The Social Cost of Congestion Games by Imposing Variable Delays

Josep Díaz[a,1], Ioannis Giotis[a,b,1,2], Lefteris Kirousis[c,2,*], Ioannis Mourtos[b,2], Maria Serna[a,1]

[a]*Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Barcelona*
[b]*Department of Management Science and Technology, Athens University of Economics and Business, Greece*
[c]*Department of Mathematics, National and Kapodistrian University of Athens, Greece and Computer Technology Institute Press "Diophantus", Patras, Greece*

Abstract

We describe a new coordination mechanism for non-atomic congestion games that leads to a (selfish) social cost which is arbitrarily close to the non-selfish optimal. This mechanism incurs no additional cost, in contrast to tolls that typically differ from the social cost as expressed in terms of delays.

Keywords: Congestion games, Price of anarchy, Coordination mechanisms.

1. Introduction

Selfish behavior is one of the primary reasons many systems with multiple agents deviate from desirable outcomes. Allowing players to prioritize solely their own benefit can lead to social inefficiency, even in outcomes where no one is better off compared to an optimal solution.

This type of behavior has been analyzed in various contexts and has often been verified in practice. A key such area is transportation and network routing where selfish selection among possible routes can lead to congestion with accompanying economical and environmental issues.

Various approaches have been proposed to steer the selfishly constructed outcome towards optimal social welfare. The main idea is usually to incentivize the users to alter their selections to ones that lead to socially better outcomes, typically through the use of tolls or similar measures.

We propose an alternative approach that alters the way users experience latency and can offer significant improvements on social cost, however drivers still get to pick their own route In more detail, instead of all users experiencing the same latency, we propose to implement variable latencies through a prioritization scheme. That is, we allow for some users to experience smaller latencies than before, while others to experience longer ones. We employ known results to show that our system, if users behave selfishly as expected, achieves the optimal social welfare. To make such a system practical we present a discretization of the theoretical continuous functions that approximates the optimal social welfare.

We also wish to emphasize the distributed and decentralized nature of our system. As explained in the next sections, each resource (road or highway in the transportation setting) implements the desired changes individually and independently. It is important to note that our system's average latency on each road, as experienced by the users, is at least equal, and actually closely matches, the road's average latency without the system in place, hence our system falls under the notion of coordination mechanisms, i.e. no "cheating" in the form of network improvements, which typically carry significant cost, is introduced. Also, there is no imposing of tolls (transfer of social cost to a different type); we simply distribute the resource differently. This holds on any instance, not just in equilibrium settings, which means that even in non-stable situations we do not get worse performance. Furthermore, we do not need to know the *demand* in advance, i.e. our system delivers close to the social optimum for all possible total amounts of traffic. Our only requirement is that the latency induced on each road is a non-negative, non-decreasing, continuously differentiable and convex function of the traffic.

We believe that our system has a strong applicability potential. For example, some countries have already implemented metered highway entrance ramps which can vary the latency of incoming drivers. Traffic lights may also be used in an urban environment to implement this

*Corresponding author

Email addresses: `diaz@cs.upc.edu` (Josep Díaz), `igiotis@cs.upc.edu` (Ioannis Giotis), `lkirousis@math.uoa.gr` (Lefteris Kirousis), `mourtos@aueb.gr` (Ioannis Mourtos), `mjserna@cs.upc.edu` (Maria Serna)

[1]Supported by funds from the Spanish Ministry of Economy and Competitiveness (MINECO) and the European Union (FEDER funds) under grant TIN2013-46181-C2-1-R (COMMAS) and grant SGR 2014 1137 (ALBCOM) of the Catalan government.

[2]This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

aspect of our mechanism. We deliberately leave the prioritization scheme generic to allow for different such approaches with our only requirement being that users choosing to alter their current selection are forced to experience maximal latency in their new selection, a reasonable requirement as typically someone that alters her selection in a running system ends up at the end of the queue.

We examine our system in the generic scheme of congestion games, to emphasize that it admits applications beyond traffic routing. One interesting application could be in the context of job scheduling on computing resources. Again, in a typical model, each user choosing to use a particular resource experiences the same latency, for example computing jobs running in parallel on a computer. By prioritizing jobs according to our proposed mechanism, such that some jobs complete faster and some slower than before, we can achieve optimal average job completion times under selfish behavior. We note that this can easily be implemented by an administrator (human or computerized) using system priorities.

2. Related work

The fact that selfish behavior can lead to inefficiency has long been studied in the context of transportation theory [1, 2]. More recently, Koutsoupias and Papadimitriou introduced the *Price of Anarchy* as a measure of this inefficiency [3, 4]. Exploration of this metric in the context of selfish routing was then greatly progressed by Roughgarden and Tardos [5, 6] who bounded the price of anarchy for different classes of latency functions.

Naturally, ways to improve inefficient outcomes have been investigated, with a prime example being the imposement of *tolls* [7, 8, 9]. While this approach achieves optimal social welfare regarding latencies, it introduces a cost separation to the players as the tolls' cost is affecting behavior but is not accounted for in the objective function.

Coordination mechanisms have recently been introduced by Christodoulou et al. [10] as a way to "shape" latency functions to steer the selfishly dictated outcome towards greater social welfare. There are two main restrictions in the type of coordination mechanisms defined in [10], namely that the latency per resource is not decreased and that the benchmark optimal social welfare, against which the mechanism is measured, is still the original one without any additional latencies possibly imposed by the mechanism. It has recently been shown that indeed such mechanisms can positively affect social welfare [11]. Our approach sustains the non-decreasing latency on average but not on every user, as a prerequisite for achieving, through 'coordination', a significantly lower price of anarchy than the mechanism of [10]. In fact, the average latency per user within our system can be made arbitrarily close to the unique latency per user without the system in place.

The approach of differentiating the latency per resource is also explored from an algorithmic perspective by Harks et al. [12] but not with the same scheme. The results of Farzad et al. [13] are closer to our work. However, in the later, it is the *strategic* equilibrium without the mechanism of [13] in place that matches the optimal under that mechanism; i.e., the strategic equilibrium under that mechanism may in general differ substantially from the non-selfish optimal.

3. Model

We define a congestion game $(E, l, \mathcal{S}, P, d)$ in the generic sense but using network routing (or alternatively, flow) terminology for convenience. First, a set E of edges with an associated non-negative, non-decreasing, continuously differentiable and convex $l_e()$ *latency* function for each edge. We note that these assumptions are typical for latency functions.

In P, we have n player types $1, 2, \ldots, n$. For each player type i we have a source-sink pair (s_i, t_i). We also have a finite set $\mathcal{S}_i$ of finite sequences of elements of E, called the strategy set of player type i. A particular element $S \in \mathcal{S}_i$ is a single strategy of player type i, also referred to as a *path* from s_i to t_i. Finally, we have a flow (or traffic) demand d_i for each player type i.

We assume that each player type corresponds to a continuum of nonatomic players, each with negligible flow, i.e. they can be arbitrarily divided into the various paths. An infinitesimal part of the flow or traffic will often referred to as a *user*. Let x_i^S denote a nonnegative real representing the part of demand corresponding to player type i that uses strategy (path) S and x_i the vector for the strategy set $\mathcal{S}_i$, i.e. $x_i = (x_i^S)_{S \in \mathcal{S}_i}$. The vector x for all x_i's is called a *flow* if for all player types i, $\sum_{S \in \mathcal{S}_i} x_i^S = d_i$. We define the part of the demand of a player type i that uses edge e as

$$x_e^i = \sum_{\{S : S \in \mathcal{S}_i, e \in S\}} x_i^S,$$

and the total flow through an edge e as

$$x_e = \sum_{i=1\ldots n} x_e^i.$$

In related literature, the *cost* induced to each player type i by a flow x is defined to be $c_i(x) = \sum_{e \in E} l_e(x_e) \cdot x_e^i$. The cost of the total flow through an edge e is defined to be

$$c_e(x_e) = l_e(x_e) \cdot x_e,$$

whereas the social cost is defined to be

$$C(x) = \sum_{e \in E} l_e(x_e) \cdot x_e.$$

For reference, we now give the notion of Wardrop equilibrium in our setting.

Definition 1. *We say that the flow vector x is in Wardrop equilibrium if for all player types i and for any pairs of strategies (paths) $S_1, S_2 \in \mathcal{S}_i$, if $x_i^{S_1} > 0$ then the following holds:*

$$\sum_{e \in S_1} l_e(x_e) \leq \sum_{e \in S_2} l_e(x_e). \tag{1}$$

4. Variable delay mechanism

Given a congestion game $(E, l, \mathcal{S}, d)$ with non-negative, non-decreasing, continuously differentiable and convex latency functions, we propose a coordination mechanism which differentiates the latency experienced by different users as follows:

Let $N = (N_e)_{e \in E}$ be a sequence of positive integers indexed by the set of elements (edges) E to be called a *batch system*. A positive integer $b \leq N_e$ is referred to as a batch index (or just batch) at edge e. At each edge e, the total flow x_e through e is split into N_e batches of *equal* size x_e/N_e each. Each batch induces a different latency cost to its corresponding flow, with batches of a larger index getting larger latencies, as formally defined below. Note that this means that different parts of the flow of some player type could receive different latencies. The way that this split is implemented does not affect our results, i.e., the assignment of flow to batches can be performed by any desired policy (e.g., randomly or first-come-first-served or through priority lists).

Now consider the following functions, known as marginal–cost latency functions:

$$\hat{l}_e(x_e) = c'_e(x_e) = l_e(x_e) + l'_e(x_e) \cdot x_e, \tag{2}$$

where $c'_e()$, $l'_e()$ are the derivatives of $c_e()$, $l_e()$, respectively. The latency induced is not going to be equal among users at an edge e. Instead, the flow of any player type and through any path S at batch b receives latency $\hat{l}_e((b/N_e)x_e)$ per unit. Users are interested in minimizing their own latency. We refer to the model of applying equal latency to all users as the uniform latency or classical model.

Since each batch b receives latency $\hat{l}_e((b/N_e)x_e)$ per unit, we define the cost with respect to the batch system at an edge with flow x_e to be:

$$\hat{c}_e(x_e) = (x_e/N_e) \sum_{b=1}^{N_e} \hat{l}_e((b/N_e)x_e)$$

and the social cost with respect to the batch system

$$\hat{C}(x) = \sum_e \hat{c}_e(x_e).$$

Note that $\hat{c}_e(x_e) \geq \int_0^{x_e} \hat{l}_e(z)dz = l_e(x_e)x_e = c_e(x_e)$, therefore we do not decrease the cost as per the coordination mechanisms' doctrine and as we shall see later, any cost increase can be made arbitrarily small.

Although we state above that all batches of the flow at a particular edge are of equal size x_e/N_e, let us note that this is not essential for our proofs, i.e., all technical arguments go through for arbitrary batch sizes. We assume batches of equal size to avoid cumbersome notation and thus improve clarity. In addition, when the number of batches tends to infinity, all batch sizes approach zero; that asymptotic case is important because, then, the Price of Anarchy under our model approaches 1 while the cost approaches the cost under the uniform latency model.

Given a path S, a sequence of batch indices $b_e, e \in S$ is called a batch assignment for S.

Definition 2. *We say that the flow vector x is in equilibrium with respect to the batch system if for all player types s i and for any pairs of strategies (paths) $S_1, S_2 \in \mathcal{S}_i$, if $x_i^{S_1} > 0$, then for every batch assignment $b_e, e \in S_1$, the following holds:*

$$\sum_{e \in S_1} \hat{l}_e((b_e/N_e)x_e) \leq \sum_{e \in S_2} \hat{l}_e(x_e). \tag{3}$$

Intuitively, if a user (infinitesimal part of flow) changes path, then we assume that it gets to the last batch of every edge of the new path. Indeed this is so because in the right hand side of the above equation the cost of the last batch appears for all edges; whereas on the left we have the cost of an arbitrary batch sequence b_e along S_1. What the above equation expresses is that under this assumption, there is no strict gain in cost a user experiences if it unilaterally implements a change of path (users are assumed to be anonymous, so the batch at an edge for a particular user is not well defined; this is the reason arbitrary batch sequences within the various edges of a path are taken on the left side).

We now have the following:

Lemma 3. *The flow vector x is in equilibrium with respect to the batch system iff it is in Wardrop equilibrium with respect to the marginal–cost latency functions $\hat{l}_e(x_e) = l_e(x_e) + l'_e(x_e) \cdot x_e$, i.e. iff for all players i and for any pairs of strategies (paths) $S_1, S_2 \in \mathcal{S}_i$, if $x_i^{S_1} > 0$, then $\sum_{e \in S_1} \hat{l}_e(x_e) \leq \sum_{e \in S_2} \hat{l}_e(x_e)$.*

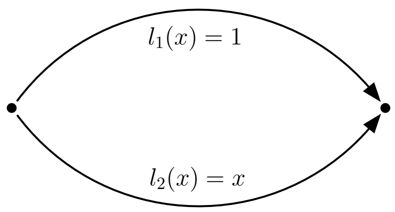
Proof. For the sufficiency of Wardrop equilibrium with respect to the marginal-cost latency notice that since $l_e()$ are convex, $\hat{l}_e$ are non-decreasing. For the necessity notice that because the inequality in Definition 2 holds for any selection of batch indices, and therefore also for $b_e = N_e$. □

We now state the following well known theorems derived from the literature [14, 6].

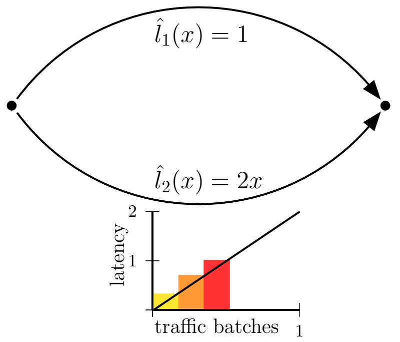
Theorem 4. *When the latencies are non-negative, continuous and non–decreasing, there always exists at least one Wardrop equilibrium.*

Theorem 5. *If x and x' are flow vectors in Wardrop equilibrium then $l_e(x_e)x_e = l_e(x'_e)x'_e$ for all edges e. This also shows a unique social cost for all Wardrop equilibria.*

Theorem 6. *A flow vector x in Wardrop equilibrium with respect to the marginal–cost latencies $l_e(x_e) + l'_e(x_e) \cdot x_e$ has optimal social cost $C(x)$ with respect to the latency functions l_e.*



(a) Original Pigou's network



(b) Pigou's network with variable delays

Figure 1: An example on Pigou's network

We will now transfer these results in our variable delay batch setting.

Theorem 7. *Under the variable delay mechanism, any batch system has a unique equilibrium (as defined in Definition 2). Moreover, there is always a suitable batch system whose cost, with respect to the batch system, when in equilibrium (in the sense of Definition 2), is arbitrarily close to its optimal social cost of the uniform latency model.*

Proof. Indeed by the preceding Theorems 4–6, and by Lemma 3, it suffices to prove that if a flow x is in equilibrium with respect to a suitable batch system, then its cost $\hat{C}(x)$ with respect to the batch system is arbitrarily close to the social cost $C(x)$ of the uniform latency model with respect to latencies l_e. This however is immediate to see since

$$\hat{c}_e(x_e) = (x_e/N_e) \sum_{b=1}^{N_e} \hat{l}_e((b/N_e)x_e)$$

can be made arbitrarily close to

$$\int_0^{x_e} \hat{l}_e(z)dz = l_e(x_e)x_e = c_e(x_e)$$

by choosing for each e a large enough N_e. Note that at this point one needs an upper bound of the demand to pick a large enough N_e but this is not required if the approximability bound is not needed. $\square$

5. Discussion

Interestingly, even a small number of batches can offer significant improvements in certain situations. We illustrate the classic Pigou's network as an example with a total traffic of 1. In the original network (see Figure 1a), it is well known that the Price of Anarchy is 4/3 since all users would pick the lower edge inducing latency of 1 to everyone while the optimal solution would be for the users to be split evenly among the edges with an average latency of 3/4. By using 3 batches with variable delays (Figure 1b), in equilibrium, half of the traffic would pick the lower edge and be distributed uniformly among the 3 batches with an induced latency of $1/3, 2/3$ and 1 respectively. The average latency would be

$$0.5 \cdot 1 + 0.5 \cdot 1/3 \cdot (1/3 + 2/3 + 1) = 5/6.$$

The resulting Price of Anarchy compared to the optimal solution in the original network would be 10/9, a significant improvement from 4/3. Even with just two batches, the resulting Price of Anarchy would be 7/6

6. Acknowledgements

We would like to thank George Christodoulou for the very helpful discussions. Also we thank an anonymous referee of a previous draft of this work for pointing to us the results by Farzad et al. [13] and Harks et al. [12]

References

[1] A. C. Pigou, The Economics of Welfare, Macmillan, 1920.
[2] J. G. Wardrop, Some theoretical aspects of road traffic research, in: ICE Proceedings: Engineering Divisions, Vol. 1, Thomas Telford, 1952, pp. 325–362.
[3] E. Koutsoupias, C. H. Papadimitriou, Worst-case equilibria, Computer Science Review 3 (2) (2009) 65–69.
[4] C. H. Papadimitriou, Algorithms, games, and the internet, in: J. S. Vitter, P. G. Spirakis, M. Yannakakis (Eds.), STOC, ACM, 2001, pp. 749–753.
[5] T. Roughgarden, É. Tardos, How bad is selfish routing?, J. ACM 49 (2) (2002) 236–259.
[6] T. Roughgarden, Selfish routing and the price of anarchy, MIT Press, 2005.
[7] R. Cole, Y. Dodis, T. Roughgarden, Pricing network edges for heterogeneous selfish users, in: L. L. Larmore, M. X. Goemans (Eds.), STOC, ACM, 2003, pp. 521–530.
[8] L. Fleischer, K. Jain, M. Mahdian, Tolls for heterogeneous selfish users in multicommodity networks and generalized congestion games, in: Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on, IEEE, 2004, pp. 277–285.
[9] G. Karakostas, S. G. Kolliopoulos, Edge pricing of multicommodity networks for heterogeneous selfish users, in: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, 2004, pp. 268–276.
[10] G. Christodoulou, E. Koutsoupias, A. Nanavati, Coordination mechanisms, Theor. Comput. Sci. 410 (36) (2009) 3327–3336.
[11] G. Christodoulou, K. Mehlhorn, E. Pyrga, Improving the price of anarchy for selfish routing via coordination mechanisms, in: Proceedings of the 19th European conference on Algorithms, Springer-Verlag, 2011, pp. 119–130.
[12] T. Harks, S. Heinz, M. E. Pfetsch, Competitive online multicommodity routing, Theory of Computing Systems 45 (3) (2009) 533–554.
[13] B. Farzad, N. Olver, A. Vetta, A priority-based model of routing, Chicago Journal of Theoretical Computer Science 1.
[14] M. Beckmann, C. McGuire, C. B. Winsten, Studies in the Economics of Transportation, Yale University Press, 1956.