How Unfair is Optimal Routing?

Tim Roughgarden*

1 Introduction

We are given a network and a rate of traffic between a source node and a destination node, and seek an assignment of traffic to source-destination paths. We assume that each network user controls a negligible fraction of the overall traffic, so that feasible assignments of traffic to paths in the network can be modeled as *network flows*. We also assume that the time needed to traverse a single link of the network is *load-dependent*, that is, the common latency suffered by all traffic on the link increases as the link becomes more congested.

We consider two types of traffic assignments. In the first, we measure the quality of an assignment by the *total latency* incurred by network users; an *optimal assignment* is a feasible assignment that minimizes the total latency. On the other hand, it is often difficult in practice to impose optimal routing strategies on the traffic in a network, leaving network users free to act according to their own interests. We assume that, in the absence of network regulation, users act in a selfish manner. Under this assumption, we can expect network traffic to converge to the second type of assignment that we consider, an assignment *at Nash equilibrium*. An assignment is at Nash equilibrium if no network user has an incentive to switch paths; this occurs when all traffic travels on minimum-latency paths.

The following question motivates our work: *is the optimal assignment really a "better" assignment than an assignment at Nash equilibrium?* While the optimal assignment obviously dominates one at Nash equilibrium from the viewpoint of total latency, it may lack desirable *fairness* properties. For example, consider a network consisting of two nodes, s and t, and two edges, e_1 and e_2, from s to t. Suppose further that one unit of traffic wishes to travel from s to t, that the latency of edge e_1 is always $2(1-\epsilon)$ (independent of the edge congestion, where $\epsilon > 0$ is a very small number), and that the latency of edge e_2 is the same as the edge congestion (i.e., if x units of traffic are on edge e_2, then all of this flow incurs x units of latency). In the assignment at Nash equilibrium, all traffic is on the second link; in the minimum-latency assignment, $1-\epsilon$ units of traffic use edge e_2 while the remaining ϵ units of traffic use edge e_1. Roughly, a small fraction of the traffic is sacrificed to the slower edge because it improves the overall social welfare (by reducing the congestion experienced by the overwhelming majority of network users); needless to say, these martyrs may not appreciate a doubling of their travel time in the name of "the greater good"! Indeed, this drawback of routing traffic optimally has inspired practitioners to find traffic assignments that minimize total latency subject to explicit *length constraints* [1], which require that no network user experiences much more latency than in an assignment at Nash equilibrium. The central question of this paper is *how much worse off can network users be in an optimal assignment than in one at Nash equilibrium?* After reviewing some technical preliminaries in the next section (all of which are classical; see [2] for historical references), we provide an exact solution to this problem under weak hypotheses on the class of allowable latency functions.

*Department of Computer Science, Cornell University, Ithaca NY 14853. Supported by an NSF Graduate Fellowship, a Cornell University Fellowship, and ONR grant N00014-98-1-0589. Research performed in part while visiting IBM Almaden. Email: `timr@cs.cornell.edu`.

2 Preliminaries

The Model. We consider a directed network $G = (V, E)$ with vertex set V, edge set E, source vertex s, and destination vertex t. We denote the set of s-t paths by $\mathcal{P}$. A *flow* is a function $f : \mathcal{P} \to \mathcal{R}^+$; for a fixed flow f we define $f_e = \sum_{P:e\in P} f_P$. With respect to a finite and positive *traffic rate* r, a flow f is said to be *feasible* if $\sum_{P\in\mathcal{P}} f_P = r$. Each edge $e \in E$ is given a load-dependent *latency function* that we denote by $\ell_e(\cdot)$. We assume that ℓ_e is nonnegative, continuous, and nondecreasing. The latency of a path P with respect to a flow f is then the sum of the latencies of the edges in the path, denoted by $\ell_P(f) = \sum_{e\in P} \ell_e(f_e)$. We call the triple (G, r, ℓ) an *instance*.

We define the *cost* $C(f)$ of a flow f in G as the total latency incurred by f, i.e., $C(f) = \sum_{P\in\mathcal{P}} \ell_P(f) f_P$. With respect to instance (G, r, ℓ), a feasible flow minimizing $C(f)$ is said to be *optimal* or *minimum-latency*.

Flows at Nash Equilibrium. A flow f feasible for (G, r, ℓ) is said to be *at Nash equilibrium* (or is a *Nash* *flow*) if for every two s-t paths $P_1, P_2 \in \mathcal{P}$ with $f_{P_1} > 0$, $\ell_{P_1}(f) \leq \ell_{P_2}(f)$. In particular, if a flow f is at Nash equilibrium then all s-t flow paths have equal latency. It is well known that Nash flows always exist and are essentially unique.

Optimal Flows. Assuming mild extra conditions on the latency functions of an instance, there is a well-known characterization of optimal flows that mirrors the definition of Nash flows. Let (G, r, ℓ) have the property that, for each edge e, the function $x \cdot \ell_e(x)$ is convex. Define the *marginal cost function* $\hat{\ell}_e$ by $\hat{\ell}_e = \frac{d}{dx}(x \cdot \ell_e(x))$. Then, a flow $\hat{f}$ feasible for (G, r, ℓ) is optimal if and only if it is at Nash equilibrium for $(G, r, \hat{\ell})$.

3 Our Results

We saw in Section 1 that some traffic in a minimum-latency flow may be routed on paths with larger latency than that incurred by all traffic in a Nash flow; our goal is to quantify this phenomenon. Define the *unfairness* of instance (G, r, ℓ) as the maximum ratio between the latency of a flow path of an optimal flow for (G, r, ℓ) and that of a flow path of a Nash flow for (G, r, ℓ). We denote the unfairness of (G, r, ℓ) by $u(G, r, \ell)$. Our first observation is that $u(G, r, \ell)$ can be arbitrarily large if we do not place additional restrictions on the class of allowable latency functions. To see this, modify the example of Section 1 as follows: for any positive integer p, define the latency of the first edge as the constant function $\ell(x) = (p+1)(1-\epsilon)$ and that of the second edge as $\ell(x) = x^p$. In this example, $u(G, r, \ell) = (p+1)(1-\epsilon)$.

Thus, we aim to quantify the worst possible unfairness as a function of the class of allowable latency functions. Toward this end, let $\mathcal{L}$ denote a class of allowable latency functions (that are continuous and nondecreasing), with the additional property that for each $\ell \in \mathcal{L}$, the function $x \cdot \ell(x)$ is convex. For $\ell \in \mathcal{L}$, define $\hat{\ell}$ as in the previous section. For $\ell \in \mathcal{L}$, define $\gamma(\ell)$ by $\gamma(\ell) = \sup_{x>0} \hat{\ell}(x)/\ell(x)$. Recalling the characterizations of Nash and optimal flows from the previous section, we may interpret $\gamma(\ell)$ as the biggest discrepancy between how optimal and Nash flows evaluate the per-unit cost of using an edge (via the "socially aware" or "conscientious" marginal cost function $\hat{\ell}$ and the "selfish" latency function ℓ, respectively). Define $\gamma(\mathcal{L})$ by $\gamma(\mathcal{L}) = \sup_{\ell \in \mathcal{L}} \gamma(\ell)$. Then, we have the following result.

THEOREM 3.1. *If (G, r, ℓ) is an instance with latency functions drawn from $\mathcal{L}$, then $u(G, r, \ell) \leq \gamma(\mathcal{L})$.*

Proof. Let (G, r, ℓ) be such an instance, admitting Nash flow f and optimal flow $\hat{f}$. We need to show that the maximum latency of a flow path of $\hat{f}$ is at most $\gamma \equiv \gamma(\mathcal{L})$ times the latency of a flow path of f.

Suppose for contradiction that P_1, P_2 are paths satisfying $f_{P_1} > 0$, $\hat{f}_{P_2} > 0$, and $\ell_{P_2}(\hat{f}) > \gamma \cdot \ell_{P_1}(f)$. First, we introduce some notation. Since f is at Nash equilibrium for (G, r, ℓ), there is a value L such that $\ell_P(f) = L$ whenever $f_P > 0$ (i.e., all flow paths of f have a common latency with respect to latency functions ℓ). Similarly, there is a value $\hat{L}$ such that all flows paths of $\hat{f}$ have latency $\hat{L}$ with respect to latency functions $\hat{\ell}$.

Now, as every latency function is nondecreasing, we have $\ell_e(x) \leq \hat{\ell}_e(x)$ for all e and x. Thus, we may derive

$$L = \ell_{P_1}(f) < \frac{1}{\gamma} \ell_{P_2}(\hat{f}) \leq \frac{1}{\gamma} \hat{\ell}_{P_2}(\hat{f}) = \frac{\hat{L}}{\gamma}.$$

We next note that the cost of the flow f is $C(f) = rL$. The cost of the optimal flow $\hat{f}$ is not so easy to compute (as flow paths have equal latency with respect to functions $\hat{\ell}$ but not with respect to ℓ). However, since every latency function is drawn from $\mathcal{L}$, we obtain $\hat{\ell}_P(\hat{f}) \leq \gamma \cdot \ell_P(\hat{f})$ for every flow path P of $\hat{f}$ and hence

$$C(\hat{f}) \geq \frac{1}{\gamma} \sum_{P \in \mathcal{P}} \hat{\ell}_P(\hat{f}) \hat{f}_P = \frac{1}{\gamma} r \hat{L} > rL = C(f),$$

which contradicts the optimality of $\hat{f}$.

For example, an instance whose latency functions are polynomials with nonnegative coefficients of degree at most p has unfairness at most $p+1$. A simple variation on the previous example shows that the theorem is best possible in the following sense: for any real number $c \geq 1$, there is a class $\mathcal{L}$ of latency functions (namely, the constant functions along with x^{c-1}) satisfying $\gamma(\mathcal{L}) \leq c$ such that there are instances with latency functions from $\mathcal{L}$ with unfairness arbitrarily close to c. In fact, this style of argument shows the following stronger statement: if $\mathcal{L}$ is a class of latency functions that includes the constant functions, then $\sup_{(G,r,\ell)} u(G, r, \ell)$ (where the supremum ranges over instances with latency function in $\mathcal{L}$) is precisely $\gamma(\mathcal{L})$ (possibly $+\infty$). The hypothesis that $\mathcal{L}$ includes the constant functions is necessary; indeed, there are classes of latency functions with arbitrarily large values of γ (e.g., latency functions of the form ax^p where p is a fixed positive integer and $a \geq 0$) with respect to which all instances have unfairness 1 (see [2]).

References

[1] O. Jahn, R. Möhring, A. S. Schulz. Optimal Routing of Traffic Flows with Length Restrictions in Networks with Congestion. *Operations Research Proceedings 1999*, pages 437–442.

[2] T. Roughgarden and É. Tardos. How Bad is Selfish Routing? *Proc. 41st FOCS*, pages 93–102.