Pricing Networks with Selfish Routing*

Richard Cole† Yevgeniy Dodis‡ Tim Roughgarden§

May 22, 2003

1 Introduction

1.1 Selfish Routing and Marginal Cost Pricing

We study the negative consequences of selfish behavior in networks and economic means of influencing such behavior. We focus on a simple model of *selfish routing*, defined by Wardrop [12] and first studied from a theoretical computer science perspective by Roughgarden and Tardos [10]. In this model, we are given a directed network in which each edge possesses a latency function, describing the common latency (delay) experienced by all traffic on the edge as a function of the edge congestion. There is a fixed amount of traffic wishing to travel from a source vertex s to a sink vertex t, and we assume that the traffic comprises a very large population of users, so that the actions of a single individual have negligible effect on network congestion. A common way to measure the quality of an assignment of traffic to s-t paths is by the sum of all travel times—the *total latency*. We assume that each network user acts selfishly and routes itself on a minimum-latency path, given the network congestion due to the other users. In general such a "selfishly motivated" assignment of traffic to paths (a *Nash equilibrium*) does not minimize the total latency; put differently, the outcome of selfish behavior can be improved upon with coordination.

The inefficiency of selfish routing (and of Nash equilibria more generally) motivates strategies for *coping with selfishness*—methods for ensuring that noncooperative behavior results in a socially desirable outcome. For selfish routing, an ancient strategy—discussed informally as early as 1920 [8]—is *marginal cost pricing*, also known as congestion, externality, or Pigouvian taxes. The principle of marginal cost pricing asserts that on each edge, each network user on the edge should pay a tax equal to the additional delay its presence causes for the other users on the edge. Several decades later, researchers showed that this principle leads to the following rigorous guarantee [1]: assuming all network users choose routes to minimize the sum of the latency experienced and taxes paid, it is possible to levy a tax on each network edge so that the resulting Nash equilibrium achieves the minimum-possible total latency. Briefly, the inefficiency of selfish routing can always be eradicated by pricing network edges appropriately.

This guarantee, while fundamental, is unsatisfying in several respects. First, it assumes a very strong *homogeneity* property: even though the model assumes a very large number of network users, all users are assumed to trade off time and money in an identical way. How should edges be priced with heterogeneous network users?

Second, the guarantee ignores the *algorithmic* aspect of edge pricing: how can edge prices be efficiently computed? When many different

*This paper surveys some of the results from two forthcoming conference papers [5, 6].

†Department of Computer Science, New York University, 251 Mercer Street, New York, NY 10012. Supported in part by NSF grant CCR0105678. Email: cole@cs.nyu.edu.

‡Department of Computer Science, New York University, 251 Mercer Street, New York, NY 10012. Supported in part by an NSF CAREER Award. Email: dodis@cs.nyu.edu.

§Department of Computer Science, Cornell University, Ithaca, NY 14853. Supported by ONR grant N00014-98-1-0589. Email: timr@cs.cornell.edu.

sets of edge prices induce a minimum-latency Nash equilibrium, can we efficiently compute the "best" one?

Finally, even assuming that all traffic is homogeneous, the principle of marginal cost pricing assumes that (possibly very large) taxes cause no disutility to network users. This assumption is only appropriate when collected taxes can be feasibly returned (directly or indirectly) to the network users, for example by refunding taxes equally to all users (a "lump-sum refund"). This assumption is not always reasonable, for example if refunding the collected taxes to network users is logistically or economically infeasible, or if taxes could represent quantities of a non-monetary, non-refundable good such as time delays. In such settings, when we aim instead to minimize the total user disutility (latency plus taxes paid)—the total *cost*—how should we price the network edges? Intuition may suggest that taxes should never be able to improve the cost of a Nash equilibrium, but the famous *Braess's Paradox* [4] shows this intuition to be incorrect.

1.2 Our Results

In this work, we study how to price the edges of a network in the absence of one or both of these assumptions. For heterogeneous traffic, with different agents trading off time and money in different ways, we prove the following.

- The edges of a single-commodity network can always be priced so that an optimal routing of traffic arises as a Nash equilibrium, even for very general heterogeneous populations of network users.
- When there are only finitely many different types of network users and all edge latency functions are convex, we show how to compute such edge prices efficiently.
- We prove that an easy-to-check mathematical condition on the population of heterogeneous network users is both necessary and sufficient for the existence of edge prices that induce a minimum-latency routing while requiring only moderate taxes.

We also consider the setting of homogeneous traffic with no possibility of refunding taxes. Our goal is then to minimize the total user disutility (latency plus taxes paid)—the total *cost*. We prove the following results.

- Taxes cannot improve the cost of a Nash equilibrium by more than a factor of $\lfloor n/2 \rfloor$, where n is the number of nodes in the network. This upper bound is tight.
- In networks with linear latency functions, taxes cannot improve over removing edges. There are networks with nonlinear latency functions, however, in which taxes are radically more powerful than edge removal.
- Taxes that minimize the cost of the Nash flow cannot be computed by an efficient algorithm. In fact, no polynomial-time heuristic can significantly outperform the trivial heuristic of assigning zero tax to every edge.

On the one hand, our results imply that taxes can be a powerful and useful tool for minimizing latency, even with heterogeneous traffic. On the other, taxes are not useful for minimizing cost with non-refundable taxes, even for homogeneous traffic (assuming only polynomial computation is allowed).

1.3 Selfish Routing and Peer-To-Peer Networks

The model of selfish routing studied here applies in two different ways to routing in peer-to-peer networks. Most obviously, this model is relevant if each machine in the network is routing its traffic on minimum-latency paths. This assumes, however, that each machine has (at least approximate) knowledge of the state of the entire network; this assumption is clearly unrealistic even in networks of moderate size.

More interestingly, the Nash equilibria of the routing game studied in this paper also arise via distributed shortest-path algorithms that are prevalent in commonly-used network protocols,

such as OSPF [7]. Specifically, these Nash equilibria are precisely the fixed points of a shortest-path protocol in which all nodes of the network define the length of their incident edges as their current latency or delay (the shortest paths computed by the protocol are of course a function of how routers define edge length) [3]. Such delay-based routing schemes are in some sense incentive-compatible, since in effect each node is acting to minimize the delay encountered by the traffic that it must route, and in particular by the traffic emanating at that node.

There are, of course, many criticisms that apply to this model of selfish routing. For example, a machine may, instead of routing purely based on delay, give priority to traffic for which it is a source (relative to traffic for which it is an intermediate node). In addition, delay information must be at least approximately correct for a fixed point of the shortest-path protocol to (approximately) correspond to the Nash equilibria studied here. Finally, the theory presented in this paper is only relevant to delay-based routing when the shortest-path protocol is assumed to (at least approximately) converge to a fixed point. Indeed, convergence issues are one the primary reasons that dynamic edge metrics such as delay are uncommon in the Internet. While reasonable sufficient conditions for convergence are known (see e.g. [3]), it is not clear that these conditions hold in practice, especially in networks where the traffic distribution and the network itself are changing rapidly over time. Nevertheless, we feel that the model studied in this paper captures the spirit of delay-based routing, a natural performance-sensitive routing scheme, and is therefore worthy of analysis.

We note that there are many additional problems in implementing a taxation scheme in a peer-to-peer network, and we do not propose any solution here. Rather, in our work we ask the more basic question: is it even worth attempting to implement such a taxation scheme in networks with selfish routing? As we shall see, the answer depends crucially on whether the collected taxes are refundable (e.g. by a lump-sum refund) or are a social loss.

2 The Model

2.1 Congested Networks and Flows

We consider a directed graph $G = (V, E)$ with source s and sink t. We denote the set of simple s-t paths in G by $\mathcal{P}$, which we assume is nonempty. We allow parallel edges but have no use for self-loops. There is one unit of traffic wishing to travel from s and t, modeled as the unit interval $[0, 1]$ endowed with Lebesgue measure λ.[1] Each point $a \in [0, 1]$ will be called an *agent*, and is thought of as a noncooperative and infinitesimal unit of traffic.

By a *flow*, we mean a Lebesgue-measurable function $f : [0, 1] \to \mathcal{P}$ describing who goes where. There are two ways to ignore some of the information provided by a flow to recover more familiar combinatorial objects. A flow naturally induces a *flow on paths*, which we define to be the vector $\{f_P\}_{P \in \mathcal{P}}$ indexed by s-t paths, with $f_P = \lambda(\{a \in [0, 1] : f(a) = P\})$ the amount of traffic assigned to the path P by f. A flow on paths then induces a *flow on edges*, defined as a vector $\{f_e\}_{e \in E}$ on edges with $f_e = \sum_{P: e \in P} f_P$ the amount of traffic using edge e en route from s to t. A flow on edges may correspond to many different flows on paths, and a flow on paths may correspond to many different flows.

The network G suffers from congestion effects; to model this, we assume each edge e possesses a nonnegative, continuous, nondecreasing *latency function* ℓ_e that describes the delay incurred by traffic on e as a function of the edge congestion f_e. The latency of a path P in G with respect to a flow f is then given by $\ell_P(f) = \sum_{e \in P} \ell_e(f_e)$. A common way to measure the quality of a flow is by its *total latency* $L(f)$, defined by $L(f) = \sum_{P \in \mathcal{P}} \ell_P(f) f_P$ or, equivalently, by $L(f) = \sum_{e \in E} \ell_e(f_e) f_e$. Evidently, any two flows inducing the same flow on edges have equal total latency. A minimum-latency flow always exists, for the set of flows on edges is a compact set and $L(\cdot)$ is continuous.

We allow a set of nonnegative *taxes* $\{\tau_e\}_{e \in E}$

[1] Allowing an arbitrary rate $r > 0$ of traffic requires only cosmetic changes to this paper.

to be placed on the edges of a network G, and denote the resulting network by G^τ. We will call a triple (G, ℓ, α) or (G^τ, ℓ, α) an *instance*.

We assume that agent a has a money/time valuation ratio of $\alpha(a)$. Thus, if a set τ of taxes is placed on the edges of a network, agent a seeks a shortest s-t path relative to edge lengths $\ell_e(f_e) + \alpha(a)\tau_e$. We will assume that agents are sorted in order of money-sensitivity, so that $\alpha : [0,1] \to [0,\infty]$ is a nondecreasing function. We call α a *distribution function*. We will not assume that distribution functions are bounded, and therefore permit functions α with $\alpha(1) = +\infty$; however, we will always assume that α is finite on $[0,1)$.

Finally, for a flow f for an instance (G^τ, ℓ, α), we define the *cost* $C(f,\tau)$ of the flow f as the total disutility caused to network users, accounting for disutility due to both latency and taxes:

$$C(f,\tau) = \int_0^1 c^a_{f(a)}(f,\tau)da$$

where $c^a_P(f,\tau) = \ell_P(f) + \alpha(a)\tau_P$ and $\tau_P = \sum_{e\in P} \tau_e$.

The functions $L(\cdot)$ and $C(\cdot,\tau)$ coincide if and only if $\tau = 0$ or $\alpha(a) = 0$ for all agents a.

2.2 Nash equilibria

We assume that noncooperative behavior results in a Nash equilibrium—a "stable point" in which no agent has an incentive to unilaterally alter its strategy (i.e., its route from s to t). To make this precise, again let $c^a_P(f,\tau) = \ell_P(f) + \alpha(a)\tau_P$ denote agent a's evaluation of path P relative to taxes τ and latencies with respect to the flow f.

Definition 2.1 A flow $f : [0,1] \to \mathcal{P}$ is *at Nash equilibrium* or is a *Nash flow* for instance (G^τ, ℓ, α) if for every agent $a \in [0,1]$ and path $P \in \mathcal{P}$,

$$c^a_{f(a)}(f,\tau) \le c^a_P(f,\tau).$$

That a Nash flow exists in every network follows from, for example, the quite general results of Schmeidler [11, Thm 2]. Nash flows are not in general unique but are all equivalent for our purposes, so we will ignore the issue of uniqueness in the sequel (see [6] for a rigorous discussion).

3 Pricing Networks with Selfish Routing

3.1 Pricing Edges to Minimize Delay for Heterogeneous Traffic

In this section we seek taxes that induce the minimum-latency routing as a Nash flow. We will call such taxes *latency-optimal*. Our central theorem for heterogeneous traffic is that as long as all agents are sensitive to taxes, then latency-optimal taxes exist. This theorem relies on Brouwer's fixed-point theorem and is thus non-constructive.

Theorem 3.1 *If instance (G, ℓ, α) satisfies $\alpha(0) > 0$, then it admits a latency-optimal set of taxes.*

Remark 3.2 An interesting open question is whether latency-optimal taxes always exist in a *multicommodity* flow network.

Under mild additional assumptions on the network latency functions and the distribution function, we can use ideas of Bergendorff et al. [2] to find a set of latency-optimal taxes efficiently.[2] In fact, we will show this in a very strong way: we will prove that the set of latency-optimal taxes can be explicitly described by a polynomial-size set of linear inequalities. Thus a latency-optimal tax can not only be efficiently found, but in fact the latency-optimal tax that optimizes some secondary linear objective function, such as minimizing the taxes paid by network users, can also be computed efficiently. This constructive result complements and does not subsume the previous existence theorem, even in the special case of finitely many distinct agent types and convex latency functions; on the contrary, this existence result provides the sole assurance that our linear description of the set of latency-optimal taxes describes a non-empty set!

Theorem 3.3 *Let (G, ℓ, α) be an instance with convex latency functions in which α takes on*

[2] We assume some reasonable encoding of latency and distribution functions.

only finitely many distinct values. Then a linear description of the latency-optimal taxes for (G, ℓ, α) can be computed in polynomial time. In particular, a set of latency-optimal taxes can be computed in polynomial time.

Theorem 3.3 implies that taxes, if implementable, can be a powerful tool for minimizing the total latency of traffic.

3.2 Pricing Edges to Minimize Cost for Homogeneous Traffic

We next state several results about pricing edges to minimize the total cost, the sum of the delay and the taxes paid, for homogeneous traffic (where $\alpha(a) = 1$ for all $a \in [0, 1]$). In what follows, we suppress dependence on the (fixed) distribution function α. We also denote by $C(G^\tau, \ell)$ the cost $C(f^\tau, \tau)$ of a flow f^τ at Nash equilibrium for (G^τ, ℓ).

That taxes can reduce the cost of a Nash flow follows from a famous example of Braess [4], which demonstrates how removing an edge from a network can improve a Nash flow. Since a sufficiently large tax on an edge effectively deletes it from the network, taxes are at least as powerful as removing edges for improving the cost of a Nash flow.

We first study the following question: how much can the cost of a Nash flow improve after levying taxes on the edges? We resolve this question by adapting techniques from [9] that bound the maximum-possible benefit from removing edges from a network.

Theorem 3.4 *Let (G, ℓ) be an instance with n vertices and τ a tax on edges. Let f and f^τ be Nash flows for (G, ℓ) and (G^τ, ℓ), respectively. Then*

$$L(f) \leq \left\lfloor \frac{n}{2} \right\rfloor \cdot C(f^\tau, \tau).$$

Examples from [9] demonstrate that the upper bound of Theorem 3.4 can be attained, even with the weaker operation of edge removal.

We next study if taxes can be more powerful than edge removal for improving the cost of a Nash flow. Formally, we will say that a set τ of taxes for the instance (G, ℓ) is $0/\infty$ if, for some Nash flow f^τ for (G^τ, ℓ), $\tau_e = 0$ or $f_e^\tau = 0$ for each edge e. We note that $0/\infty$ taxes are no more powerful than removing edges from the network, since if τ is $0/\infty$ then $C(G^\tau, \ell) = C(H, \ell)$, where H is the subgraph of G of the edges with zero tax. A tax is *cost-optimal* for an instance (G, ℓ) if it minimizes $C(G^\tau, \ell)$ over all nonnegative tax vectors τ.

We show that taxes are no more powerful than edge removal in a special class of networks, but are far more powerful in general networks.

Theorem 3.5 *(a) An instance with linear latency functions (all of the form $\ell(x) = ax+b$ for $a, b \geq 0$) admits a cost-optimal set of taxes that is $0/\infty$.*

(b) For each integer $n \geq 2$, there is an instance (G, ℓ) with general latency functions with $C(H, \ell) = \lfloor n/2 \rfloor$ for all subgraphs H of G but $C(G^\tau, \ell) = 1$ for some tax $\tau \geq 0$.

The question now arises: since cost-optimal taxes can be so powerful, can we compute them efficiently? Our next result, together with Theorem 3.4, answers this question negatively in a strong way: no polynomial-time heuristic can significantly outperform the trivial heuristic of assigning zero tax to every edge.

Theorem 3.6 *Unless $P = NP$, for every $\epsilon > 0$ there is no $O(n^{1-\epsilon})$-approximation algorithm for the problem of computing the cost-optimal tax in n-node networks with arbitrary latency functions.*

This computational complexity inherent in cost-optimal taxes obviously casts doubt on their potential for reducing cost in networks with selfish routing and unrefundable taxes.

3.3 Pricing Edges to Minimize Cost for Heterogeneous Traffic

In this section, we study the cost of latency-optimal taxes for heterogeneous traffic. Our contribution is a complete characterization of the distribution functions α for which the disutility

due to latency-optimal taxes is always at most a constant factor times the disutility due to latency. This is an extremely strong guarantee, and there is no reason a priori to believe that *any* distribution function has this property.

We first formalize the property we desire of a distribution function.

Definition 3.7 A distribution function α is *ρ-cheap* with parameter $\rho \geq 1$ if the following property holds: for every instance (G, ℓ, α) with $\alpha(0) > 0$, there is a set τ of latency-optimal taxes and a minimum-latency flow f^τ at Nash equilibrium for (G^τ, ℓ, α) such that

$$C(f^\tau, \tau) \leq \rho \cdot L(f^\tau).$$

A distribution function is *cheap* if it is ρ-cheap for some finite $\rho \geq 1$.

We now state our characterization, employing the notation $\alpha(z^-)$ to mean the left limit $\lim_{z_n \uparrow z} \alpha(z_n)$ of a distribution function α at a point z.

Theorem 3.8 *A distribution function α with $\alpha(0) > 0$ is ρ-cheap if and only if*

$$\int_0^z \alpha(a)\, da \leq (\rho - 1) \cdot \alpha(z^-)[1 - z]$$

for all $z \in (0, 1)$.

Remark 3.9 The condition of Theorem 3.8 is quite strong and is not satisfied by most distribution functions—the simplest distribution functions satisfying it for some value of ρ are the functions $\alpha(a) = (1 - a)^{-k}$ for $k > 1$—and thus latency-optimal taxes are in general quite costly. Nonetheless, we find it surprising that *any* natural distribution function is cheap, and satisfying that cheap distribution functions admit such a crisp mathematical characterization.

References

[1] M. Beckmann, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation.* Yale University Press, 1956.

[2] P. Bergendorff, D. W. Hearn, and M. V. Ramana. Congestion toll pricing of traffic networks. In P. M. Pardalos, D. W. Hearn, and W. W. Hager, editors, *Network Optimization*, pages 51–71. Springer-Verlag, 1997.

[3] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation.* Prentice Hall, 1989.

[4] D. Braess. Uber ein paradoxon der verkehrsplanung. *Unternehmensforschung*, 12:258–268, 1968.

[5] R. Cole, Y. Dodis, and T. Roughgarden. How much can taxes help selfish routing? In *Proceedings of the Fourth Annual ACM Conference on Electronic Commerce*, 2003.

[6] R. Cole, Y. Dodis, and T. Roughgarden. Pricing network edges for heterogeneous selfish users. In *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing*, 2003.

[7] S. Keshav. *An Engineering Approach to Computer Networking.* Addison-Wesley, 1997.

[8] A. C. Pigou. *The economics of welfare.* Macmillan, 1920.

[9] T. Roughgarden. Designing networks for selfish users is hard. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science*, pages 472–481, 2001. Full version available from `http://www.cs.cornell.edu/timr`.

[10] T. Roughgarden and É. Tardos. How bad is selfish routing? *Journal of the ACM*, 49(2):236–259, 2002. Preliminary version in *FOCS '00.*

[11] D. Schmeidler. Equilibrium points of nonatomic games. *Journal of Statistical Physics*, 7(4):295–300, 1973.

[12] J. G. Wardrop. Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers, Pt. II*, volume 1, pages 325–378, 1952.