Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
ISTTT 00 (2012) 1–19
Procedia
Social and Behavioral Sciences
www.elsevier.com/locate/procedia

Optimization-based Framework for Rerouting a Subset of Users with Mixed Lagrangian-Eulerian Demand

Jack Reilly, Walid Krichene, Saurabh Amin, Alexandre M. Bayen

Abstract

Socially optimal routing strategies are often developed with the assumption that all demand on the network can be cooperative. When only partial cooperation is possible, new difficulties arise: origin-destination demand data is difficult to obtain for noncooperative demand, and noncooperative demand can change routes to respond to the cooperative demand's routing decisions if there are significant changes in route travel times. The present article develops a framework for providing routing strategies with partial cooperation of demand on a flow network, which is based on convex optimization and addresses the above difficulties.

In what we refer to as Lagrangian-Eulerian demand modeling, we assume that, while origin-destination demands are known for the cooperative flow (Lagrangian information), only link-level flows are known for the total flow on the network (Eulerian information). This setting is compatible with the use of preexisting stationary flow sensors to measure the "baseline" flow of the network. One only needs origin-destination information on the limited set of cooperative demand. While the non-uniqueness of origin-destination demand imputation from link flows does not allow us to consider noncooperative flow rerouting, we develop an alternative approach to account for noncooperative response. We model the noncooperative flow to have bounded tolerance, which assumes that noncooperative flow has no desire to seek alternative routes if the nominal (original) route's resulting latency is within a given bound of the nominal latency. We then show that under these constraints, total latency on the network can be minimized by solving a convex optimization problem. The optimization program formulation is used to find optimal rerouting strategies on networks with horizontal queues (e.g. traffic networks) with extensions to vertical queues (e.g. communications networks). The framework is applied to a multiple-destination network with horizontal queues, which demonstrates the generality of the method and how the tolerance model parameters affect the optimal re-routing solution. An optimization approach is presented that preconditions historical input data to be compatible with the presented models.

Keywords: traffic assignment, convex optimization, social optimum, Lagrangian flow, Eulerian flow

Nomenclature

$\mathcal{L}$	Set of links.
$\mathcal{O} \subset \mathcal{L}$	Set of origins (sources).
$\mathcal{D} \subset \mathcal{L}$	Set of destinations (sinks).
$\mathcal{J}$	Set of junctions.
$\mathcal{R}$	Set of routes.
$r \in \mathcal{R}$	Sequence of contiguous links $(r_1, \ldots, r_{\lvert r \rvert}) : r_i \in \mathcal{L}$
$\Gamma_j \subset \mathcal{L}$	Set of incoming links for junction $j \in \mathcal{J}$.
$\Gamma_j^{-1} \subset \mathcal{L}$	Set of outgoing links for junction $j \in \mathcal{J}$.
$\mathbf{r}_j \subset \mathcal{R}$	Set of routes passing through junction $j \in \mathcal{J}$.

$f_l, \bar{f}_l$	Flow (resp. nominal flow) on link $l \in \mathcal{L}$.
f_r	Total flow on route $r \in \mathcal{R}$
f_r^{c}	Cooperative flow on route $r \in \mathcal{R}$.
$\rho_l, \bar{\rho}_l$	Density (resp. nominal density) on link $l \in \mathcal{L}$
$\mathbf{f}^{\mathrm{c}}$	Assignment of cooperative flows across all routes $\in \mathcal{R}$.
$\bar{f}_l^{\mathrm{nc}}$	Nominal noncooperative flow on link $l \in \mathcal{L}$.
$\bar{f}_{o,d}$	OD flow demand of cooperative (Lagrangian) users from origin $o \in \mathcal{O}$ to destination $d \in \mathcal{D}$.
$\ell_l, \bar{\ell}_l$	Latency (resp. nominal latency) on link $l \in \mathcal{L}$.
α	Tolerance scale factor.

1. Introduction

1.1. Traffic assignment: selfish routing vs. social routing

The problem of traffic assignment handles users' route and departure time decisions and how individual behaviors impact the performance of the underlying traffic network. If all user decide in a self-optimizing manner, then the resulting network state is a *user equilibrium (e.g. Wardrop 1952).* If every user acts in a manner that is beneficial to societal goals, it is said to be a *system* or *social optimum.* Socially optimal schemes are studied under the assumption that a central agency controls *all* the users, while on the other extreme, user equilibrium is is a good model to describe selfish behavior *in the absence* of a central agency. A complete characterization user equilibrium model requires complete information of the origin-destination demands on the network. This information is often too expensive to obtain. Specifically, origin-destination information may only be available for a fraction of users because collecting such information requires participation/consent of the travelers and technological capability of the central agency. Lo & Szeto 2002 give a variational inequality approach to solving user equilibrium, while Papageorgiou 1990 presents an optimal control framework, and both methods require full information of origin-destination demands on the network.

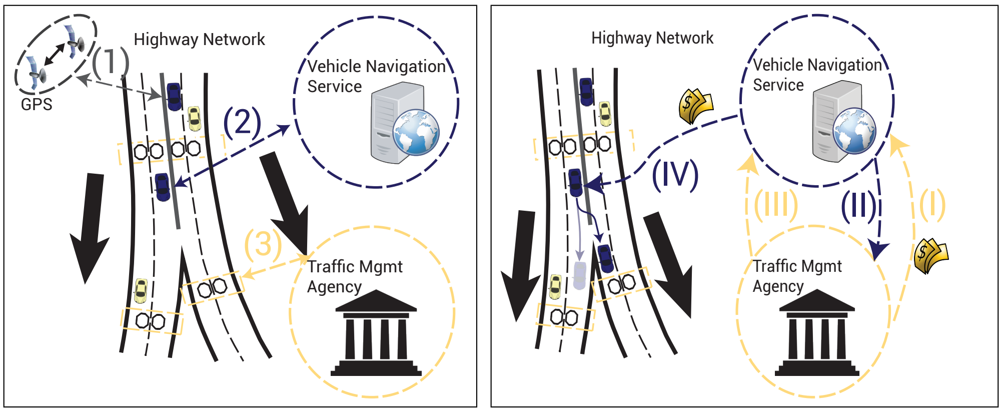
These technologies can be broadly categorized into two categories. First, there are recommendation systems, such as variable message signs that suggest particular routes based on estimated travel times or general dissemination of information to better inform users of network conditions. Second, there are direct control systems that restrict behavior of users via ramp metering or detours. Gomes *et al.* 2008 discusses the effectiveness of ramp metering as a means of achieving a social optimum. These direct control mechanisms are generally applied at a specific point and time and do not distinguish between users who have different routes or destinations. The effectiveness of such active control schemes usually depends on complete origin-destination demands. An exception is that boundary flow demands may be sufficient for evacuation-type problems (e.g. Ziliaskopoulos 2000).

1.2. Using mobile phones to control routes of individual users

With the emergence of GPS-enabled cell phones and their widespread adoption in populated city areas, a third category of control has become possible: one that communicates directly with users and permits a central agency or a private entity to engage individual users to shift their travel choices. Such a high granularity of control would allow specific origin-destinations or routes to be targeted by the control scheme and could even be customized to the route preferences of the individual users.

Vehicle navigation services that collect, aggregate, and process information from a large number of GPS-equipped mobile devices have become increasingly popular. Such services include Waze, Google Maps, and other such mobile applications. While these services are popular for their utility to individual drivers, the service providers are also able to collect information on behavior of the fraction of users that are equipped with these devices. Once the data has been anonymized to protect the privacy of individual users, the origin-destination information could be interpreted as a subset of the total demand on a network. Additionally, route guidance decisions could be made to benefit their user-base as a whole, rather than on an individual level. Papadimitriou 2001 discusses the inherent inefficiencies of selfish routing versus the social optimum.

Individually-applied control schemes have many advantages, but a limitation is that the user-base of a particular vehicle navigation service would only constitute a subset of the total users of the network. A



(a) (b)

Figure 1.1: Route guidance system architecture for using both route-based and link-level flow information. **(a)** An illustration of Lagrangian and Eulerian traffic information collection systems. **(1)** Path-based information is collected via GPS-equipped vehicles, from either onboard route-guidance systems or cell phones. **(2)** Route-based information is sent back to a vehicle navigation service that aggregates traffic information from many GPS-equipped users. **(3)** Eulerian-based loop detectors collect flow counts and send the information to a traffic management agency. **(b)** An illustration of proposed interaction the traffic management agency, vehicle navigation services, and network users. **(I)** Contracting of the vehicle navigation services by the traffic management agency. This may involve monetary compensation or tolling. **(II)** Anonymized Lagrangian information (owned by the vehicle navigation agency) is transferred to the traffic management agency. **(III)** The traffic management service provides route guidance to the vehicle navigation service to improve overall traffic conditions. **(IV)** The vehicle navigation service provides individual network users with alternate route suggestions, with potential incentivization. Users may be guided to switch from their previously preferred (nominal) route.

significant number of users of the network may not have access to or prefer not to use a GPS-enabled device. Also, a complete understanding of the origin-destination demands on a network by a single entity would still be difficult or very expensive to obtain.

1.3. Combining route-based demands with link-level flow information

While collecting route information on individual users suffers from limited penetration, existing traffic monitoring systems, such as loop detectors or cameras, are able to capture all vehicle flows for particular locations on networks. These stationary systems are often monitored by public, traffic management agencies, that are interested in the welfare of all users on the network. It is apparent that the two methods for capturing traffic information are complementary: GPS-based methods have limited penetration but more detailed origin-destination information, while stationary sensors have full penetration of flow, but cannot give route-level information on the demands. Figure 1.1a depicts the vehicle navigation service collecting aggregate GPS data from their "dark" users on the network **(1)**, while the traffic management agency collects flow count data, accounting for *all* users ("dark" and "light") **(2)**, from the loop detectors embedded in the road. Collectively, the route-based flow data could be used as a source of *re-routable* traffic flow, while the link-based flow data could be used to better estimate the expected travel times that all users will experience before and after re-routing a subset of users. This article proposes a method for using both information types to improve traffic conditions across the network. For now, we motivate our work with a scenario in which such a technique could be employed.

Figure 1.1b depicts the scenario in which a traffic management agency partners with a vehicle navigation service to take advantage of their different information sources. Initially, a contractual phase may take place **(I)**, where the public agency compensates the vehicle navigation service for access to their data **(II)**. The information from the vehicle navigation service would be aggregated and anonymized, in order to protect the privacy of the individual users of the service. Then, the traffic management agency would input the route-based demand data and the stationary, link-based loop detector data into the algorithm we have developed (Sections 2 and 3). The algorithm outputs a new set of aggregated routing suggestions, which are then sent to the vehicle navigation service **(III)**. Finally, the service relays the routing suggestions to their users **(IV)**.

For this last point of communication between the agency and the user, there are two notes. First, it is likely that a fraction of users will be suggested routes with larger travel times than to which they have become accustomed. In order to incentivize the user to accept the route suggestion, the traffic management agency may require the vehicle navigation service to compensate these users, enforceable by the initial contractual agreement between the two organizations. Merugu *et al.* 2009 describes an experiment which utilized incentives to move commuters' departure times to less congested times. Alternatively, one can consider that socially optimized routing policies may decrease the travel time for all users *on average*. Then, if all users get assigned desirable routes some days and less desirable routes other days (in order to reduce congestion on desirable routes), then every user could expect to have an improved average travel time. Such an argument could potentially remove the need for monetary compensation or other types of incentivizes. The second note is that we have assumed, given enough incentive from the vehicle navigation service, a user will *always* comply with the suggested route. We do not discuss the method of incentives in this article, but note that the assumption can be relaxed by limiting the amount of re-routable flow.

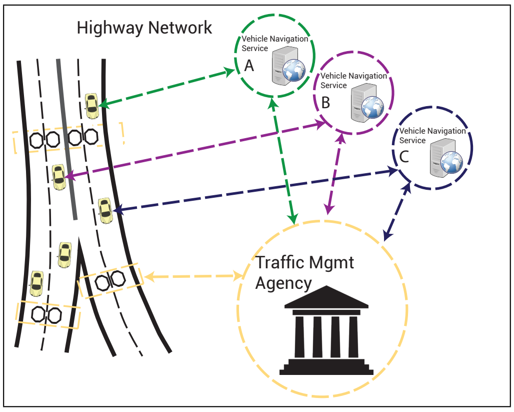


Figure 1.2: Scenario with multiple vehicle navigation services. Individual service's data is aggregated by the central agency and not shared with other agencies. The routing strategies are calculated by the traffic management agency, and the suggested routes are partitioned between the different user groups.

Due to the decoupled nature of the system described in Figure 1.1, we can generalize the scenario to include multiple vehicle navigation services (Figure 1.2). Without sharing information between services, more route-based flow information can be used as input into the algorithm, thus providing more complete information on the origin-destination preferences of the users and collecting a larger pool of re-routable users, while maintaining the privacy of the services which wish to provide socially optimal routes to users.

1.4. Accounting for untracked users' response

There are a number of reasons why a user would participate in the socially optimized routing guidance program described above. As already stated, they could be incentivized through monetary compensation. They could also simply be altruistic, and willing to sacrifice personal optimality for the greater good. What is unknown is the behavioral aspects of those users of the network whose route cannot be tracked. How can one predict the response of these untracked (which we refer to as noncooperative later) network users to the routing schemes being implemented for the tracked users?

A standard approach, described as a Stackelberg game (e.g. Roughgarden 2001; Krichene *et al.* 2012a; Aswani & Tomlin 2011), assumes that the users outside the control of the central agency will respond with a user equilibrium assignment. Since the origin-destination demands of the untracked users is unknown, solving for a user equilibrium is not possible.

In order to address this lack of information on preferences of untracked users, we develop an alternative model of behavior. Related to the concept of bounded rationality in Guo & Liu (2011); Hu & Mahmassani (1997), we assume that the untracked users lack the full information of the state of the network, and cannot make fully rational decisions on their optimal route. Alternatively, the untracked users could possess some *inertia* towards switching routes, and will be content with their previously chosen (nominal) routes, as long as the experienced travel time on the route does not change "too much". This concept of inertia can be practically motivated by considering that some users may appreciate the scenic beauty of a particular suboptimal route, or others have a favorite café along another route. Thus, in order to reasonably assume that the untracked users will not switch their routes, the routing suggestions provided by the algorithm are guaranteed to not significantly deteriorate the quality of existing routes, beyond an a priori specified bound.

A bounded rationality argument in the context of drivers' route selections was made in (Hu & Mahmassani, 1997), where drivers only seek utility gains outside of a certain threshold. Guo & Liu 2011 give some empirical evidence of bounded rationality on road networks. Our model differs from these because our model lacks origin-destination information on the noncooperative users, and to make this distinction, we refer to our model as *bounded tolerance* model.

1.5. Contributions and overview

There is relatively little work done on how partial control schemes can be practically implemented on flow networks. Additionally, inconsistent estimations of traffic between GPS-based data and link-level data can complicate the analysis of the problem. In this article, we present a single methodology for accommodating both origin-destination based and link-level flow information for a general, multi-origin, multi-destination, static network (parameters are unchanging with time), while guaranteeing that the two sources of data are consistent with mass balance across junctions (Section 2). Furthermore, we present a behavioral model on the untracked users based on the concept of bounded rationality (Section 3). This bounded rationality model permits one to cope without origin-destination demands for all users on the network, while still addressing the behavioral aspects of self-routing users.

As our main contribution, we demonstrate how the models presented in this article lead to an elegant, optimization-based solution to the socially optimal routing strategy problem (Section 4). The optimization problem is proven to be convex for a specific instance of horizontal queues that model highway traffic and extended to a general class of vertical queues . As a corollary, we show that for the discretized LWR network model, the social optimum can be solved exactly for both the purely Eulerian flow and the purely Lagrangian flow cases.

The generality of our method is given by applying the framework to a multiple-destination network with horizontal queues and investigating how changes in the tolerance model impact the routing advice (Section 5). The article finishes with a conclusion and discussion of the practical importance of the framework and models developed here-within (Section 6).

2. Modeling partial cooperation with Lagrangian-Eulerian demands

We present the general setting of the routing problem considered, as summarized in Figure 2.1. Consider a setting in which a subset of users are equipped with GPS-enabled devices and are connected to a central coordinator through a *Routing interface* (e.g. a mobile phone application). We refer to this subset as *cooperative* users. First, the cooperative users provide their desired routes to the coordinator through the routing interface. This allows the coordinator to have individual route information, i.e. *Lagrangian information* for the cooperative users. Second, the loop-detectors (or other sensors capable of measuring aggregate link-level flows) provide Eulerian information. We refer to the historical estimates of Lagrangian and Eulerian information as the *nominal* state of the network.

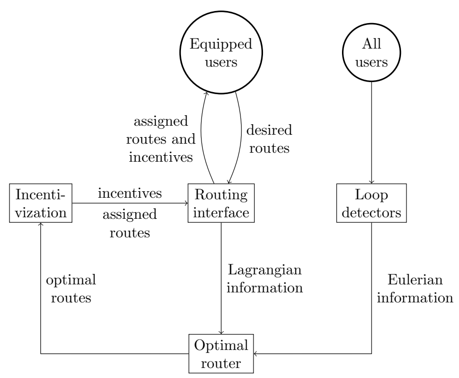


Figure 2.1: Data-flow diagram

Given the nominal Eulerian flow measurements for the entire network and the nominal Lagrangian information for the equipped vehicles, the central coordinator determines the optimal route assignment for the equipped vehicles (Section 2.3). This optimization problem is represented by the *optimal router* block. Since only the cooperative users follow the optimal route assignments provided by the central coordinator, we will refer to this problem as a *partial cooperation* problem.

The next step is an *incentivization* step: given the target optimal routes, and possibly additional constraints (such as a total available budget) a second problem (not discussed in this article) determines an incentive for each equipped vehicle and the corresponding target route. The incentivization problem is outside of the scope of the present article. More information on how to solve incentivization and traffic demand management can be found in Leblanc & Walker (2012).The assigned routes and the corresponding incentives are then offered to the equipped drivers, who can either accept or refuse the offer. The subset of vehicles that do accept the offer (thus taking the route assigned by the central coordinator) are called *cooperating vehicles*. In the present article, we focus our attention on the optimal route assignment with information on mixed Lagrangian-Eulerian demands.

Considering the route optimization goals stated above, we give a declaration of the problem statement to direct the model development of the proceeding sections.

Problem statement: Find a mathematical framework for flow networks which can encompass:

- Two different types of demand information: Lagrangian information, which is specified by the route traversed by the flow, and Eulerian information, which is specified by the flow-count across a link.
- Socially optimal routing strategies which can encompass both information types, given their limitations:
  - Lagrangian information is only known for the cooperative flow, which can be rerouted from its nominal route to improve network conditions.
  - Only Eulerian information is known for the noncooperative flow, which is assumed to maintain its nominal state.

2.1. Network model

Using standard network notation, the network model is defined by the tuple, $(\mathcal{L}, \mathcal{J})$, where $\mathcal{L}$ is the set of links, and $\mathcal{J}$ is the set of junctions. A junction $j \in \mathcal{J}$ has a set of incoming links $\Gamma_j \subseteq \mathcal{L}$ and outgoing links $\Gamma_j^{-1} \subseteq \mathcal{L}$. An origin $o \in \mathcal{O} \subseteq \mathcal{L}$ is a link with no upstream junction. A destination $d \in \mathcal{D} \subseteq \mathcal{L}$ is a link with no downstream junction. A route $r = (r_1, \ldots, r_{|r|}) \in \mathcal{R}$ is a set of adjacent links where $r_1 \in \mathcal{O}$, $r_{|r|} \in \mathcal{D}$, and $\forall i \in (1, \ldots, |r| - 1)$, $\exists j_i^r : r_i \in \Gamma_{j_i^r}, r_{i+1} \in \Gamma_{j_i^r}^{-1}$.

2.2. Cooperative demand vs. total demand

Let the network in Section 2.1 contain flows f_l on every link $l \in \mathcal{L}$. Furthermore, we assume that the network is in steady state, i.e. all state on the network is stationary with respect to time (e.g. flows). We further differentiate two types of demands: Lagrangian and Eulerian.

- We assume that the cooperative users have provided their desired origin and destination. Therefore, for every origin-destination pair $(o, d) \in \mathcal{O} \times \mathcal{D}$, there is a nominal flow demand $\bar{f}_{o,d}$ from the cooperative users, where the bar notation refers to nominal state values. Since this type of demand concerns the routes taken by the flow, we describe this type of demand as *Lagrangian* demand.

- For the noncooperative users, i.e., the users who do not (or choose not to) interact with the routing interface, we do not assume knowledge of Lagrangian demand. Thus, we assume that the only the aggregate link-level flows are available via loop detectors. This aggregate level information does not include OD and route information, and is therefore defined as *Eulerian* demand.

To recover the nominal Eulerian demand of the noncooperative vehicles, we further assume that the nominally used routes of the cooperative vehicles are known. For each link $l \in \mathcal{L}$, we specify a nominal total link flow $\bar{f}_l$, and for each route $r \in \mathcal{R}$, we can specify a nominal route flow for cooperative vehicles, $\bar{f}_r^{\mathrm{c}}$. Then, the nominal noncooperative Eulerian demand, $\bar{f}_l^{\mathrm{nc}}$, is obtained for each link $l \in \mathcal{L}$ by subtracting cooperative flow from the total link flow:

$$\bar{f}_l^{\mathrm{nc}} = \bar{f}_l - \sum_{r|l\in r} \bar{f}_r^{\mathrm{c}} \tag{2.1}$$

For the remainder of the article we use the noncooperative link flows ($\bar{f}_l^{\mathrm{nc}}, l \in \mathcal{L}$) as the input data for nominal flow, but it is understood that this data is derived from the more practically measurable *total* nominal flow values ($\bar{f}_l, l \in \mathcal{L}$) and the cooperative nominal route flows ($\bar{f}_r^{\mathrm{c}}, r \in \mathcal{R}$) via Equation (2.1).

Since we have subtracted off the flow of nominal cooperative flow to obtain the noncooperative flow, we study properties of the network flow when the rerouted cooperative flow is added back into the network. We introduce the decision variable: f_r^{c}, the amount of cooperative flow assigned to route $r \in \mathcal{R}$. To enforce that the entire flow across a link is accounted for and same origin-destination demands of the cooperative users are satisfied, we have the following constraints:

$$\sum_{r|o,d\in r} f_r^{\mathrm{c}} = \bar{f}_{o,d} \qquad \forall o \in \mathcal{O}, d \in \mathcal{D} \tag{2.2}$$

$$f_l = \sum_{r|l\in r} f_r^{\mathrm{c}} + \bar{f}_l^{\mathrm{nc}} \qquad \forall l \in \mathcal{L} \tag{2.3}$$

where $\bar{f}_{o,d} = \sum_{r|o,d\in r} \bar{f}_r^{\mathrm{c}}$ is cooperative flow between origin o and destination d.

A requirement of the Eulerian flow is that noncooperative flow must be conserved across junctions. If the flow across a link $l \in \mathcal{L}$ is f_l, then the following must hold:

$$\sum_{l\in\Gamma_j} f_l = \sum_{l\in\Gamma_j^{-1}} f_l \qquad \forall j \in \mathcal{J} \tag{2.4}$$

Since we partitioned flow on each link into two classes (cooperative and noncooperative), flow conservation must hold across both classes independently. We will see shortly that flow conservation across the cooperative class will be guaranteed by the condition that all cooperative flow must be assigned to a route. Then, for the noncooperative class of nominal flow, we must always have the condition that flow conservation holds across junctions. Since this is a condition on the input to the problem we only state it here once and assume the condition for the rest of the article.

Model Consistency Condition: For every junction $j \in \mathcal{J}$, we assume: $\sum_{l\in\Gamma_j} \bar{f}_l^{\mathrm{nc}} = \sum_{l\in\Gamma_j^{-1}} \bar{f}_l^{\mathrm{nc}}$.

Equations (2.2)-(2.4) define a route-allocation policy $\mathbf{f}^{\mathrm{c}} = \{f_r^{\mathrm{c}} : r \in \mathcal{R}\}$ for all cooperative users that satisfies all demand requirements. There are three main requirements that we have from the set of constraints: non-compliant (Eulerian) demand is satisfied, compliant (Lagrangian) demand is satisfied, and mass balance across junctions is satisfied. The first two are obvious from the above constraints, while the third one needs proof.

Proposition 1. *For a feasible* $\mathbf{f}^{\mathrm{c}}$ *to the set of Equations (2.2) and (2.3),* $\forall j \in \mathcal{J}$, $\sum_{l\in\Gamma_j} f_l = \sum_{l\in\Gamma_j^{-1}} f_l$.

Proof. From the model consistency condition above, we only need to prove the following statement:

$$\sum_{l\in\Gamma_j} \sum_{r|l\in r} f_r^{\mathrm{c}} = \sum_{l\in\Gamma_j^{-1}} \sum_{r|l\in r} f_r^{\mathrm{c}}$$

Let $\mathbf{r}_j^{\mathrm{in}}$ be the routes that pass through links in the incoming links of junction j. Let $\mathbf{r}_j^{\mathrm{out}}$ be the same for outgoing links. Then $\mathbf{r}_j^{\mathrm{in}} = \left\{r \in \mathcal{R} : \Gamma_j \cap r \neq \emptyset\right\}$. We also know that by the definition of a route, any route that passes through an incoming link of a junction (not a source or sink) must pass through an outgoing link, and therefore $\mathbf{r}_j^{\text{in}} \subseteq \mathbf{r}_j^{\text{out}}$. A similar argument can be made to show $\mathbf{r}_j^{\text{out}} \subseteq \mathbf{r}_j^{\text{in}}$. This shows that $\mathbf{r}_j^{\text{in}} = \mathbf{r}_j^{\text{out}}$. Then,

$$\sum_{l \in \Gamma_j} \sum_{r | l \in r} f_r^{\text{c}} = \sum_{\mathbf{r}_j^{\text{in}}} f_r^{\text{c}} = \sum_{\mathbf{r}_j^{\text{out}}} f_r^{\text{c}} = \sum_{l \in \Gamma_j^{-1}} \sum_{r | l \in r} f_r^{\text{c}}$$

□

2.3. Reducing total latency by rerouting cooperative users

We now formulate the problem of minimizing total latency (or equivalently, total travel time) with route assignments of cooperative vehicles as the decision variable. There are two classes of latency functions studied in the literature: first Nie & Liu 2010; Roughgarden 2001; Papadimitriou 2010, where the link latency is assumed to be the function of flow in the link, and second Tong & Wong 2000; Lo & Szeto 2002; Daganzo 1994; Lighthill & Whitham 1955; Richards 1956, where density is assumed to affect link latencies. We generically introduce latency as a value ℓ_l associated with the state and properties of a link $l \in \mathcal{L}$ and discuss the different models of latency as it pertains to different flow models in Section 4. We can therefore express the total latency on a link as the flow times the latency, or $f_l \ell_l$.

We can now express a general form of the Lagrangian-Eulerian flow, route assignment problem in a standard optimization program formulation:

$$\begin{array}{lrll} \underset{f_l | l \in \mathcal{L}, f^{\text{c}} | r \in \mathcal{R}}{\text{minimize}} & & \sum_{l \in \mathcal{L}} f_l \ell_l & \\ \text{subject to:} & & & \\ & f_l = & \sum_{r | l \in r} f_r^{\text{c}} + \bar{f}_l^{\text{nc}} & \forall l \in \mathcal{L} \\ & \sum_{r | o,d \in r} f_r^{\text{c}} = & \bar{f}_{o,d} & \forall o \in \mathcal{O}, d \in \mathcal{D} \end{array} \qquad (2.5)$$

where the objective represents the total latency in the network, the first constraint relates cooperative route flows and noncooperative link flows to total link flows, and the last constraint states that the compliant route flows must be partitioned in a way that satisfies the nominal origin-destination demand.

3. Accounting for response of noncooperative demand via bounded tolerance

Recall from our discussion in Section 2, that due to imperfect perceptions of the travel times, the travelers forming the non-cooperative demand may be assumed to be boundedly rational, and may not change their nominal paths. This assumption is especially valid when the travel times of non-cooperative users do not significantly change when the cooperative users are routed in a socially optimal manner. In contrast, if the cooperative vehicles cause an excessive increase in latency on some routes for noncooperative vehicles, then this assumption may not be realistic (see Aswani & Tomlin (2011); Roughgarden (2001); Krichene *et al.* (2012b)). For this reason, we have to enhance the model for rerouting cooperative vehicles under stricter conditions.

3.1. Bounded tolerance

A traditional approach to predicting vehicle route choice comes from the field of traffic assignment (e.g. Wardrop (1952); Lo & Szeto (2002); Papageorgiou (1990)) and Nash equilibria in game theory (e.g. Roughgarden & Tardos (2002); Papadimitriou (2010)), often described as *user equilibrium* in the context of traffic assignment and introduced in Wardrop (1952). The congestion games literature considers Stackelberg games, which are used to analyze how selfish users respond to a centrally-controlled subset of users in a principled way (see Krichene *et al.* (2012a); Roughgarden (2001)). Our approach in this paper is simpler.

The reasonability of our approach can be argued using a bounded tolerance assumption on the part of noncooperative vehicles. We replace the assumption of stationarity with a stronger assumption that stationarity is only achieved if all routes on the network do not have a latency increase greater than a certain amount, proportional to the nominal latency experienced before rerouting the cooperative vehicles. Tolerance is assumed in the sense that if latencies on a route do not noticeably increase, then noncooperative vehicles do not seek better paths. However, the tolerance to increase delay is still limited in the sense that as latency increases on a route, noncooperative vehicles will eventually switch. The term *tolerance* is used to address the fact that the bounded rationality models assumed in Hu & Mahmassani (1997); Guo & Liu (2011) allow decisions to be made by the noncooperative users, while our bounded tolerance model specifies how much perturbation of the nominal is allowed, before the assumption the assumption is no longer valid that noncooperative users do not change routes.

3.2. Modeling bounded tolerance

The discussion in Section 3.1 dictates that one must have knowledge about a nominal network state with which to compare final network state conditions. Therefore, we introduce as input, the nominal latency $\bar{\ell}_r$ for every route $r \in \mathcal{R}$. We then select as a model of bounded tolerance the condition that the final latency on a route may not be a factor $(1 + \alpha)$ greater than the nominal latency $\bar{\ell}_r$, where $\alpha \in \mathbb{R}_+$ can be seen as the *tolerance scale factor*. The latency on a route can be computed by summing the latencies experienced on all links in $r = \{r_1, \ldots, r_{|r|}\}$:

$$\ell_r = \sum_{l \in r} \ell_l$$

We can now express the bounded tolerance condition constraints:

$$\sum_{l \in r} \ell_l \quad \leq (1 + \alpha)\, \bar{\ell}_r \quad \forall r \in \mathcal{R} \tag{3.1}$$

Adding this constraint to Problem 2.5 completes the *partial cooperation, bounded tolerance* problem. Section 4 describes how this problem applies to different flow models and latency models. We now introduce another tolerance model and describe the class of problems to which it can be applied.

3.3. Comparative tolerance

The model for bounded tolerance described above places a limit on the increase of latency on a particular route. An alternative approach would be to limit the increase in utility that alternative routes gain over a particular route. In other words, the model developed in Section 3.2 assumes that a particular route flow would be complacent on its original route as long as its own latency does not increase too much, *while not considering the possibility that the utilities of alternative routes may have increased significantly*. To address this limitation, we introduce a *comparative tolerance* model and discuss the underlying modeling assumptions.

We first assume for a given route $r \in \mathcal{R}$ with origin $o_r \in \mathcal{O}$ and destination $d_r \in \mathcal{D}$, and assuming a tolerance scale factor $\alpha = 0$ (no tolerance to delays induced by cooperative flows), that the allowable difference between the route's final latency and the final latency of all other routes sharing the same origin and destination is the largest difference in nominal latencies between itself and all other routes, or 0 if the considered route has the smallest nominal latency. Then, if the scale factor α is greater than 0, this allowable difference is increased by $\alpha\bar{\ell}_r$. Mathematically, we can express this condition as follows:

$$\begin{aligned} \ell_r - \ell_{\tilde{r}} \leq \quad & \max\left(0, \max_{\tilde{r}|o,d\in\tilde{r},\tilde{r}\neq r}\left(\bar{\ell}_r - \bar{\ell}_{\tilde{r}}\right)\right) + \alpha\bar{\ell}_r \quad \forall \tilde{r} : o, d \in \tilde{r}, \tilde{r} \neq r \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall r : o, d \in r, \forall o \in \mathcal{O}, \forall d \in \mathcal{D} \end{aligned}$$

While this formulation more accurately captures the concept of traveler behavior under improved comparative information about alternative routes, it introduces many more constraints than the bounded tolerance formulation in Section 3.2. Additionally, since the LHS of the constraint is a less-than inequality that contains the subtraction of two functions of decision variables, common assumptions on the latency functions will typically lead to a non-convex constraint.

One assumption that will guarantee convexity of the above constraints is if all links have affine latency functions. It can be seen by considering that the LHS is the summation of link latencies along a particular route, and the RHS is a constant that can be computed a priori. In Section 5.1 we will give a numerical example of a simple network with linear latency functions, comparing the output of our model assuming first simple bounded tolerance, and then considering comparative tolerance.

4. Formulating bounded tolerance as a convex optimization problem

The preceding sections discussed a generic model for route-based flow optimization on a flow network with mixed Lagrangian-Eulerian demands, without identifying any specific flow model. In this section, we discuss two types of flow models, horizontal queues and vertical queues. We show for each case how the modeling assumptions can be made into convex constraints, enabling one to solve the partial cooperation, bounded tolerance model as a convex optimization problem. We begin our discussion showing how vertical queues fit cleanly within our model (Section 4.1). However, modeling horizontal queues (e.g. highway networks) requires some additional theoretical setup. What has worked for modeling internet, supply chains, etc. does not work for highway networks, as they are nonlinear systems with non-convex constraints that depend on the density of the links, rather than the flows. Its discussion constitutes the bulk of the section (Section 4.2).

4.1. Vertical queues

Several types of networks, such as communication networks or machine queues (e.g. Roughgarden 2001), can model link latencies as a function of the aggregated flow on the link. To contrast with the model discussed in Section 4.2, we refer to such networks as *vertical queues*. In this section, we show an example of how the concepts of partial cooperation and bounded tolerance can be modeled as a convex optimization program for a specific class of vertical queues, and give a brief discussion on how the results extend to a more general class of vertical queues.

4.1.1. Example: M/M/1 queueing model

A common way to model latencies for communication networks is the M/M/1 queue (e.g. Aswani & Tomlin 2011), which assumes Poisson arrivals and exponential service times. On a link $l \in \mathcal{L}$, the average latency as a function of the rate of Poisson arrivals (the flow) f is given by the equation:

$$\ell_l(f) = \frac{\beta_l}{\mu_l - f} \tag{4.1}$$

where β_l is the occupation rate and μ_l is the processing rate. The flow on a link must be less than the processing rate for the system to be stable. The function $\ell_l(\cdot)$ is convex in $[0, \mu_l)$. We can now substitute Equation (4.1) into (2.5) and (3.1) to obtain the following program:

$$\begin{aligned}
\underset{f_l|l\in\mathcal{L}, f_r|r\in\mathcal{R}}{\text{minimize}} \quad & \textstyle\sum_{l\in\mathcal{L}} \frac{\beta_l f_l}{\mu_l - f_l} \\
\text{subject to:} \quad & \\
\textstyle\sum_{l\in r} \frac{\beta_l}{\mu_l - f_l} \leq \; & \bar{\ell}_r (1+\alpha) && \forall r \in \mathcal{R} \\
f_l = \; & \textstyle\sum_{r|l\in r} f_r^{\mathrm{c}} + \bar{f}_l^{\mathrm{nc}} && \forall l \in \mathcal{L} \\
\sum_{r|o,d\in r} f_r^{\mathrm{c}} = \; & \bar{f}_{o,d} && \forall o \in \mathcal{O}, d \in \mathcal{D}
\end{aligned} \tag{4.2}$$

where again, α is given and corresponds to the maximal threshold tolerable by users if Lagrangian (cooperative) demand perturbs the nominal flow. This program is convex and can be solved by standard convex solvers (with some algebraic manipulations for disciplined convex programming solvers). Indeed, the objective is the summation of convex functions, and the first constraint is a convex inequality *(less-than* inequality with a summation of convex functions on the LHS).

4.1.2. Class of convex vertical queues

This section shows that if all link latencies are convex, increasing functions of flow (e.g. following the modeling assumptions of Roughgarden (2001)), then the partial cooperation, bounded tolerance problem is guaranteed to be convex.

From the discussion in Section 4.1.1, we can generalize the class of latency functions for vertical queues, which lead to a convex formulation. In Equation (2.5), only the objective contains latency terms, and in Equation (3.1), the LHS of the inequality contains latency terms. Therefore, we need to verify the convexity of the objective and the bounded tolerance constraints.

A well known-result of convex analysis is that the summation of convex functions preserves convexity. Therefore, the convexity of the bounded tolerance constraint is guaranteed if the latency function on every link is convex. Additionally, for a link $l \in \mathcal{L}$, the total link latency, $f_l \ell_l (f_l)$, is convex from the assumptions on ℓ_l, and therefore the sum of all total link latencies (which is the objective) is guaranteed to be convex.

4.2. Horizontal queues

A standard assumption in transportation networks is that latencies are not determined uniquely by the flow on the link, but rather how densely populated the link is. Such latency models are referred to as *horizontal queues*, as the congestion on a link occupies physical space which may propagate in the horizontal direction. In this section, we show how the partial cooperation and bounded tolerance models can be extended to horizontal queues. First we develop the relationship between link flow and link density, the resulting latency model, and how a convex optimization problem can be formulated for networks with horizontal queues.

4.2.1. Link model

Constraining flow by link densities. For each horizontal queuing link $l \in \mathcal{L}$, in addition to having a link flow f_l, a horizontal queue link also has a density of vehicles ρ_l, expressing the number of vehicles occupying a link divided by the length L_l. Relating the density of a link to its flow, each link also has a trapezoidal fundamental diagram specified by three parameters: free-flow velocity v_l, congestion velocity w_l, and max flow $f_l^{\max}$ (see Figure 4.1 and references Daganzo (1994, 1995); Garavello & Piccoli (2006)). From these parameters, one can compute the critical density ρ_l^c and jam density $\rho_l^{\max}$. Given that we are assuming the network is in equilibrium, then outflow must equal inflow for each link (see Gomes *et al.* (2008) for a detailed analysis of horizontal queue equilibria). Therefore, only need to consider the single variable f_l when analyzing flow on a link, rather than considering both the inflow and outflow of a link. We express the ρ_l (as traditionally assumed by the LWR equation Daganzo (1994, 1995); Lighthill & Whitham (1955); Richards (1956)), we have two coupled variables f_l and ρ_l, with the following constraints:

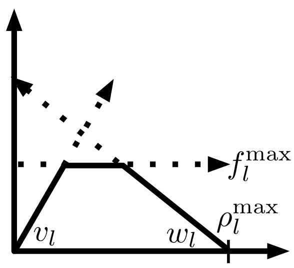


Figure 4.1: Trapezoidal fundamental diagram used to model flow-density relationships in horizontal queues.

$$f_l \leq v_l \rho_l \tag{4.3}$$

$$f_l \leq w_l \left(\rho_l^{\max} - \rho_l\right) \tag{4.4}$$

$$0 \leq f_l \leq f_l^{\max} \tag{4.5}$$

where (4.3) restricts the outflow of link l, (4.4) restricts the inflow, and (4.5) is a physical capacity of the link. These constraints are a relaxation of the fundamental diagram, initially introduced by Gomes & Horowitz (2006).

Latencies. For a link $l \in \mathcal{L}$, the latency ℓ_l is obtained by multiplying the length and velocity of the link, where the velocity of the link is a function of both flow and density. With a notational change (the latency now depending on two variables), the latency function is given by:

$$\ell_l(f, \rho) = \frac{L_l \rho}{f}$$

and the total latency $f_l \ell_l(f_l, \rho_l)$ on a link is given simply by the number of vehicles on a link, or $L_l \rho$. Note that a nominal link latency must be determined from both a nominal flow and nominal density, requiring more information than the point queue model, which only needs nominal link flows.

4.2.2. Relaxation of Junction Model

In order to guarantee uniqueness of solutions of junction problems in LWR networks, it is common to assume that the sum of flows across junctions is maximal, while respecting the prescribed turning ratios. Daganzo (1995) describes a junction model for 2-to-1 merges and 1-to-2 diverges tailored to the CTM model, while Garavello & Piccoli (2006) describe a more general junction model allowing n-to-m merges for the continuous LWR network model, which includes the Daganzo model as a special case with triangular fundamental diagrams and limited merge/diverge types. We refer to the flow-maximizing junction models as the *unrelaxed* junction model, and the flow-density relationship in Section 4.2.1 as the *relaxed* junction model as it does not include a flow maximization condition.

One technical reason why the relaxed model is used is that a flow-maximization condition would lead to a non-convex problem formulation. Another argument that can be made is that for certain junction types, some *split-ratio vector* or *priority vector* (see Coclite & Piccoli (2002)) may exist that would lead to the flow solution given by the optimization problem. Therefore, since this problem has no fixed split-ratios, it can be considered a free variable and the optimization problem has discovered one of the many possible solutions to some junction. This argument has limits, as there is no such free parameter for 1-to-1 junction types, for instance.

There have been methods proposed for dealing with the implicit "car holding" issue introduced from the relaxation, such as adding penalty terms in the objective (see Ziliaskopoulos (2000)), but we do not consider these in our analysis.

4.2.3. Optimization program

For the horizontal queues network, we can now express the total latency minimization problem expressed in Section (3.2):

$$\begin{array}{rlll}
\underset{f_l, \rho_l | l \in \mathcal{L}, f_r | r \in \mathcal{R}}{\text{minimize}} & & \sum_{l \in \mathcal{L}} L_l \rho_l & \\
\text{subject to:} & & & \\
\sum_{l \in r} \frac{L_l \rho_l}{f_l} & \leq & (1+\alpha)\,\bar{\ell}_r & \forall r \in \mathcal{R} \\
f_l & \leq & v_l \rho_l & \forall l \in \mathcal{L} \\
f_l & \leq & w_l \left(\rho_l^{\max} - \rho_l\right) & \forall l \in \mathcal{L} \\
0 & \leq & f_l \leq f_l^{\max} & \forall l \in \mathcal{L} \\
f_l & = & \sum_{r | l \in r} f_r^{\mathrm{c}} + \bar{f}_l^{\mathrm{nc}} & \forall l \in \mathcal{L} \\
\sum_{r | o, d \in r} f_r^{\mathrm{c}} & = & \bar{f}_{o,d} & \forall o \in \mathcal{O}, d \in \mathcal{D}
\end{array} \tag{4.6}$$

This formulation is not convex, specifically the bounded tolerance constraint is not convex. There is a superseding problem with the formulation, that the bounded tolerance constraints and outflow constraints

($f_l \leq w_l \left(\rho_l^{\max} - \rho_l\right)$) are guaranteed to be non-binding. Several of the constraints can be shown to be non-binding by observing that a solution must satisfy $\rho_l = \frac{f_l}{v_l}$. We prove that next.

Lemma 2. *If the solution* $\mathbf{f}^{c*}, \rho^*$ *is optimal for Problem 4.6, then* $\forall l \in \mathcal{L}$: $\rho_l^* = \frac{f_l^*}{v_l}$

Proof. Assume $\exists \rho_l > \frac{f_l^*}{v_l}$. Reducing ρ_l to $\frac{f_l^*}{v_l}$ only decreases the LHS of the first constraint in Problem (4.6). The second constraint becomes an equality by construction. The RHS of the third constraint increases. Since the flow terms are not changed, we see that the feasibility of the problem is maintained. Additionally, the objective strictly decreases, thus proving that a solution with such a ρ_l is sub-optimal. □

We can now simplify Problem 4.6 by substituting in the value of ρ from Lemma 2, and using the following notational change for the parameters L_l and $a_l = \frac{L_l}{v_l}$:

$$
\begin{aligned}
&\min_{f_l | l \in \mathcal{L}, f_r^c | r \in \mathcal{R}} && \textstyle\sum_{l \in \mathcal{L}} a_l f_l && (4.7)\\
&\text{subject to:} \\
&&& \textstyle\sum_{l \in r} L_l v_l \leq (1+\alpha)\,\bar{\ell}_r && \forall r \in \mathcal{R}\\
&&& f_l \leq v_l \left(\frac{f_l}{v_l}\right) && \forall l \in \mathcal{L}\\
&&& f_l \leq w_l \left(\rho_l^{\max} - \frac{f_l}{v_l}\right) && \forall l \in \mathcal{L}\\
&&& 0 \leq f_l \leq f_l^{\max} && \forall l \in \mathcal{L}\\
&&& f_l = \textstyle\sum_{r | l \in r} f_r^c + \bar{f}_l^{nc} && \forall l \in \mathcal{L}\\
&&& \textstyle\sum_{r | o,d \in r} f_r^c = \bar{f}_{o,d} && \forall o \in \mathcal{O}, d \in \mathcal{D}
\end{aligned}
$$

We can now detect non-binding constraints easily. The first constraint in Problem (4.7) must be satisfied because the LHS is the free-flow travel time of the route and is minimal, while the RHS must be greater or equal to free-flow (keeping in mind $\alpha \geq 0$). The second constraint is always an equality, by Lemma Lemma 2. The third constraint is guaranteed from the assumption of a trapezoidal fundamental diagram. The simplified problem is now:

$$
\begin{aligned}
&\min_{f_l | l \in \mathcal{L}, f_r^c | r \in \mathcal{R}} && \textstyle\sum_{l \in \mathcal{L}} a_l f_l && (4.8)\\
&\text{subject to:} \\
&&& 0 \leq f_l \leq f_l^{\max} && \forall l \in \mathcal{L}\\
&&& f_l = \textstyle\sum_{r | l \in r} f_r^c + \bar{f}_l^{nc} && \forall l \in \mathcal{L}\\
&&& \textstyle\sum_{r | o,d \in r} f_r^c = \bar{f}_{o,d} && \forall o \in \mathcal{O}, d \in \mathcal{D}
\end{aligned}
$$

The above problem is now in a linear program formulation. We now show that the concept of noncooperative flow can be replaced by a capacity reduction on all the links. Let us rework some of the expressions in terms of the cooperative and noncooperative vehicles:

$$
\begin{aligned}
&\min_{f_r^c | r \in \mathcal{R}} && \textstyle\sum_{l \in \mathcal{L}} a_l \bar{f}_l^{nc} + \sum_{l \in \mathcal{L}} a_l \sum_{r | l \in r} f_r^c && (4.9)\\
&\text{subject to:} \\
&&& -\bar{f}_l^{nc} \leq 0 \leq \textstyle\sum_{r | l \in r} f_r^c \leq f_l^{\max} - \bar{f}_l^{nc} && \forall l \in \mathcal{L}\\
&&& \textstyle\sum_{r | o,d \in r} f_r^c = \bar{f}_{o,d} && \forall o \in \mathcal{O}, d \in \mathcal{D}
\end{aligned}
$$

We can now simplify further. The first term in the objective is constant, since f_r^{nc} is not a decision variable. Then, the second constraint can be simplified by introducing a reduced capacity constant, $\bar{f}_l^{\max} =$

$f_l^{\max} - \bar{f}_l^{\mathrm{nc}}$. If we drop the *cooperative* pretense from the decision variable, then we have reduced the problem to a modified capacity, constant latency, Lagrangian system optimal problem, which is simplified and linear:

$$\begin{aligned} \min_{f_r \mid r \in \mathcal{R}} \quad & \textstyle\sum_{l \in \mathcal{L}} a_l \sum_{r \mid l \in r} f_r \\ \text{subject to:} \quad & \\ & 0 \leq \textstyle\sum_{r \mid l \in r} f_r \leq \bar{f}_l^{\max} \quad \forall r \in \mathcal{R} \\ & \textstyle\sum_{r \mid o,d \in r} f_r = \bar{f}_{o,d} \quad \forall o \in \mathcal{O}, d \in \mathcal{D} \end{aligned} \tag{4.10}$$

Lemma 3. *Let* $\mathbf{f}^* = \{f_r^* : r \in \mathcal{R}\}$ *be a solution to Problem* (4.10)*. Then*

$$\begin{aligned} \mathbf{f}^{\mathrm{c}\prime} &= \mathbf{f}^* \\ \rho_l' &= \frac{\bar{f}_l^{\mathrm{nc}} + \sum_{r \mid l \in r} f_r^*}{v_l} \end{aligned}$$

is a solution to Problem (4.6).

Proof. Using Lemma 2, the equality $f_l = \bar{f}_l^{\mathrm{nc}} + \sum_{r \mid l \in r} f_r^{\mathrm{c}}$, and the variable name substitution made in Problem 4.10, the result follows immediately. □

Corollary 4. *An optimal solution to Problem* (4.6) *is a feasible solution of the unrelaxed junction model in Section 4.2.2.*

Proof. Since the flow on every link is in free flow ($f_l = v_l \rho_l, \forall l \in \mathcal{L}$), the supply $\sum_{l \in \Gamma_j} v_l \rho_l$ at every junction $j \in \mathcal{J}$ is equal to the flow across the junction $\sum_{l \in \Gamma_j} f_l$, and is therefore maximal. □

This corollary shows that solving for the static social optimum on networks with horizontal queues does not encounter the non-convexity issues typically associated with the CTM constraints in dynamic traffic problems. For instance, Ziliaskopoulos (2000) uses the relaxed junction model (which allows "car-holding") that we present in Section 4.2.1 to solve the single destination social optimum problem as a linear program, and Gomes & Horowitz (2006) use a relaxed junction model to solve an optimal ramp metering problem as a linear program (with a zero-relaxation gap under certain conditions).

Commonly considered problems in traffic assignment such as social optimum for purely Lagrangian flow ($\bar{f}_l^{\mathrm{nc}} = 0, \forall l \in \mathcal{L}$) and purely Eulerian flow ($f_r^{\mathrm{c}} = 0, \forall r \in \mathcal{R}$) serve as special cases of Corollary 4 and therefore an optimal solution can be found for both problems for the unrelaxed junction model by solving the linear program in Problem (4.10).

4.2.4. Limiting deviations in density

There are limitations in the expressiveness of the current horizontal queues model under total latency minimization. To circumvent these issues, this section proposes the addition of constraints that restrict the allowable densities to be within the locality of the nominal densities that are used to compute nominal latencies.

The purpose of these constraints is to prevent the optimization program from setting all links to be in the free-flow state, which has the negative effect of over-simplifying the model developed here-within (Section (4.2.3)). Instead, total latencies across the network can be minimized *while considering likely congestion patterns*. To motivate the usefulness of such a model, one can make a physical argument that rerouting may only cause bounded deviations in the density, and that congestion may not be cleared due to rerouting because of additional issues such as weaving or the physical road conditions.

To restrict the densities to only take certain values, we require that each link $l \in \mathcal{L}$, includes an upper and lower density bound, $\rho_l^{\uparrow}$ and $\rho_l^{\downarrow}$ respectively. We append to the program in Equation (4.6), the set of constraints bounding the allowable densities:

$$\rho_l^{\downarrow} \leq \rho_l \leq \rho_l^{\uparrow} \quad \forall l \in \mathcal{L}$$

In Section 5, we show an example of a network with horizontal queues with density bounds that has the bounded tolerance constraint as a tight constraint. This demonstrates that bounding the allowable densities can capture the characteristics of bounded tolerance for networks with horizontal queues.

4.3. Algorithm for data preconditioning

If input data into our problem is taken from a physical network with inherent sources of noise, it is likely that there will be a number of conditions that will cause the raw data to be incompatible with the problem constraints, thus making the problem infeasible. For instance, a link's estimated density may not lie within the fundamental diagram constraints in Section 4.2.1, or there may not be exact mass balance across junctions. If the estimates from the stationary sensors are reasonable, then these constraint violations will not be severe, but even small deviations will render the optimization program infeasible. Therefore the input data must be filtered to be preconditioned to meet the requirements. We propose an optimization program formulation.

The constraints that concern noncooperative flow are the following:

$$f_l \leq v_l \rho_l \qquad \forall l \in \mathcal{L} \tag{4.11}$$
$$f_l \leq w_l \left(\rho_l^{\max} - \rho_l\right) \quad \forall l \in \mathcal{L} \tag{4.12}$$
$$0 \leq f_l \leq f_l^{\max} \qquad \forall l \in \mathcal{L} \tag{4.13}$$
$$f_l = \textstyle\sum_{r|l\in r} f_r \qquad \forall l \in \mathcal{L} \tag{4.14}$$

These constraints are all convex (indeed, linear) in the decision variables f_l, f_r. Then, let $\hat{f}_l, \hat{\rho}_l$ be the input flow and density respectively on link $l \in \hat{\mathcal{L}}$, where $\hat{\mathcal{L}} \subseteq \mathcal{L}$ are the links with input data available. Our objective will be to minimize some definition of distance from the input data to the selected data that violates none of the above constraints. If we select as the distance measurement the *n-norm*, $n \geq 1$, then we have the following convex optimization program for obtaining amenable input data:

$$\begin{array}{ll} \underset{f_l,\rho_l : l\in\mathcal{L}}{\text{minimize}} & \sum_{l\in\bar{\mathcal{L}}} \left\|\hat{f}_l - f_l\right\|_n + \|\hat{\rho}_l - \rho_l\|_n \\ \text{subject to:} & \text{Constraints (4.11) – (4.14)} \end{array}$$

The result of the optimization problem is a set of route-based flows $\{f_r : r \in \mathcal{R}\}$. Finally, the route flows would then be partitioned into both cooperative and noncooperative flows, which then gets the data in a suitable format.

While the formulation presented specifically discusses horizontal queue constraints, the same methods can be extended to other problems with convex constraints, such as M/M/1 queues presented in Section 4.1.1.

5. Numerical results

We demonstrate the highly practical nature of our work by applying the model to two different problems. We focus on a multi-destination network of horizontal queues. This problem demonstrates the generality of our method to the multi-commodity case and the ability to solve real-world transportation planning problems on a regional level. First, we demonstrate the simplicity of the model on a small network of vertical queues, to which we apply both models of tolerance and compare the benefits gained by re-routing.

5.1. Linear Latency

Name	$a\left(\frac{s^2}{\text{units}}\right)$	$b\,(s)$	$f^{\max}\left(\frac{\text{units}}{s}\right)$
source	1	0	1
sink	1	0	1
left	1	0	1
right	0.5	0.5	1

(a)

Type	Description	$f\,(\frac{\text{cars}}{\text{s}})$
O-D (Lagrangian)	source-sink	0.8
Link (Eulerian)	source	0.2
Link (Eulerian)	sink	0.1
Link (Eulerian)	left	0.1
Link (Eulerian)	right	0.2

(b)

Table 2: Summary of illustrative network properties. **2a:** Link-level input parameters. **2b:** Network-level input demands

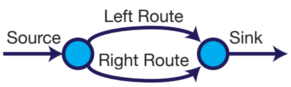


Figure 5.1: Network diagram

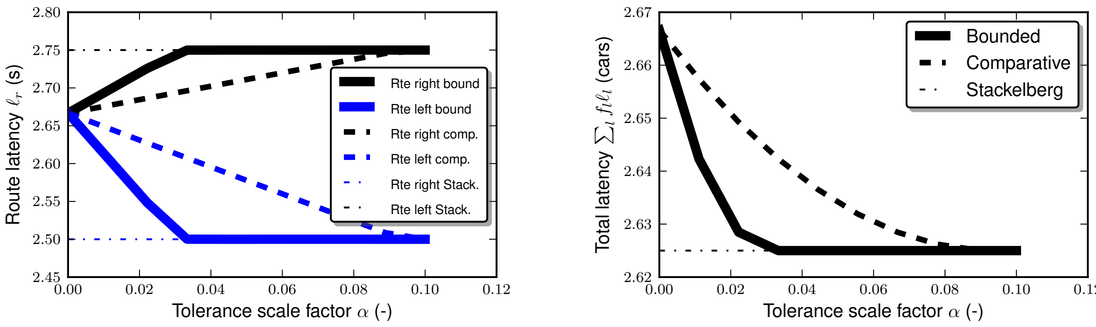
Figure 5.1 depicts an illustrative, two parallel routes network. Flow enters at the source and exits at the sink and can travel along either the "left" route or the "right" route. Each link $l \in \mathcal{L}$ has a linear latency function $\ell_l(f) = a_l f + b_l$. The link properties given in Figure 2a show that the left link has a lower *zero-flow* latency than the right link, but has a higher marginal cost per unit flow. As the left route becomes more congested, its latency will eventually increase until the right route has equal utility. From a user equilibrium viewpoint, as the network is loaded with additional flow, the latencies across the two routes will remain the same. But from a social optimum viewpoint, additional flow will always be routed to the right route since it will always have a lower marginal cost than the left route.

As described in Figure 2b, we assume the network is loaded with both cooperative and noncooperative flow. There is a total of 0.2 units-per-second of noncooperative flow, with 0.1 units-per-second of flow on both the left and right link. In addition, there is 0.8 units-per-second of flow on the network, which we assume is initially distributed amongst the left and right routes in a manner that achieves user equilibrium.

We now show how our route optimization framework can be applied to this network to optimally route the cooperative flow. Results are given over a range of bounded rationality scale factor to show the sensitivity of our results to the scale factor. Since the latency functions are linear, both the standard bounded rationality model and comparative bounded rationality model are solved using well-developed and highly efficient convex optimization tools (see Rubira (2011)). In addition to comparing the two models against each other, we compare them both against the Stackelberg game solution of the routing problem. The Stackelberg game solution gives a minimum social cost with the assumption that the noncooperative flow will be routed in a user equilibrium manner. Stackelberg analysis is only possible in the case when origin-destination demands can be uniquely determined for all users of the network, which holds for our simple network. This does not hold in general, and this case will studied subsequently (Section 5.2).

Figure 5.2 summarizes the numerical results on the simple network. The route latencies as a function of the bounded rationality scale factor are shown in Figure 5.2a, while total latencies are shown in Figure 5.2b. As expected, as the bounded rationality scale factor increases, so do the benefits of re-routing. Additionally, the comparative bounded rationality model improves at a slower rate than the standard bounded rationality model. This is due to the fact that the comparative model permits the right route to be "aware" of the latency improvements on the left route, while the standard model only limits deviations in route latencies in comparison to a route's individual *nominal* latency and ignores the improvements on the left route.

The results tell us that as the scale factor increases, the model converges to the Stackelberg solution. It may appear counter-intuitive that the model with inherently no tolerance factor could perform better than the tolerance models. The explanation is that the tolerance models are overly-conservative due to the assumption that no noncooperative flow changes routes, and the routing strategy will not drastically improve one route over another route. On the other hand, the Stackelberg solution shifts all noncooperative flow to the left route and the cooperative flow accommodates this shift in a socially optimal manner. Since all noncooperative flow is on a single route and improves upon its nominal latency, discrepancies in route latencies are no longer a behavioral issue, allowing the Stackelberg to be as liberal as necessary with latency increases on



(a) (b)

Figure 5.2: Comparison of simple bounded tolerance and comparative tolerance. **5.2a:** Route latencies. Comparative tolerance allows smaller deviations in route latencies than bounded tolerance. **5.2b:** Total latencies. The total latencies decrease more slowly with the comparative tolerance model versus the bounded tolerance model. The total route flows approach the Stackelberg equilibrium as the tolerance scale factor goes to infinite.

the right route.

5.2. Horizontal queueing network

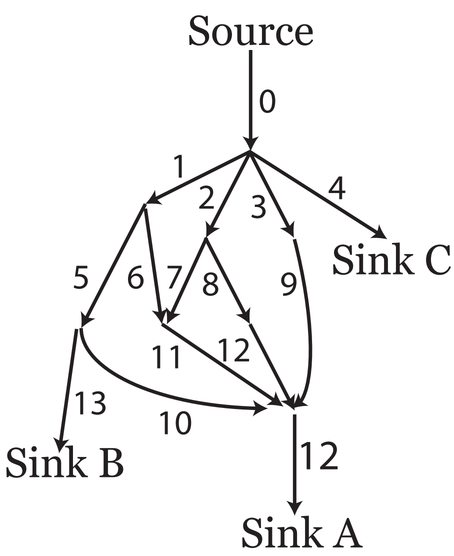


Figure 5.3: Multiple-destination network with horizontal queues. There are many overlapping routes between *Source* and *Sink A*, while *Sink B* and *Sink C* are origin-destinations which have demands on the same network as *Sink A* demands.

As discussed in Section 4.2, given the nonlinear dynamics of horizontal queueing cells, modeling horizontal queues is in general a more difficult process than vertical queues. It is also important to consider a more general network than the compact one in Section 5.1, one with multiple destinations, and therefore multiple Lagrangian demand types. In this section, we model a mid-sized multi-destination network of horizontal queues within the partial cooperation, bounded tolerance framework. We follow with numerical results on how the routing strategies change with respect to the parameters of the tolerance model.

5.2.1. Network properties and demands

Figure 5.3 shows a topological description of the network. Since only *Sink A* can be reached through multiple routes, the algorithm only decides how to partition the demand across the routes originating from *Source* and leading to *Sink A*. The algorithm will take as input some network and link level properties, as recorded in Table 3. The nominal state given in Table 3a shows that links 3, 8 and 9 were heavily congested, while the other links were close to free flow. Furthermore, as discussed in Section 4.2.4, the densities must have more constraints than just the fundamental diagram constraints (Equations (4.3)-(4.5)). Table 3a tells us that this problem assumes that links do not shift from their nominal state (links with a nominal free flow/congestion state must maintain this state). Table 3b

Name	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Length (m)	0.5	0.1	1.0	1.5	0.3	0.1	0.6	0.4	1.0	1.5	1.0	0.7	0.8	1.0	0.1
Nom. flow ($\frac{\text{units}}{s}$)	0.15	0.05	0.03	0.02	0.05	0.05	0	0.02	0.01	0.02	0	0.02	0.01	0.05	0.05
Nom. density ($\frac{\text{units}}{m}$)	0.15	0.05	0.03	4.99	0.05	0.05	0	0.02	4.99	4.99	0	0.02	0.01	0.05	0.05
State constraint	FF	FF	FF	Cong	FF	FF	FF	FF	Cong	Cong	FF	FF	FF	FF	FF

(a)

Sink Dest.	A	B	C
Flow ($\frac{\text{users}}{s}$)	0.5	0.5	0.5

(b)

Table 3: Multi-destination network properties. **3a:** Link properties, including nominal state. **3b:** demand input into network

tells us that there is 0.5 $\frac{\text{users}}{sec}$ demand between all origin-destination pairs on the network. For simplicity, we assume that all demand is cooperative as well to focus analysis.

5.2.2. Numerical results

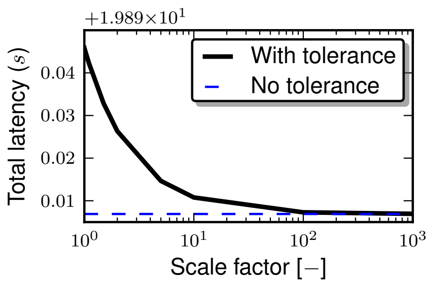


Figure 5.4: Total latency on network of horizontal queues as a function of tolerance scale.

The results of our numerical calculations are summarized in Figure 5.4. As supported by the results for vertical queues in Section 5.1, the relief of network congestion is greater the more tolerance is assumed in the users. Additionally, it is noted that the network does not immediately push into free flow (social optimum), but rather decongests links to an amount dependent on the tolerance scale factor. This is a desirable behavior of the model, as it is not reasonable to assume that congestion can be completely avoided just through re-routing schemes. Lastly, we see the intuitive result that the bounded tolerance model will converge to the more familiar social optimum as the scale factor increases.

6. Conclusion

We have presented a framework rerouting flow in an socially optimal way with mixed Lagrangian-Eulerian information. The cooperative flow has known nominal routes, while the noncooperative flow has known flow counts across links. In order to anticipate network conditions for all users after re-routing has been applied, the model combines the two types of information in a complementary way; by only allowing the cooperative flow to change routes, we have removed the necessity of having origin-destination demand information for all users on the network. Furthermore, by looking at the static flow problem, we can study practical networks with multiple origins and multiple destinations, where dynamic multi-commodity problems often suffer from intractability issues.

The framework also addresses the behavioral nature of the noncooperative users, which we call *bounded tolerance* and *comparative tolerance*, by only allowing perturbations of the nominal state of the network that boundedly impact the noncooperative flow in a negative manner. The tolerance model comes about as a response to the lack of origin-destination information that does not permit the game-theoretic Stackelberg game analysis, but does allow us to require only Eulerian information across the majority of the network. We show that the comparative tolerance model will in general limit network latency improvements more so than the bounded tolerance model, but since the comparative tolerance model allows individual routes to compare latencies with other routes, it is arguably a more accurate model of noncooperative flow behavior.

By taking a convex optimization-based approach, the framework is shown to efficiently solve many classes of network flow problems. The horizontal queue, highway network problem can be modeled as a convex optimization program, which permits one to study highway networks of practical size. The multi-destination network of horizontal queues gives an example of how the framework can be applied to multi-commodity type networks such as highways with multiple onramps and off-ramps. A live data feed of Lagrangian GPS sensors and Eulerian loop detectors, in conjunction with the data pre-conditioning algorithm, would enable the framework to run in an "online" sense, and provide automatic, daily routing advice for a traffic management agency during rush hour periods. We conclude that the partial cooperation, bounded tolerance model can allow a traffic management operator to make beneficial re-routing decisions with much less origin-destination demand input required.

References

Aswani, A., & Tomlin, C. 2011. Game-theoretic routing of GPS-assisted vehicles for energy efficiency. *Pages 3375–3380 of: American Control Conference (ACC), 2011*. IEEE.

Coclite, G.M., & Piccoli, B. 2002. Traffic Flow on a Road Network. *Arxiv preprint math/0202146*.

Daganzo, C. F. 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological*, **28**(4), 269–287.

Daganzo, C. F. 1995. The cell transmission model, part II: network traffic. *Transportation Research Part B: Methodological*, **29**(2), 79–93.

Garavello, M., & Piccoli, B. 2006. *Traffic flow on networks*. American institute of mathematical sciences Springfield, MA, USA.

Gomes, G., & Horowitz, R. 2006. Optimal freeway ramp metering using the asymmetric cell transmission model. *Transportation Research Part C: Emerging Technologies*, **14**(4), 244–262.

Gomes, G., Horowitz, R., Kurzhanskiy, A. A., Varaiya, P., & Kwon, J. 2008. Behavior of the cell transmission model and effectiveness of ramp metering. *Transportation Research Part C: Emerging Technologies*, **16**(4), 485–513.

Guo, X., & Liu, H.X. 2011. Bounded rationality and irreversible network change. *Transportation Research Part B: Methodological*, **45**(10), 1606–1618.

Hu, T., & Mahmassani, H.S. 1997. Day-to-day evolution of network flows under real-time information and reactive signal control. *Transportation Research Part C: Emerging Technologies*, **5**(1), 51–69.

Krichene, W., Reilly, J., Amin, S., & Bayen, A.M. 2012a. On the Characterization and Computation of Nash Equilibria in Horizontal Queueing Networks. *In: Conference on Decision and Control*.

Krichene, W., Reilly, J., Amin, S., & Bayen, A.M. 2012b. Stackelberg Routing on Horizontal Queueing Networks. *In: Conference on Decision and Control*.

Leblanc, R., & Walker, J. L. 2012. *Which is the biggest carrot? Comparing non-traditional incentives for demand management*.

Lighthill, M.J., & Whitham, G.B. 1955. On kinematic waves. II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, **229**(1178), 317.

Lo, H. K., & Szeto, W. Y. 2002. A cell-based variational inequality formulation of the dynamic user optimal assignment problem. *Transportation Research Part B: Methodological*, **36**(5), 421–443.

Merugu, D., Prabhakar, B.S., & Rama, N.S. 2009. An incentive mechanism for decongesting the roads: A pilot program in bangalore. *NetEcon, ACM Workshop on the Economics of Networked Systems*.

Nie, Y., & Liu, Y. 2010. Existence of self-financing and Pareto-improving congestion pricing: Impact of value of time distribution. *Transportation Research Part A: Policy and Practice*, **44**(1), 39–51.

Papadimitriou, C. 2001. Algorithms, games, and the internet. *Pages 749–753 of: Proceedings of the thirty-third annual ACM symposium on Theory of computing - STOC '01*. New York, New York, USA: ACM Press.

Papadimitriou, C. 2010. Selfish Routers and the Price of Anarchy. *Science*.

Papageorgiou, M. 1990. Dynamic modeling, assignment, and route guidance in traffic networks. *Transportation Research Part B: Methodological*, **24**(6), 471–495.

Richards, P.I. 1956. Shock waves on the highway. *Operations research*, **4**(1), 42–51.

Roughgarden, T. 2001. Stackelberg scheduling strategies. *Pages 104–113 of: Proceedings of the thirty-third annual ACM symposium on Theory of computing*. ACM.

Roughgarden, T., & Tardos, É. 2002. How bad is selfish routing? *Journal of the ACM (JACM)*, **49**(2), 236–259.

Rubira, Tomas Tinoco De. 2011. *Cvxpy*.

Tong, C. O., & Wong, S. C. 2000. A predictive dynamic traffic assignment model in congested capacity-constrained road networks. *Transportation Research Part B: Methodological*, **34**(8), 625–644.

Wardrop, J.G. 1952. Some Theoretical Aspects of Road Traffic Research. *Proceedings of the Institution of Civil Engineers*, **1**, 325–378.

Ziliaskopoulos, A. K. 2000. A linear programming model for the single destination system optimum dynamic traffic assignment problem. *Transportation science*, **34**(1), 37.