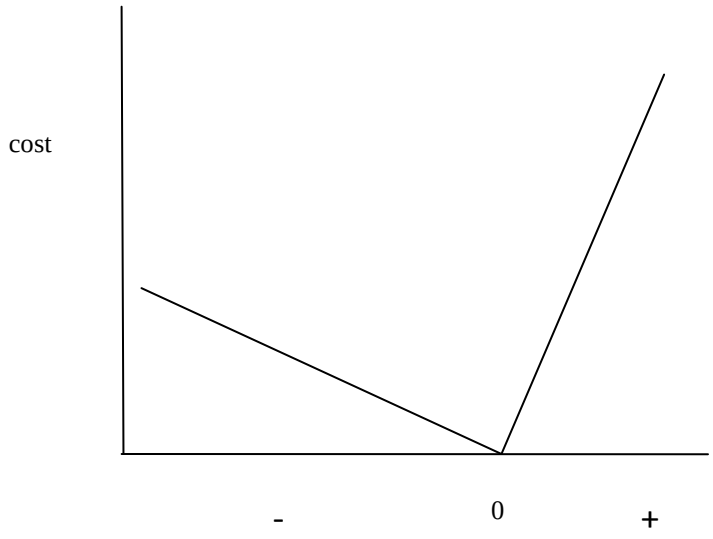
In this situation, we can make a couple of generalizations about the buffer time the firm will choose. First, if there is not variability in crossing times the shipper can always schedule the truck so that the

deviation from the PAT is zero, therefore the buffer time will be zero. Second, even if there is variability in the crossing time, if the slopes of the lines representing the costs of being early and late are the same (i.e. the early and late penalty rates are equal), the shipper is indifferent between being early and late and will therefore assign a buffer time of zero. However, if the slope of the late cost line is greater than the slope of the early cost line, the shipper will assign a buffer time that is increasing in the difference between the late and early costs and in the magnitude of the variability in crossing times.

The problem of choosing the optimal buffer time has been treated formally, first by Gaver (1968) and later with a variety of extensions reviewed in Noland and Polak (2002). While most of the literature is focused on commuters' departure time choice, Small *et al* (1999) note that the framework is equally applicable to truck and other freight scheduling. We will return to this literature later, but the consistent finding is that the buffer time is increasing in variance of travel times and the ratio of the late penalty rate to the early penalty rate.

Measurement of the early penalty rate is relatively straightforward. It may be set equal to cost per minute of free flow travel (the cost per km multiplied by the average speed in km per minute) or set slightly lower since an idle truck does not consume as much fuel as a moving truck. It may also be adjusted downward because in some circumstances it is possible to deliver goods early and reassign the truck to other service. Measuring the late penalty rate is more involved. The receiver may build its own buffer time between its specified PAT and the time when the components are actually needed, in which case there may be a grace period during which the penalty is zero. If the receiver were a retailer, the cost might be the loss of profits on sales not made because goods were out of stock. For a manufacturer, however, the penalty must be defined on the basis of the costs of production disruption. Furthermore, failure to consistently deliver goods on time could result in the shipper being "de-sourced," a possibility that may also be factored in to the shipper's calculation of late penalty.

Figure 1: Cost of Deviation from PAT



While it is difficult to find information on the effective late penalty, a study by the Center for Automotive Research (2002, p. 14) provides at least some perspective. The study estimates that an automotive assembly plant generates about US$1.5 million per hour. At a return on sales of 4%, the lost income due to shutting down the plant is US$60,000 per hour, or $1000 per minute. This is an extreme case, because assembly plants are very large. A disruption in a tier one part supplier, for example, would have a lower cost. Furthermore, the effective late penalty would probably be smaller than this, since the assembly plant can make up for the production at a later time (with the increased cost of overtime labour rates.) But the main point is that the cost of potential production disruptions is large relative to the $25 - $100 per hour that is normally applied to delay time for commercial vehicles in cost-benefit analyses. Thus the late penalty rate for trucks

in just-in-time supply chains will be much higher than the early penalty rate, so shippers will assign relatively high buffer times.

Border Crossing Time Data

The data we present below are from the Border Wait Time Measurement Project, an innovative relationship between Transport Canada, Ontario Region, and Turnpike Global Technologies (TGT), a firm that provides GPS vehicle tracking services to trucking firms. With the agreement of TGT clients, travel times across a number of major crossings are extracted from digital trip logs. The crossing times cover movement from the bridge approach to a point beyond the inspection plaza on the far side, so they capture both the time spent crossing the bridge and the time spent passing through the plaza. (Times in zones approaching the bridges are also available but we only use the crossing time in this paper.)

The data set consists of individual records for each crossing truck. Data elements include the date and clock time at the beginning of the crossing, and the crossing time in minutes Data collection commenced in 2004, but the data reported on here are from July 1, 2008 to June 31, 2009. Descriptive statistics for crossing times at the four major bridges are shown in Table 1.

The mean values for all bridges are less than fourteen minutes and the medians are less than 8 minutes. However, standard deviations are quite high. The data are positively skewed, which is not surprising since the observed values are truncated at zero.

The Peace Bridge has a particularly high standard deviation and skewness statistic. This appears to arise from the fact that there are 21observations with crossing times over 200 minutes, compared with two each for the Blue Water and Ambassador Bridges and one for the Lewiston-Queenston Bridge. Times in this range usually represent trucks that are singled out for secondary inspection, which means they are directed to a separate area where they can be more closely scrutinized without holding up the queues at the primary inspection booths.

Table 1 Descriptive Statistics for Crossing Time in Minutes at Four Border Crossings, July1, 2008 to June 31, 2009.

Statistic	Bridge			
	Ambassador	Blue Water	Peace	Lewiston - Queenston
Mean	11.3	13.8	13.2	10.8
Median	7.6	7.5	7.9	5.2
Mode	4.9	3.3	5.6	2.3
Standard Deviation	9.8	18.3	24.6	14.2
Skewness	4.4	4.5	13.0	3.55
Min	0.8	1.0	1.1	1.0
Max	238.4	288.6	732.1	217.5
observations	20,883	5,398	8,273	29,335

Casual observers may be surprised that the mean values for these crossing times are so low. To some extent this reflects the slowdown in the economy, especially the automotive sector, during the period in question. But it also suggests that expectations tend to be conditioned on the unusually long crossing times that occur occasionally. Thus variability in crossing times, which is shown by the standard deviations to be quite high, is a critical part of the overall picture.

Figure 2 shows the mean and median crossing times by month over the period July 2008 to June 2009 for the bridges with the largest numbers of observations: the Ambassador and Lewiston-Queenston Bridges. None of the bridges show a general decreasing trend, despite the onset of the economic slowdown during the study period.

Figure 3 shows the mean crossing time by time of day for the same two bridges. Neither bridge shows the typical morning and evening peaks associated with urban traffic congestion.

**Figure 2a Ambassador Bridge to US,
Average Crossing Times by Month,
July 2008 - June 2009**

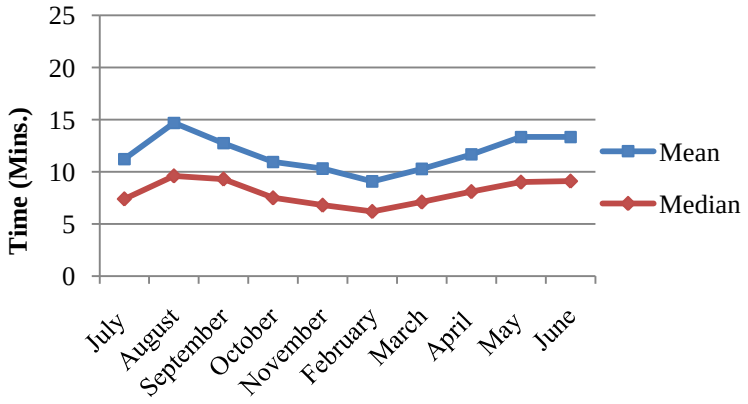


**Figure 2b Lewiston-Queenston Bridge to US
Average Crossing Times by Month
July 2008 - June 2009**

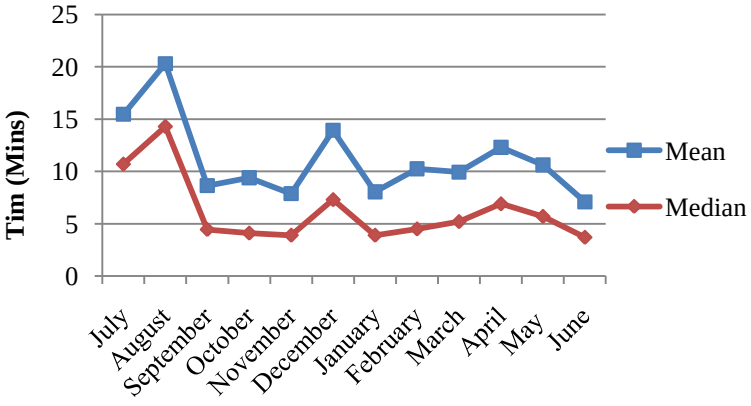


Figure 3a Ambassador Bridge to US
Average Crossing Times by Hour
July 08 - June 09

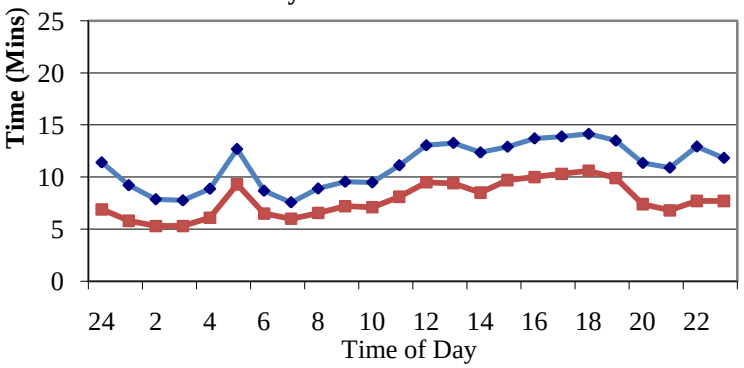
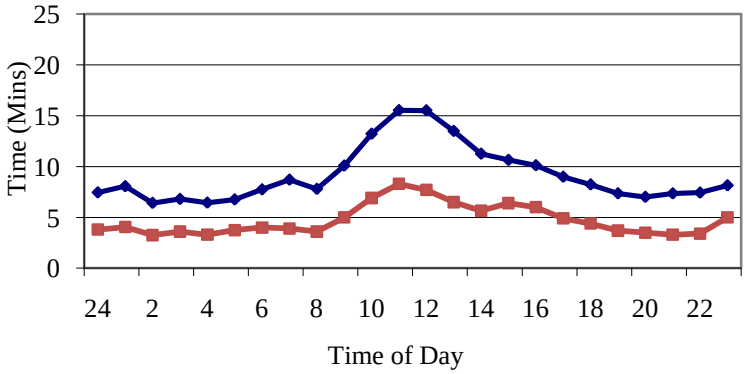


Figure 3b Lewiston-Queenston Bridge to US
Average Crossing Times by Hour
June 2008 - June 2009



Variability in Crossing Times

The buffer index is an intuitive measure of variability in travel time as represented by required buffer time (Cambridge Systematics *et al* (2008)). It essentially tells you how much extra time must be budgeted to ensure that a shipment has a 95% chance of arriving on or before the PAT. It is calculated as follows:

$$BI = \frac{t_{95} - \overline{t}}{\overline{t}} \times 100$$

where t_{95} is the 95th percentile crossing time and $\overline{t}$ is the average crossing time. Its interpretation is as follows: if the BI = 100, then the required buffer time is 100% of the average crossing time. In other words, the total time budgeted for the crossing is twice the average crossing time.

The choice of 95% as the on-time confidence level is arbitrary. Therefore, in Table 2 the buffer index is defined at three confidence levels, 90%, 95% and 99%. Clearly the choice of percentile is very important – for the Blue Water and Peace Bridges the 99th percentile value is around six times as high as the 90% value and for all but the Ambassador Bridge the 99th percentile is more than twice the 95th percentile.

Table 2 Buffer Index at Four Border Crossings, July 1, 2008 to June 31, 2009

Confidence level	Bridge			
	Ambassador	Blue Water	Peace	Lewiston - Queenston
90%	101.4	102.9	92.0	141.2
95%	163.5	189.1	179.7	258.2
99%	312.6	609.3	525.9	536.4

Optimal Buffer Times

From our buffer index calculations it is clear that the cost of uncertainty of border crossing times depends upon the shipper's tolerance for being late. If the shipper judges that it is acceptable to arrive after the PAT 5% of the time a much smaller buffer time is needed than if it is acceptable to be late only 1% of the time. It stands to reason that this tolerance depends on the cost of being late relative to the cost of being early. A model that provides a rigorous way to determine travel decision making in an environment of uncertainty due to traffic congestion, and that takes account of the relative costs of being early and being late, was developed by Noland and Small (1995). While this framework was originally intended to address an individual's morning commute, Small *et al* point out that it is equally applicable to problems of freight scheduling.

In this model, a commuter has to choose a home departure time t_d in order to minimize an expected utility function

$$E[U(t_d)] = \alpha E[T(t_d)] + \beta E[SDE(t_d)] + \gamma E[SDL(t_d)] + \theta p_L(t_d)$$

Where *T* is total travel time, *SDE* is schedule delay early and *SDL* is schedule delay late, α, β and θ are per minute costs of total travel time, being early and being late. (β and γ are equivalent to the early penalty rate and late penalty rate discussed earlier.) *SDE* and *SDL* are calculated as the absolute values of differences between the actual arrival time and the PAT. Finally, θ is a discrete lateness penalty, which is multiplied by the probability of being late p_L . Note that all time variables are a function of the departure time t_d . This is because the model was developed for the morning commute period, when both travel time and uncertainty are increasing in departure time up to a peak and declining thereafter.

The transfer of the underlying logic to freight transportation is straightforward. The expected utility is always negative, since the

terms on the right hand side are all part of a generalized cost of commuting. Thus, maximizing expected utility is equivalent to minimizing expected cost. In the case of freight transportation, elements of travel time are not necessarily functions of departure time, which can simplify the model significantly.

Maximizing the expected utility requires the separation of travel time into a constant component (free flow time plus recurrent delay) and a stochastic component (non recurrent delay) and specifying a probability distribution for the latter. Noland and Small (1995) specify an exponential distribution with parameter *b*, which is both the mean and standard deviation.[iii] Among the results of their derivation is an expression for the optimal probability of being late:

$$p_L^* = \frac{b(\beta - \alpha\Delta)}{\theta + b(\beta + \gamma)}$$

Where Δ is the rate of change in delays with respect to the departure time. Bates *et al* (2001) point out that if it is assumed that $\Delta = \theta = 0$, meaning that delays are not time of day dependent and there is no discrete lateness penalty, the optimal probability simplifies to

$$p_L^* = \frac{\beta}{\beta + \gamma}$$

The implication is simple: if the cost of being late is 9 times as great as the cost of being early, then the trip will be scheduled based on the expectation of being late 10% of the time. This corresponds to choosing the buffer time based on the the 90th percentile buffer index. (Note that while the optimal probability is no longer dependent on the variance of the probability distribution, the departure time is, because the required buffer time is increasing in the variance.) If being late is 99 times as costly as being early, the trip will be scheduled based on the 99th percentile buffer index.

For the extreme example mentioned above (where the late penalty rate is $1000 per minute) the schedule will be based on an even

higher basis than the 99th percentile. Only if the early penalty rate were set at an unrealistically high $10 per minute ($600 per hour), would the ratio $\beta/\beta+\gamma$ yields an optimal late probability of .01, corresponding to the 99th percentile. For most shipments the late penalty will be much lower. However at a more realistic early penalty of $1 per minute ($60 per hour) a late penalty of $100 per minute ($6000 per hour) would call for a buffer time based on the 99th percentile.

Taking the Blue Water Bridge as an example, the buffer index for the 99th percentile would imply that a total of one hour and 37 minutes would be scheduled even though the average crossing time is only 13.8 minutes. While this may seem unrealistic, it is not out of line with what limited empirical information is available. For example Taylor *et al* (2004) found through interviews of industry participants that a two hour rule of thumb for border crossings was common. (It should be noted, however, that this was in 2004 when delays were considerably worse that in 2008.)

Concluding Comments

Using the Buffer Index in conjunction with the optimal late probability derived from the model of Noland and Small (1995) provides a theoretically-grounded explanation for why times scheduled for crossing may be much higher than observed average times and suggests that variability in delay may be significantly more costly than delay *per se.* As a predictive tool, however, it is still in need of significant refinement. There are three important weaknesses of the analysis as it now stands:

First, actual buffer times depend critically upon the late penalty rate, which may vary significantly across shipments. Certainly it is higher for goods in just-in-time supply chains than for other goods, but even within that category it may vary significantly depending on the value of the production disruption that would occur if shipments were late.

Second, inventory policy provides a substitute for adding buffer time to schedules as a strategy to insure against crossing time uncertainty. At some level of required buffer time it may be less expensive to

stockpile inventory on the opposite side of the border so that it can be used in place of a delayed shipment. Taylor *et al* (2004) in the most comprehensive empirical study of border delays to date, found that extra inventory carrying costs and buffer time costs made roughly equivalent contributions to aggregate border crossing costs.

Third, the observed distribution of the travel times may not accurately represent the distribution faced by a specific truck. For example, almost 20 percent of shipments crossing the Ambassador Bridge are compliant with the Free And Secure Trade (FAST) program, which pre-screens shipments and allows them to use faster queues. Carriers associated with automotive supply chains are more likely than others to incur the substantial costs necessary to become compliant with the FAST program. Thus, many carriers with high late penalty rates are able to reduce buffer times through FAST membership.

Addressing these issues will require both extensions to the modeling framework and survey-based empirical research to learn more about the costs and choices faced by manufacturers and carriers. If sufficient refinements can be made, the framework will provide a useful method for quantifying the cost of crossing time uncertainty for benefit-cost analyses and other policy analyses.

References

Anastakis, Dimitry (2005) *Auto Pact: Creating a Borderless North American Auto Industry, 1960-1971*, Toronto: University of Toronto Press.

Bates, John, John Polak, Peter Jones and Andrew Cook (2001) The valuation of reliability to personal travel, *Transportation Research E*, 37:191-229.

Cambridge Systematics, Dowling Associates, System Metrics Group and Texas Transportation Institute (2008) *Cost Effective Performance Measures for Travel Time Delay, Variation ad Reliability*, NCHRP Report 618, Washington: Transportation Research Board.

Canadian and US Chambers of Commerce (2009) *Finding the Balance: Shared Border of the Future.* Washington and Ottawa.

Gaver, Donald P. (1968) Head start strategies for combating congestion, *Transportation Science*, 2: 172-181.

Kergin, Michael and Birgit Matthiesen (2008) *A New Bridge for Old Allies*, Ottawa: Canadian International Council.

Noland, Robert B. And John W. Polak (2002) Time travel variability: a review of theoretical and empirical issues, *Transport Reviews*, 22: 39-54.

Noland, Robert B. And Kenneth A. Small (1995) Travel-time uncertainty, departure time choice and the cost of the morning commute, Institute of Transportation Studies, University of California, Irvine, UCI-ITS-WP-95-1.

Sands, Christopher (2009) *Toward a New Frontier: Improving the US-Canadian Border*, Washington: The Brookings Institution.

Small, Kenneth A., Robert B. Noland, C. Chu and D. Lewis (1999) *Valuation of Travel Time Savings and Predictability in Congested Conditions for Highway User-Cost Estimation*, NCHRP Report 431, Washington: The Transportation Research Board.

Taylor, John, Douglas R. Robideaux and George C. Jackson (2004) U.S.-Canada Transportation and Logistics: Border Impacts and Costs, Causes and Possible Solutions, *Transportation Journal*, 43(4):5-21.

---

[i] We wish to thank Tony Shallow, Transport Canada, for help with data access and for crucial suggestions and advice. Any errors are the exclusive responsibility of the authors.

[ii] The FAST program for fright and the NEXUS program for personal transportation are administered jointly by the US and Canadian governments. Both conduct extensive security checks in advance so that trucks and drivers in the case of FAST and individual travelers in the case of NEXUS can pass through the border with reduced inspection requirements.

[iii] Based on empirical studies of non-recurrent delay and log-normal distribution would have been preferable, however the authors found it to be intractable in their model.