Transportation Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Impact of Optimized Fleets in Transportation Networks

Matthew Battifarano, Sean Qian

To cite this article:
Matthew Battifarano, Sean Qian (2023) The Impact of Optimized Fleets in Transportation Networks. Transportation Science

Published online in Articles in Advance 10 Jan 2023

. https://doi.org/10.1287/trsc.2022.1189

https://pubsonline.informs.org/journal/trsc

TRANSPORTATION SCIENCE
Articles in Advance, pp. 1–22
ISSN 0041-1655 (print), ISSN 1526-5447 (online)

The Impact of Optimized Fleets in Transportation Networks

Matthew Battifarano,[a] Sean Qian[a,b,*]

[a] Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [b] Heinz College, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

*Corresponding author

Contact: mbattifa@andew.cmu.edu (MB); seanqian@cmu.edu, https://orcid.org/0000-0001-8716-8989 (SQ)

Received: December 23, 2021
Revised: May 27, 2022; September 4, 2022
Accepted: November 20, 2022
Published Online in Articles in Advance: January 10, 2023

https://doi.org/10.1287/trsc.2022.1189

Abstract. Connected technologies have engendered a paradigm shift in mobility systems by enabling digital platforms to coordinate large sets of vehicles in real time. Recent research has investigated how a small number of connected vehicles may be coordinated to reduce total system cost. However, platforms may coordinate vehicles to optimize a fleet-wide objective which is neither user nor system optimal. We study the behavior of optimized fleets in mixed traffic and find that, at small penetrations, fleets may worsen system cost relative to user equilibrium, and provide a concrete example of this paradox. Past a critical penetration level, however, optimized fleets reduce system cost in the network, up to achieving system optimal traffic flow, without need for an external subsidy. We introduce two novel notions of fleet-optimal mixed equilibria: critical fleet size for user equilibrium (CFS-UE) and critical fleet size for system optimum (CFS-SO). We demonstrate on the Sioux Falls and Pittsburgh networks that 33% and 83% of vehicles, respectively, must participate in the fleet to achieve system optimum. In Pittsburgh, we find that, although fleets permeate the network, they accumulate on highways and major arterials; the majority of origin-destination pairs are either occupied exclusively by users or by the fleet. Critical fleet size offers regulators greater insight into where fleet and system interests align, transportation planners a novel metric to evaluate road improvements, and fleet coordinators a better understanding of their efforts to optimize their fleet.

Funding: This work was supported by the U.S. Department of Transportation [Mobility21] and the National Science Foundation [CMMI-1931827].

Supplemental Material: The online appendices are available at https://doi.org/10.1287/trsc.2022.1189.

Keywords: transportation network equilibrium • fleet optimum • network flow • system optimum • user equilibrium

1. Introduction

The central claim of this paper is that groups of vehicles coordinated to achieve a group-level objective have significant and diverging impacts on total travel cost on the road network. On the one hand, vehicle coordination may increase total travel cost on the network. On the other hand, if deployed on certain origin-destination pairs in sufficient volume, a fleet can reduce system cost on the network, up to achieving the system optimal (SO) network flow. In this work, we consider the static network effect of a single fleet that routes its vehicles in a fleet optimal (FO) manner to minimize the total fleet travel cost in mixed traffic with individual vehicles who seek to minimize their own travel cost.

Technological advances have allowed mobility and information service providers to influence traveler decision making in real time and on an unprecedented scale. Ride-sourcing vehicles, connected and autonomous vehicles, vehicles using real-time navigation devices, and vehicles using carpool matching services are all examples of what we call ad hoc fleets currently operating on road networks. We refer to the platforms who coordinate ad hoc fleets interchangeably as "service providers" or "fleet coordinators."

It is often in the interest of the service to coordinate the vehicles in their fleet to advance a service-level goal that may not always align with the goal of an individual user. In this work, we examine the goal of minimizing average fleet travel cost or, equivalently, travel time. This is perhaps the simplest fleet-level goal lying in the intersection of plausible and interesting fleet behavioral principles; it is by no means the only one.

Ride-sourcing platforms like Uber and Lyft coordinate drivers in a variety of ways to benefit the platform. Matching riders and drivers on the Uber platform, for example, is optimized in batches over a local fleet (Uber Technologies 2022). With the same information, an individual driver could likely find a better match than the one they were assigned precisely because the matching was done to *minimize* a fleet level metric rather than *equalize* a driver level metric. Minimizing travel time in particular is an important existing goal of strategy at Uber. For a fixed demand and fixed driver pool, quicker service means increased capacity and higher level of

service, which in turn can demonstrate the value of the platform to new riders and drivers. In another example, Uber uses "pickup spots" to reduce congestion *among Uber vehicles* in small areas of high demand. Many airports, for example, have dedicated entire wings of near-terminal parking structures to coordinate ride-hailing pickups. This is fundamentally a fleet (and likely system) optimal rather than user optimal solution: For each individual rider-driver pair, terminal curb-side pickup would offer travel-time savings, but measured across the fleet, organized parking lot pickups offer less total time wasted.

Real-time navigation systems with substantial user pools like Google Maps also have the means and motivation to induce FO behavior. Conventional reasoning among network equilibrium theorists is these systems will lead to a user equilibrium (UE) by providing travelers with accurate day-to-day traffic information to find their least cost route. However, the goal of Google Maps is to generate revenue. It will only help users find least cost paths insofar as it drives ad revenue: by keeping users on the platform. In this light, UE behavior is but one of many possible results of Google Maps use; FO is another possibility with compelling rationale. Google Maps currently balances several factors in selecting routes including travel time, future predicted traffic, emissions (i.e., "green routes"; Alcántara 2021), road quality, directness, and safety, among others (Lau 2020). The fact that Google Maps is predicting traffic while simultaneously directing a subset of it means that it has, intentionally or not, answered the question: How will our traffic predictions take into account the directions provided? If the predictions do not take into account the directions at all, then the predictions will be inaccurate when a large volume of travelers use the service. If they do, then the two should be mutually consistent: the traffic predictions used to generate the directions remain accurate when the traffic volume of users following those directions is fully incorporated into traffic prediction. There are many models that achieve mutual consistency, and FO is a particularly compelling one. Under FO, because the fleet rather than the individual cost is minimized, the directions are not always fair: some users end up on better routes than others, but its users are, on average, better off than they otherwise would have been. For travelers who travel many origin-destination pairs over their lifetime on the platform, this on-average benefit is what matters. However, Google Maps has a great deal of flexibility in how it presents routes to users so that it may not be obvious to a user, nor may they care, and is in fact very difficult to validate, that their route is not best. Green routes are used by many on the platform even though they are explicitly labeled as being slower than alternative routes. FO in this setting offers a competitive advantage to real-time navigation providers with large user pools: In the same way that large firms negotiate lower unit prices, these services can extract lower unit travel costs via coordination. For users then, Google Maps may be able to offer them lower travel costs on average than individuals and competitors with smaller market share, thereby keeping users on its platform.

In short, fleet coordinators may, and in many cases already do, influence the behavior of their fleets to improve a metric computed over the fleet for the benefit of the service. In this work, we take this metric to be the total travel cost of the fleet and present the system-level implications of its use. Travelers who minimize travel time when driving alone may opt to give up their ability to choose their own route in return for some benefit. In the case of ride-sourcing, this benefit is the ability to use one's time more productively. In the case of real-time navigation, the benefit is the ability to follow directions instead of keeping track of directions in one's head. In both cases, the service offers a benefit that makes the traveler more flexible in terms of the travel time of the routes they are willing to accept. From the perspective of the platform, this flexibility is an opportunity to provide acceptable routes that further a business goal. We find travel time minimization in this context to be a parsimonious choice. In minimizing total fleet travel time, the platform uses its market power to squeeze to extract smaller average travel costs, possibly at the expense of other road users. The average travel time minimization is a clear benefit to the fleet coordinator but is also an attractive proposition to many road users who might not benefit on every origin-destination (OD) pair but on average come out ahead. For users of ride-hailing services and real-time navigation devices who in general will traverse many different OD pairs on the network, this on-average travel time benefit, in addition to the nontravel time benefits previously discussed, present a compelling case for the adoption of the particular service. This is especially notable in light of the subscription model ride-hailing companies like Uber are pursuing, whereby users may opt in to a monthly fee to gain access to platform benefits. Regardless, platforms who coordinate FO behavior across their fleet *can* pass total travel time savings onto their users via incentive schemes. The design and evaluation of such schemes are out-of-scope for this work but are investigated in detail in upcoming research. This paper illustrates and highlights the importance of considering fleet's goals in system-level planning and operation, and the methodology and solutions can be extended to incorporate other system-level metrics in future work.

If ad hoc fleets are to remain a fixture on road networks, how should transportation planners understand and anticipate their use of transportation infrastructure? For example, in analogy to oligopoly models in economics, if a service provider wished to leverage its market power to extract a better deal on travel cost from the

network via route choice coordination of its fleet, what would happen to network efficiency? Furthermore, are there network designs that align the interests of the fleet and society at large? By that same token, fleet coordinators also have an interest in understanding the network impacts of coordinating their fleet: what discount on travel cost does their market power allow them to extract? Perhaps there is an opportunity for service providers to align their goals with traffic managers or pass congestion-relief incentives to riders and drivers. Both the identification of such opportunities and the measurement of their benefits rely on a framework for understanding how vehicles with fleet-level goals and those with individual goals interact on a network.

The remainder of this paper is organized as follows. Section 2 examines the line of inquiry in which this work participates while also placing our work in the context of other perspectives on the relationship between ad hoc fleets and total system cost. Section 3 introduces the notation and fundamental concepts used throughout this paper. Section 4 presents an example of the "fleet optimality paradox," in which the presence of a fleet increases total system travel time relative to UE. It is followed by examination of conditions under which fleets *do* improve total system travel time. In Section 5, we introduce two important mixed equilibria: the smallest fleet to induce SO, termed the critical fleet size for SO (CFS-SO), and the largest fleet to induce UE, termed the critical fleet size for UE (CFS-UE). CFS-SO and CFS-UE are examined analytically in a parallel network. Solution methodologies are then developed to solve both CFS-SO and CFS-UE in general networks. Section 6 presents the critical fleet size solution on two networks and provides an analysis of the results. Section 7 discusses our findings, outlines potential areas of future research, and discusses the relevance of CFS-SO and CFS-UE as a practical tool for transportation planners, traffic managers, and fleet coordinators to understand and tune the impact of fleets on road networks.

2. Background and Related Work

In this work, we examine a mix of individual users and fleets on the network through the lens of static network equilibrium. The application of equilibrium theory to transportation networks is attributed to Wardrop (1952) who advocates for its use as a principled heuristic to estimate the impact of a road network improvement on the future distribution of traffic flow. If travelers are "user-optimal" decision makers (they choose the quickest route), we must contend with the fact that travel time both affects and is affected by traveler route choices. The framework proposed by Wardrop, and widely used by transportation planners to this day (Boyles, Lownes, and Unnikrishnan 2021), acknowledges and reconciles this circular dependency by finding the route choices and travel costs that are mutually consistent: an equilibrium. This is preferable, Wardrop argues, to the "arbitrary assumptions" one would otherwise have to use.

It is important to note that the use of "optimal" and "optimized" in the traffic network equilibrium literature is different from the way the terms are used in operations research. In network equilibrium, road users are infinitesimal units of flow each with a fixed origin and destination on the network. The only decision the road users make is which route they should use to travel from their origin to destination. In Wardrop's UE, road users select the route with minimal travel cost; they are said to be "user optimal" in the sense that minimizing their travel cost is solely beneficial to the user. When a network of user optimal road users achieve an equilibrium, it is referred to as a UE. In this paper we use the term "individual user" or "UE user" to refer to user optimal road users. We also study two other route choice principles: SO and FO. So-called SO users choose routes to minimize the total travel cost over all road users, and FO users choose routes to minimize the total travel cost of their own fleet. When a network of system optimal users reach an equilibrium, it is known as the SO flow. For an in-depth introduction to UE and SO, see Sheffi (1985) or Boyles, Lownes, and Unnikrishnan (2021).

In our setting, we are interested in the equilibrium achieved when individual users share the network with an optimized fleet. The formulation of network equilibrium for multiple classes of vehicles each with their own behavior (a multiclass or mixed equilibrium) was introduced by Dafermos (1972). Mixed equilibrium has historically been applied in cases where the vehicles within each class are still user optimal but use different notions of travel cost. It is not until Harker (1988) that coordination is introduced to the mixed equilibrium setting. Harker (1988) computes a mixed equilibrium on networks with both individuals and "Cournot-Nash" players who behave identically to our fleets and are routed to minimize a collective rather than individual travel cost. Harker (1988) specifically identifies "privatized urban mass transit" in addition to freight transportation as relevant domains for the application of this mixed equilibrium. Yang, Zhang, and Meng (2007) extend this analysis to examine a network with UE users, fleets, and SO users. Although Yang, Zhang, and Meng (2007) primarily serve to introduce the formulation and solution algorithm for this particular kind of mixed equilibrium, it contains an important empirical observation that sparked a continuing line of inquiry: When individuals control enough of the demand, neither SO nor fleet users may change the total system cost, and conversely, if there are few enough individuals on the network, a combination of fleets and SO users can achieve SO flow. To summarize our work as a single question, we ask the following: Under what conditions, if any, will a network achieve SO traffic flow

with a mixed equilibrium of fleet and individual users alone, *without* the aid of SO users?

There has recently been renewed interest in the line of literature established by Harker (1988) and Yang, Zhang, and Meng (2007), due largely to the advent of ad hoc fleets. Indeed, to realize the Cournot-Nash player of Harker (1988), the fleet must be able to coordinate itself to minimize collective cost. By comparison, there is extensive literature on ways to induce SO behavior by manipulating the individual notion of cost via infrastructure (e.g., ramp metering; Sheffi 1985). In our setting, coordination comes not from the physical infrastructure but from the digital infrastructure used by the service provider. In parallel work, Sharon et al. (2018) and Chen et al. (2020) identify a minimum-control ratio (MCR) that they define as the smallest volume of SO users capable of inducing SO flow on a network shared with UE users. In both cases, the fleet is centrally routed to minimize total system cost, corresponding to the second half of the observation of Yang, Zhang, and Meng (2007): that SO flow can be achieved even if not all users are SO users. Although the system as a whole is better off, the SO users themselves are relatively, if not absolutely, worse off precisely because they prioritized the travel cost of others over that of their own. We examine this idea more precisely in Section 4.3. This may not be realistic for private service providers who have no inherent altruistic motivation. In contrast, our work focuses on the interaction between fleets (Cournot-Nash players) and individual users, both of which are self-interested.

A second line of inquiry addresses the regulations required to realize the benefits of a centrally routed fleet. Zhang and Nie (2018) view the fleet as a direct government intervention and balance the benefits of a fleet of SO users against their deployment cost. In this view, the fleet is a "mobile actuator" (Wang et al. 2020) and becomes a traffic management tool. Another perspective is offered by Mansourianfar et al. (2021) and Delle Site (2021), who view the fleet as a third party that must be compensated or tolled to align their interests with that of the system's. Last, Mehr and Horowitz (2019) investigate the impact of platooning autonomous vehicles (AVs) on network equilibrium via their impact on road capacity. Similar to our work, they find that a mixed equilibrium of platooning AVs and individual users may increase total delay in the network.

Perhaps most closely related to our work, Cominetti, Correa, and Stier-Moses (2009) investigate the impact of a set of the Cournot-Nash players of Harker (1988) on the network but do not consider a mix of fleet and individuals. Our work explicitly considers both fleets and individual travelers coexisting on the network. We focus not only on when fleet optimal behavior breaks classical UE and increases system-level travel cost but also when it enables SO network flow, producing insights, and policy implications via application to real-world networks.

Separately, there has been substantial research into the congestion effects of the most prevalent example of ad hoc fleets on the road today: ride-sourcing fleets. The studies fall broadly into two categories: statistical analysis of ride-sourcing data and simulation-based studies. Among the reasons why ride-sourcing could affect congestion, Erhardt et al. (2019) identifies shared rides, integration with mass transit, and lower rates of car ownership as potentially beneficial and deadhead cruising, impeding traffic flow during pickups and dropoffs, and modal shift away from less congesting modes as potentially harmful. Several empirical studies, including Erhardt et al. (2019), have indicated that ride-sourcing increases congestion. Ward et al. (2019) and Hall, Palsson, and Price (2018) have further investigated effects of ride-sourcing fleets on public transit ridership, car ownership, and vehicle miles traveled. In contrast, simulation studies have offered ways in which ride-sourcing fleets could possibly be leveraged to reduce congestion. Fagnant and Kockelman (2014) simulate a fleet of shared autonomous vehicles (SAVs) to conclude that such a system would require substantially fewer vehicles on the road.

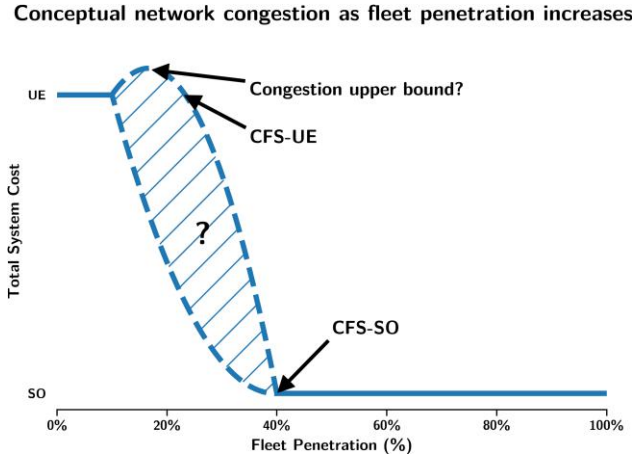
Our work contributes to the understanding of ride-sourcing congestion effects but differs substantially from prior work. In none of the prior work reviewed here could the identified congestion effects be plausibly tied to route choice. For example, the empty vehicle miles that arise as a side effect of ride-sourcing services have nothing to do with whether route choice is coordinated or not. Our work instead offers a complementary view of the effect of ad hoc fleets by isolating the effect that coordinated route choice might have on network congestion. Our work does not confirm nor refute prior work in the area; rather, it aims to expand our understanding of how coordinated fleets may impact network congestion. In prior work, it is the scale of the fleets that makes the congestion an issue: None of the congestion effects are unique to ride-sourcing services. In the same way, we ask that, when these fleets become large enough to coordinate their market power to extract lower average travel costs, will they contribute to or ease congestion?

This paper aims to fill in some key details of the conceptual diagram in Figure 1 and extends the literature in the following ways:

1. We demonstrate that there exist networks on which mixed equilibrium with fleets will increase system cost over UE. This is marked on the conceptual diagram as "congestion upper bound?." We leave open the question of whether there exists an upper bound, but if there is one it must be, at least for some networks, strictly greater than UE.

2. We show that when there is a single fleet, there must exist a fleet demand pattern for which the resultant mixed equilibrium is SO; in this work, we are concerned with the smallest fleet to induce SO ("CFS-SO" in Figure 1). We also show that there must exist a fleet

Figure 1. (Color online) Conceptual Diagram of System Cost at Mixed Equilibrium with One Fleet as Fleet Penetration, as a Percent of Total Demand, Increases



Notes. In this paper, we show fleets can worsen system cost relative to UE and that UE and SO may be achieved at certain penetration levels. Between these critical levels, however, it is unclear how total system cost behaves. Because demand is generally high-dimensional, this diagram does not directly map to general networks on which oftentimes CFS-UE is larger than CFS-SO.

demand pattern for which the resultant mixed equilibrium is UE; we are interested here in the largest fleet to induce UE ("CFS-UE" in Figure 1).

3. We formulate novel mathematical programs to find CFS-SO and CFS-UE. Exact and heuristic algorithms are developed to efficiently solve those programs in large-scale transportation networks.

4. We solve for CFS-SO and CFS-UE on two real-world networks, finding that not all vehicles need to participate in an optimized fleet for the system to attain its minimum system cost. Moreover, we show that such a fleet would need no external subsidy to benefit from optimizing itself.

5. We demonstrate the practical relevance of critical fleet size to regulators, transportation planners, and fleet coordinators.

3. Preliminaries

In this section, we define our notion of network equilibrium that we term "mixed equilibrium with fleets" and show that this mixed equilibrium has UE and SO as special cases.

We consider a road network represented as a graph $\mathcal{G}$ with nodes N and edges (or links) A. On this network, there is a volume of travelers, each seeking to travel from one node to another. We refer to the set of all such ordered node pairs as the set of OD pairs, $W \subseteq N \times N$. The travel demand is segmented into *flow classes*: the individual user flow class, denoted (typically as a superscript) by $\mathbf{u}$, and $k \geq 1$ fleet flow classes, denoted $\mathbf{c}_i$ for $i = 1, \ldots, k$. Although in this paper we will consider only one fleet, we define mixed equilibrium with multiple fleets here for completeness. The set of all fleet flow classes is written $\mathcal{F}$. The travel demand for each flow class is represented as a vector of OD travel volume: $\mathbf{q}^{\mathbf{u}} \in \mathbb{R}_+^{|W|}$ for the user flow class and $\mathbf{q}^{\mathbf{c}_i} \in \mathbb{R}_+^{|W|}$ for each fleet flow class $\mathbf{c}_i \in \mathcal{F}$. For each flow class and OD pair, travel is represented by the assignment of travel demand across the paths connecting the OD pair. The set of all paths on the network is denoted by P. This assignment may be represented using a trip path incidence matrix, $\mathbf{M} \in \{0,1\}^{|W| \times |P|}$, where the element at (w, p) is one if and only if path p starts and ends at the origin and destination, respectively, of the pair w. A user path assignment $\mathbf{f}^{\mathbf{u}} \in \mathbb{R}_+^{|P|}$ is feasible if $\mathbf{M}\mathbf{f}^{\mathbf{u}} = \mathbf{q}^{\mathbf{u}}$. Similarly for each fleet $\mathbf{c}_i \in \mathcal{F}$, the fleet path flow $\mathbf{f}^{\mathbf{c}_i} \in \mathbb{R}_+^{|P|}$ is feasible if $\mathbf{M}\mathbf{f}^{\mathbf{c}_i} = \mathbf{q}^{\mathbf{c}_i}$. The relation between path flow and link flow is represented using a link path incidence matrix, $\mathbf{D} \in \{0,1\}^{|A| \times |P|}$, where the element at (a, p) is one if an only if link a lies on path p. A user link assignment $\mathbf{x}^{\mathbf{u}} \in \mathbb{R}_+^{|A|}$ is feasible if there exists a feasible user path flow $\mathbf{f}^{\mathbf{u}}$ such that $\mathbf{D}\mathbf{f}^{\mathbf{u}} = \mathbf{x}^{\mathbf{u}}$. Similarly for each fleet $\mathbf{c}_i \in \mathcal{F}$, the fleet link flow $\mathbf{x}^{\mathbf{c}_i} \in \mathbb{R}_+^{|A|}$ is feasible if there exists a feasible fleet path flow $\mathbf{f}^{\mathbf{c}_i} \in \mathbb{R}_+^{|P|}$ such that $\mathbf{D}\mathbf{f}^{\mathbf{c}_i} = \mathbf{x}^{\mathbf{c}_i}$. Feasibility is always with respect to an OD demand vector.

Travelers incur a nonnegative travel cost on each link traversed represented as a link-separable monotone nondecreasing and differentiable function of aggregate link flow $\mathbf{t} : \mathbb{R}_+^{|A|} \to \mathbb{R}_+^{|A|}$.

We assume that individual users each wish to minimize the cost of their own travel, corresponding to Wardrop's first principle (Wardrop 1952). Each fleet, as the Cournot-Nash players in Harker (1988), is assumed to minimize the average travel cost over the fleet. We may now define mixed equilibrium with fleets.

Definition 1 (Mixed Equilibrium with Fleets). Let $\Omega^{\mathbf{u}}$ denote the set of $\mathbf{q}^{\mathbf{u}}$-feasible user link flows and $\Omega^{\mathbf{c}_i}$ for each $\mathbf{c}_i \in \mathcal{F}$ the set of $\mathbf{q}^{\mathbf{c}_i}$-feasible link flows for fleet $\mathbf{c}_i$. The tuple of feasible link flows $(\mathbf{x}^{\mathbf{u}*}, \mathbf{x}^{\mathbf{c}_1*}, \ldots, \mathbf{x}^{\mathbf{c}_k*})$ is a mixed equilibrium if the following holds:

$$\langle \mathbf{t}(\mathbf{x}^*), \mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{u}*} \rangle \geq 0 \quad \forall \mathbf{x}^{\mathbf{u}} \in \Omega^{\mathbf{u}}, \tag{1}$$

$$\langle \tilde{\mathbf{t}}(\mathbf{x}^*, \mathbf{x}^{\mathbf{c}_i*}), \mathbf{x}^{\mathbf{c}_i} - \mathbf{x}^{\mathbf{c}_i*} \rangle \geq 0 \quad \forall \mathbf{x}^{\mathbf{c}_i} \in \Omega^{\mathbf{c}_i} \quad \forall \mathbf{c}_i \in \mathcal{F}, \tag{2}$$

where $\mathbf{x}^* = \mathbf{x}^{\mathbf{u}*} + \sum_{\mathbf{c}_i \in \mathcal{F}} \mathbf{x}^{\mathbf{c}_i*}$ represents the aggregate link flow at the mixed equilibrium and $\tilde{\mathbf{t}}(\mathbf{x}, \mathbf{x}^{\mathbf{c}_i}) = \mathbf{t}(\mathbf{x}) + \mathbf{x}^{\mathbf{c}_i}\mathbf{t}'(\mathbf{x})$ represents the marginal cost of fleet travel (referred to as fleet marginal cost), and $\mathbf{t}'(\mathbf{x})$ is the element-wise derivative of the link cost function.

It can easily be seen via the Beckmann transformation (Beckmann, McGuire, and Winsten 1956, Sheffi 1985) that an equilibrium of the fleet marginal cost in (2) is equivalent to a minimization over the total fleet cost, $\mathbf{x}^{\mathbf{c}_i}\mathbf{t}(\mathbf{x})$. It is also useful to point out that the difference between our fleet users and the system optimal

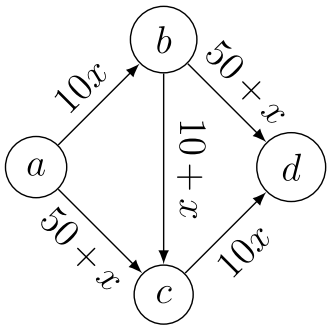
users in Yang, Zhang, and Meng (2007), Zhang and Nie (2018), Chen et al. (2020), and Sharon et al. (2018) is that the system optimal users seek to equalize the *system marginal cost*, expressed as $\mathbf{t}(\mathbf{x}) + \mathbf{x}\mathbf{t}'(\mathbf{x})$, or in our notation, $\tilde{\mathbf{t}}(\mathbf{x}, \mathbf{x})$.

From Definition 1 we immediately see that SO and UE are special cases of mixed equilibrium with fleets. In particular, if all fleet demand is zero, then mixed equilibrium is UE, and if one fleet accounts for all of the demand, then mixed equilibrium is SO.

We are interested in demonstrating that on some networks when neither individual users nor the fleet control all the demand, SO can be achieved in the aggregate flow. It is straightforward to imagine how a fleet might *reduce* congestion on the network: The fleet minimizes the total cost for a subset of the flow so one would hope that this effort also reduces travel cost for nonfleet users. We can generate an intuition for why we should expect a mix of selfish behaviors to *ever* induce SO by considering Yang, Zhang, and Meng (2007) in a simpler setting: adding SO users to aggregate UE on the Braess network. The Braess network (Sheffi 1985), shown in Figure 2, contains three paths: upper, lower, and shortcut. At UE, two units of flow use each path, at SO, and three units will take each of the upper and lower paths, with no flow on the shortcut. If we were to replace some UE demand with SO users, they would simply replace UE flow on the upper and lower paths; all remaining flow would still use the shortcut path. Therefore, despite the fact that the SO users are choosing routes to benefit the system (not themselves), they still choose paths they would have chosen as users. At small SO penetration, these two notions of cost are aligned: The least marginal cost paths are also least cost paths.

Now suppose we are at aggregate SO, and we replace some SO flow with a fleet. The total marginal cost and fleet marginal cost differ by the product of the user flow and the link cost derivatives. If this difference is uniform enough across paths at aggregate SO, then the two notions of cost may be aligned: Least fleet marginal cost paths are also least total marginal cost paths. We would then expect a fleet to make the same route choices that they would have made as SO flow, and therefore, the aggregate network state will not change when SO flow is replaced with fleet flow. If we can do the same with UE users, finding paths where cost and marginal cost align and replace SO users with UE users, and then UE users would similarly make the same decisions SO users would. In effect, we have partitioned the network at SO into paths for which cost and marginal cost are aligned and paths for which fleet marginal cost and marginal cost are aligned so that the combination of FO and UE behaviors in aggregate achieves SO on the network.

Figure 2. Braess Network, Annotated with Link Cost Functions



Notes. Six units of demand wish to travel from node a to node d. The three paths on this network are "upper" ($a \to b \to d$), "lower" ($a \to c \to d$) and "shortcut" ($a \to b \to c \to d$).

4. Fleet Optimality Paradox

In this section, we first analyze total system cost at mixed equilibrium in general. We then demonstrate a concrete example of an optimized fleet which, at mixed equilibrium, *increases* total system cost relative to UE. We refer to this phenomenon as the fleet optimality paradox.

4.1. Total Delay Under Mixed Equilibrium with Fleets

On the one hand, total delay at mixed equilibrium with fleets has a trivial but tight lower bound of the total delay at SO. Only under certain conditions, however, is the total travel time at UE an upper bound of mixed equilibrium with fleets. We give a sufficient condition in (3), the proof of which is provided in Online Appendix A:

$$\sum_{a\in A} \mathbf{x}_a t_a'(\mathbf{x}_a)(\mathbf{y}_a - \mathbf{x}_a) \geq \sum_{a\in A}\sum_{i\in\mathcal{F}} \mathbf{x}_a^{\mathbf{c}_i} t_a'(\mathbf{x}_a)(\mathbf{y}_a^{\mathbf{c}_i} - \mathbf{x}_a^{\mathbf{c}_i}), \quad (3)$$

where, $\mathbf{x}_a = \mathbf{x}_a^u + \sum_{i\in\mathcal{F}} \mathbf{x}_a^{\mathbf{c}_i}$ and $\mathbf{y}_a = \mathbf{y}_a^u + \sum_{i\in\mathcal{F}} \mathbf{y}_a^{\mathbf{c}_i}$ are the link flows at mixed and user equilibrium, respectively. This condition is not of practical use because it depends on the link flows at UE and mixed equilibrium: One could simply compare the total system costs. It does, however, yield some helpful theoretical insight.

The left-hand side of (3) is a first-order estimate of the change in total cost when the aggregate link flow shifts from mixed equilibrium to UE; the right-hand side is the first-order estimate of the change in total *fleet* cost when the network shifts *fleet* link flow from mixed equilibrium to UE. The coefficients, $\mathbf{x}_a t_a'(\mathbf{x}_a)$ and $\mathbf{x}_a^{\mathbf{c}_i} t_a'(\mathbf{x}_a)$, measure the impact of a given change in flow. Roughly speaking, (3) is fulfilled if links with high total impact that receive additional aggregate flow are also those with high *fleet* impact and receive additional fleet flow. When UE behavior shifts fleet flow onto high fleet impact links that are not also high total impact links, then the right-hand side can exceed the left, possibly by enough so that the total system cost at mixed equilibrium exceeds that at UE. This idea will be exploited in the following example.

4.2. Example of the Fleet Optimality Paradox

In this section, we will examine the network in Figure 3 as a concrete example of the fleet optimality paradox. Intuitively, the paradox arises when the fleet chooses a path with high system marginal cost but low *fleet* marginal cost. In the example network, this is caused by an imbalance of fleet flow between two alternative paths. The fleet only wishes to avoid interfering with other vehicles in the fleet and therefore does not consider the effect of its route choice on the individual users. As a result, the path with fewer fleet vehicles has low *fleet* marginal cost, although its many users induce a large system marginal cost, causing the total system cost to increase over UE.

Consider the network in Figure 3. The OD pairs are $a \to b, c \to d$, and, $e \to f$. The only real route choice in this network exists for users travelling from a to b. The other OD pairs have only one route available to them The links leaving a and entering b have zero cost to simplify the arithmetic, but this choice neither fundamentally changes the problem nor the paradox.

We define the link costs t for cd and ef as follows:

$$t_{cd}(x_{cd}) = 1 \cdot x_{cd} + 50, \tag{4}$$

$$t_{ef}(x_{ef}) = 10 \cdot x_{ef} + 0. \tag{5}$$

We consider the mixed equilibria resulting from the two OD demand scenarios shown in Table 1. The left-hand column under each OD pair in Table 1 present the OD demand for first scenario (UE); the right-hand columns (in gray) show the OD demand for the mixed equilibrium that produces the paradox, the fleet optimal scenario (FO). Total volume of vehicles between each OD pair, the aggregate demand, remains unchanged.

We summarize the UE and FO equilbria in Table 2.

At the UE link flow, given in the left half of Table 2, the link cost of ef is minimal: All demand from a to b selects the route through link ef. At the FO link flow, given in the right half of Table 2, the users and fleet traveling from a to b prefer different paths. The individual users still prefer the path through ef as it remains least cost. The fleet, however, prefers the path through cd because it is least *fleet marginal cost*. By selecting the least fleet marginal cost path, the fleet *reduces* its total cost relative to its total cost under UE as shown in Table 3 (bottom row, in bold); however, the total system cost increases from UE to FO (Table 2 bottom row, in bold). The fleet operator makes routing choices to minimize the travel delay impact only on members of its own fleet, regardless of the impact on other travelers. In this case, the fleet has so few vehicles on link cd that the fleet marginal cost of the path through cd is lower than the fleet marginal cost of the path through ef, where the fleet already has vehicles.

Figure 3. Fleet Optimality Paradox: An Illustrative Toy Network

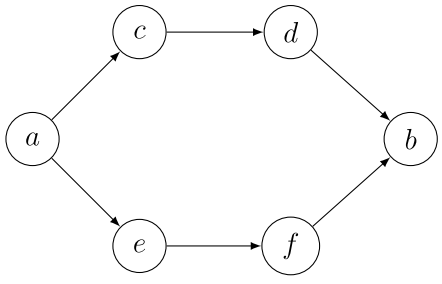


Table 1. Fleet, User, and, Aggregate Demand for the Fleet Optimality Paradox

	$c \to d$		$e \to f$		$a \to b$	
Fleet	0	1	0	2.75	0	0.05
User	13	12	2.75	0	1	0.95
Aggregate	13		2.75		1	

To see why this occurred, we note that not only is the path through ef user optimal, it is also system optimal, which means that the SO link flow and UE link flow are the same for the given aggregate travel demand on this network. What is then also true of this example is that the fleet is better off then it would have been at SO; this is related to a more general phenomenon we call the system optimal paradox discussed further in Section 4.3.

4.3. System Optimality Paradox

The fact that the fleet may actually reduce its total travel cost relative to SO in the fleet optimality paradox results from a property of SO traffic assignment that we call the system optimality paradox. For a network with nonunit price of anarchy, the following is true of SO traffic flow:

1. There is *always* volume that is *relatively* worse off compared with available paths, and
2. There is *possibly* volume that is *absolutely* worse off compared with the travel cost of the same OD pair at UE.

In short, when SO path flow is partitioned, the guarantees that SO provides for the network as a whole may not apply to each partition individually.

The first point follows directly from the UE principle that requires that all used paths between each OD pair have equal and minimal travel cost. Any traffic assignment that is not UE must therefore have at least one OD pair for which some volume occupies a path that is not least cost: This volume is worse off *relative* to some other available path.

The second point occurs only on some networks. On the Braess paradox network, for example, the cost of every used path at SO is strictly less than the cost of every used path at UE. Thus, although some users at SO are *relatively* worse off (there is a lesser cost path available), they are all better off than they would have been under UE. However, we can easily construct a network where this is not the case. A concrete example is given in

Table 2. Demonstration of the Fleet Optimality Paradox

	UE			FO						
	Aggregate (users)			Aggregate			User		Fleet	
	Flow	Cost	Total cost	Flow	Cost	Total cost	Flow	Total cost	Flow	Total cost
$c \to d$	13.00	63.0	819.000	13.05	63.05	822.8025	12.00	756.60	1.05	66.2025
$e \to f$	3.75	37.5	140.625	3.70	37.00	136.9000	0.95	35.15	2.75	101.7500
Total	16.75		**959.625**	16.75		**959.7025**	12.95	791.75	3.80	167.9525

Note. Bold indicates the outcome highlights.

Online Appendix B, but the essential intuition is that when UE is different from SO on a parallel network one path gains volume (and therefore travel cost) from the other in moving from UE to SO. In this setting, not only are some users not using the least cost path (they are relatively worse off), they are also experiencing higher absolute travel costs than they would have under UE (they are absolutely worse off). This solution is SO because the users that are better off (both relatively and absolutely) accrue total travel cost savings that exceed the increased travel cost of the worse off travelers so that on average the system is better off.

4.4. Concluding Remarks

Taken together, this section shows that, at extreme penetration levels of 0% and 100%, mixed equilibrium with fleets achieves a total system cost equal to the total system cost at UE and SO respectively. However, as penetration levels increase from 0% to 100%, total system cost will not necessarily monotonically decrease and may in fact become larger than the total system cost under UE. This has strong policy implications: Under certain demand and roadway network conditions, increasing fleet penetration (e.g., individuals shift to use a mobility service that is centrally planned by a private entity) may increase congestion. Only if the fleet penetration is adequately high would its presence be guaranteed to reduce network congestion, up to achieving SO. Transportation network companies, such as Uber/Lyft, who decide to optimize their fleet may paradoxically lead to more congestion in some areas in their initial development stage (when the penetration is low), even if they would have replaced some private driving trips and OD demand remains the same.

Table 3. Demonstration That Fleet Decreases its Total Cost at FO

		UE		FO	
OD pair	Fleet demand	OD cost	Total cost	OD cost	Total cost
$c \to d$	1.00	63.0	63.000	63.05	63.0500
$e \to f$	2.75	37.5	103.125	37.00	101.7500
$a \to b$	0.05	37.5	1.875	63.05	3.1525
Total	3.80		**168.000**		**167.9525**

Note. Bold indicates the outcome highlights.

5. Critical Fleet Size

In this section, we will examine how, in certain networks, SO can be achieved through mixed equilibrium where not all vehicles are fleet vehicles. The smallest fleet size for which mixed equilibrium achieves aggregate SO traffic assignment is expressed as a fraction of total demand and termed the critical fleet size for system optimum (CFS-SO). There is a complementary notion, termed the critical fleet size for user equilibrium (CFS-UE), which measures the largest fleet size for which mixed equilibrium achieves aggregate UE traffic assignment.

CFS-SO is bounded by zero and one; a value of one indicates that all vehicles *must* be in the fleet for mixed equilibrium to result in SO, and a value of zero indicates that no fleet vehicles are required. A network has a CFS-SO of zero if and only if its UE and SO assignments are the same; in other words, the network has a price of anarchy (Roughgarden 2005) of one. Similarly, CFS-UE is bounded by zero and one; a value of zero indicates that the presence of any fleet vehicles on the network "breaks" UE (induces a mixed equilibrium whose aggregate link flow is different from UE). A CFS-UE of one indicates that all vehicles may participate in the fleet without changing UE. A CFS-UE of one implies that UE and SO assignments are identical on this network, and as a result, CFS-UE is one if and only if CFS-SO is zero. Outside of this case, the relationship between CFS-SO and CFS-UE on a network is not at all clear and is left for future research.

CFS-UE is different from the notion of the smallest fleet for which mixed equilibrium is different from UE. As a somewhat counter-intuitive result with respect to the simplified one-dimensional Figure 1, CFS-UE can be larger than CFS-SO. In representing demand as a unidimensional quantity, Figure 1 is in a sense overly simplistic: When demand is higher dimensional, CFS-UE can occupy an entirely different set of OD pairs than CFS-SO and simply have larger magnitude. In one dimension, this is only possible when, as we discuss in the next paragraph, the price of anarchy on the network is one. The smallest fleet to "break" UE may also be of interest but is strictly a different question than the one CFS-UE seeks to answer.

The example network in Figure 3 on which we demonstrated the fleet optimality paradox also provides instructive examples of CFS-SO and CFS-UE. We noted

that the UE and SO traffic assignments are the same on the example network. As a result, CFS-SO is zero because users alone achieve SO, and CFS-UE is one because when all the volume on the network is an optimized fleet, the network is at SO, which is also UE. The paradox presented in Section 4 is thus also an example of the following three phenomena:

1. CFS-UE can be greater than CFS-SO,
2. Fleet OD demand that is element-wise greater than the fleet OD demand at CFS-SO is not necessarily SO, and,
3. Fleet OD demand that is element-wise less than the fleet OD demand at CFS-UE is not necessarily UE.

What CFS-SO and CFS-UE *do* imply, however, is that at mixed equilibrium with fleets,

1. SO cannot be achieved with a fleet smaller than CFS-SO, and
2. UE cannot be achieved with a fleet larger than CFS-UE.

5.1. Critical Fleet Size on a Parallel Network

In this section, we will analyze CFS-SO and CFS-UE on a parallel network. We derive analytical results that hold for all separable, monotonic increasing link performance functions on a parallel network. Although we should not expect these results to generalize, the parallel network is useful in developing an intuition for critical fleet size and for mixed equilibrium more generally. In particular, we are interested in whether critical fleet size results in paths, OD pairs, or entire networks that are exclusively fleet vehicles or individual users. In what follows, consider a parallel network with one OD pair connected by n parallel links.

5.1.1. Mixed Equilibrium Preserving UE Flow.

Let $\bar{\mathbf{x}}^{\mathbf{ue}} \in \mathbb{R}^n_+$ denote the UE link flow on the network. We wish to find the mixed equilibrium with the largest fleet share such that UE link flow is preserved in aggregate. Proof of Propositions 1–4 can be found in Online Appendix C.

Proposition 1. *Let* $(\mathbf{x}^{\mathbf{c}}, \mathbf{x}^{\mathbf{u}})$ *denote a mixed equilibrium on a parallel network whose aggregate link flow is the UE link flow (i.e.,* $\mathbf{x}^{\mathbf{c}} + \mathbf{x}^{\mathbf{u}} = \bar{\mathbf{x}}^{\mathbf{ue}}$*). If fleet demand is strictly positive then fleet flow on a link is positive if and only if aggregate link flow is positive. That is,*

$$\bar{\mathbf{x}}^{\mathbf{ue}}_a > 0 \Leftrightarrow \mathbf{x}^{\mathbf{c}}_a > 0 \quad \forall a \in A. \tag{6}$$

Proposition 2. *Given a fleet demand* $q^{\mathbf{c}} \in \mathbb{R}_+$ *such that* $0 < q^{\mathbf{c}} \le q$ *we can find the mixed equilibrium on the parallel network analytically:*

$$\tilde{t}^* = \frac{q^{\mathbf{c}}}{\sum_{a \in A_+} \frac{1}{\mathbf{t}'_a}} + t^*, \tag{7}$$

$$\mathbf{x}^{\mathbf{c}}_a = \frac{\tilde{t}^* - t^*}{\mathbf{t}'_a}, \tag{8}$$

provided that the following holds:

$$\tilde{t}^* \le \min_{a \in A} \mathbf{t}_a + \bar{\mathbf{x}}^{\mathbf{ue}}_a \mathbf{t}'_a, \tag{9}$$

where A_+ *denotes the set of all links where aggregate flow is strictly positive,* $\mathbf{t}_a$ *is the link cost on link* $a \in A$ *evaluated at mixed equilibrium,* $\mathbf{t}'_a$ *is the derivative of the link cost on link* $a \in A$ *evaluated at mixed equilibrium,* $\tilde{t}^*$ *is the minimum fleet marginal link cost at mixed equilibrium, and* t^* *is the minimum link cost at mixed equilibrium.*

Proposition 3. *The largest fleet demand that preserves UE is given by,*

$$\sum_{a \in A_+} \frac{\tilde{t}^* - t^*}{\mathbf{t}'_a} = \sum_{a \in A_+} \mathbf{x}^{\mathbf{c}}_a = q^{\mathbf{c}}, \tag{10}$$

where $\tilde{t}^*$ *satisfies*

$$\tilde{t}^* = \min_{a \in A} \mathbf{t}_a + \bar{\mathbf{x}}^{\mathbf{ue}}_a \mathbf{t}'_a. \tag{11}$$

5.1.2. Mixed Equilibrium Preserving SO Flow.

We now wish to find the smallest fleet demand such that SO is preserved at mixed equilibrium.

Proposition 4. *The minimum fleet demand required to induce SO flow on a parallel network with aggregate demand* q *is either*

1. *Zero if the UE flow is the same as the SO flow; or*
2. q *if UE flow is different from SO flow (CFS-SO = 1).*

In short, Proposition 4 ensures that for any parallel network, CFS-SO is either zero or one.

5.1.3. Extending to a General Network.

Although the parallel network provides a useful demonstration of fleet optimal mixed equilibrium, Proposition 4 ensures that it will not be an interesting one. In general networks, any value of CFS-SO is possible. Perhaps the simplest, albeit unsatisfying, way to demonstrate this fact is to consider a network composed of two parallel networks one of which is entirely fleet and the other, entirely individual users. Any value of critical fleet size can be achieved by varying the demand on the two subnetworks.

Nevertheless, the parallel network does provide us some valuable intuition for CFS on general networks. Key to the proof of Proposition 4 is the realization that if the set of least cost and least marginal cost paths are not the same, the fleet will need to fill those paths that are least marginal cost but not least cost. However, to maintain fleet optimal assignment, the fleet marginal cost must be equalized over all paths in use by the fleet, which in turn requires each path to be filled by the fleet. In a general network, because many path flows may induce the same aggregate link flow, it is not necessarily the case that user flow removed from a path must be replaced by fleet flow on *that specific path*; rather, the fleet

may occupy a completely different set of (fleet-optimal) paths that still induces the same aggregate link flow. It is for this reason that we should not expect general networks, even at the OD level, to only have the two extremal values of critical fleet size.

In total, it is possible, but difficult to imagine, that OD pairs may not be exclusively user or exclusively fleet at critical fleet size. Precise conditions under which paths, OD pairs, or entire networks are exclusively user or fleet are left for subsequent work. In this paper, we will attempt to shed light on the question of exclusivity empirically on real world networks in Section 6.

5.2. Critical Fleet Size Formulation and Solution Algorithm

In this section, we formulate critical fleet size as a mathematical program with equilibrium constraints (MPEC) and develop efficient ways to solve it in practice.

5.2.1. Problem Formulation. The fleet vehicles we are examining wish to minimize the fleet travel time: $\mathbf{z}(\mathbf{x^c}, \mathbf{x^u}) = \sum_a \mathbf{x}_a^{\mathbf{c}} \mathbf{t}_a(\mathbf{x}_a^{\mathbf{u}} + \mathbf{x}_a^{\mathbf{c}})$. Given fleet OD demand $\mathbf{q^c}$ and the link flow of the individual users $\mathbf{x^u}$, the fleet path (link) flows may be computed as the solution to the following mathematical program (Sheffi 1985):

$$\mathrm{FO}(\mathbf{x^u}, \mathbf{q^c}) = \arg\min_{\mathbf{f^c}} \ \mathbf{z}(\mathbf{x^c}, \mathbf{x^u}) \tag{12a}$$

$$\text{s.t.} \sum_{p \in P_w} \mathbf{f}_p^{\mathbf{c}} = \mathbf{q}_w^{\mathbf{c}} \quad \forall w \in W, \tag{12b}$$

$$\mathbf{f}_p^{\mathbf{c}} \geq 0 \quad \forall p \in P_w, w \in W, \tag{12c}$$

$$\sum_{p \in P} \mathbf{f}_p^{\mathbf{c}} D_{a,p} = \mathbf{x}_a^{\mathbf{c}} \quad \forall a \in A. \tag{12d}$$

The program is parameterized by the user link flow and the fleet demand, which are considered constant within in this program.

We now consider the critical fleet size problem that imposes the external constraint that aggregate link flow at mixed equilibrium must match a given aggregate link flow. For CFS-SO, that aggregate link flow, denoted $\bar{\mathbf{x}}^{\mathbf{so}}$, is link flow at SO. CFS-SO is expressed as the solution to the mathematical program with equilibrium constraints (MPEC) given by Program (13), and CFS-UE is expressed as the solution to the MPEC given by Program (14), with $\bar{\mathbf{x}}^{\mathbf{ue}}$ as the UE link flow.

$$\min_{\mathbf{f^u}} \sum_{w \in W} \mathbf{q}_w^{\mathbf{c}} \tag{13a}$$

$$\text{s.t.} \sum_{p \in P_w^*} \mathbf{f}_p^{\mathbf{u}} = \mathbf{q}_w^{\mathbf{u}} \ \ \forall w \in W \tag{13b}$$

$$\mathbf{f}_p^{\mathbf{u}} \geq 0 \quad \forall p \in P \tag{13c}$$

$$\sum_{p \in P^*} \mathbf{f}_p^{\mathbf{u}} \ \delta_{a,p} = \mathbf{x}_a^{\mathbf{u}} \quad \forall a \in A \tag{13d}$$

$$\mathbf{x^c} = \mathbf{D}\mathrm{FO}(\mathbf{x^u}, \mathbf{q^c}) \tag{13e}$$

$$\mathbf{x^c} + \mathbf{x^u} = \bar{\mathbf{x}}^{\mathbf{so}} \tag{13f}$$

$$\mathbf{q^c} + \mathbf{q^u} = \bar{\mathbf{q}} \tag{13g}$$

$$\max_{\mathbf{f^u}} \sum_{w \in W} \mathbf{q}_w^{\mathbf{c}} \tag{14a}$$

$$\text{s.t.} \sum_{p \in P_w^*} \mathbf{f}_p^{\mathbf{u}} = \mathbf{q}_w^{\mathbf{u}} \ \ \forall w \in W \tag{14b}$$

$$\mathbf{f}_p^{\mathbf{u}} \geq 0 \ \ \forall p \in P \tag{14c}$$

$$\sum_{p \in P^*} \mathbf{f}_p^{\mathbf{u}} \delta_{a,p} = \mathbf{x}_a^{\mathbf{u}} \ \ \forall a \in A \tag{14d}$$

$$\mathbf{x^c} = \mathbf{D}\mathrm{FO}(\mathbf{x^u}, \mathbf{q^c}) \tag{14e}$$

$$\mathbf{x^c} + \mathbf{x^u} = \bar{\mathbf{x}}^{\mathbf{ue}} \tag{14f}$$

$$\mathbf{q^c} + \mathbf{q^u} = \bar{\mathbf{q}} \tag{14g}$$

Because the aggregate link flow is fixed under UE or SO and the link performance function depends only on the aggregate link flow, the travel cost (and marginal travel cost) for each link is a constant with respect to both the user and fleet link flows. As a result, the set of least cost paths between each OD pair, P^*, is fixed and known a priori. Constraints (13d) and (14d) therefore are sufficient to describe the UE condition in their respective programs completely: Individual users may only use least cost paths.

5.2.2. Heuristic Solution Algorithm via Sensitivity Analysis. Typically, MPECs are solved via a huersitic solution methodology known as sensitivity analysis (Tobin and Friesz 1988). However, for this specific MPEC, sensitivity analysis is not the best choice. In this section, we outline how one would apply sensitivity analysis before identifying its limitations in this setting. A detailed treatment of our sensitivity analysis approach is given in Online Appendix D.

Sensitivity analysis is typically applied to an MPEC for which the objective depends on the network at equilibrium and the network equilibrium, in turn, is affected by the decision variables. Sensitivity analysis provides the gradient of the equilibrium link flow with respect to the decision variables and is used to compute the descent (ascent) direction. In Programs (13) and (14), the objective is an input of the equilibrium operator rather than its output. To apply sensitivity analysis, we solve the programs in their augmented Lagrangian forms (Hestenes 1969), treating the equilibrium constraint as a penalty term. The algorithm, given by Algorithms 2 and 3, alternates between descent steps on the augmented lagrangian via sensitivity analysis and sequential linear programming (Wright, Nocedal, and Wright 1999) known as the primal update, and ascent steps for the penalty weights, known as the dual updates.

The main theoretical deficit of solving CFS as an MPEC via sensitivity analysis is that the MPEC itself is nonconvex, and, as such, the solution algorithm is guaranteed

only to return a local optimal solution. The main practical limitation is that this particular application of sensitivity analysis requires many more equilibrium computations than a typical application: once every iteration within each primal update that itself is simply one descent step. This can take significant amounts of time. Taken together, on neither practical nor theoretical grounds does solving CFS via sensitivity analysis offer benefits over our methodology.

5.2.3. Exact Solution Algorithm via Mixed Integer Programming. As an alternative to sensitivity analysis, we can leverage the structure of the problem to develop an exact solution methodology. Conceptually our approach reverses the order that sensitivity analysis imposes on the solution algorithm: Instead of searching the space of mixed equilibria for those that happen to meet an aggregate link flow constraint, we search the space of feasible user-fleet link (path) flow partitions that satisfy the conditions of mixed equilibrium. This approach takes advantage of the problem structure to not only guarantee a global optimum solution but provide upper and lower bounds on the optimal value. In comparison with the previous sensitivity analysis approach, this approach avoids computing mixed equilibria entirely and instead forms mathematical programs to descend directly in the space of feasible equilibria. In particular, the constraint that aggregate link flow for CFS-SO is fixed to be the aggregate SO link flow and the aggregate link flow for CFS-UE is fixed to be aggregate UE link flow implies two important facts:

1. The set of paths the fleet *could possibly use* is known, finite, and substantially smaller than set set of all (simple) paths, and,

2. The fleet marginal link (path) cost is linear in fleet link (path) flow.

First for fact 2, when we consider the aggregate link flow fixed, the fleet marginal link cost is a linear function of fleet link flow: $\tilde{\mathbf{t}}_a(\mathbf{x}^{\mathbf{c}}_a) = \mathbf{t}_a + \mathbf{x}^{\mathbf{c}}_a \cdot \mathbf{t}'_a$. If we consider the fleet marginal path cost as a vector and denote by $diag(\mathbf{t}')$ the $|A| \times |A|$ diagonal matrix with the elements $\mathbf{t}'_a$ on the diagonal, then we may write the vector of fleet marginal path costs as a linear function of fleet flow $\mathbf{f}^{\mathbf{c}}$: $\tilde{\mathbf{c}}(\mathbf{f}^{\mathbf{c}}) = \mathbf{c} + \mathbf{D}^T \text{diag}(\mathbf{t}')\mathbf{D}\mathbf{f}^{\mathbf{c}}$, or equivalently as a function of user flow $\mathbf{f}^{\mathbf{u}}$, $\tilde{\mathbf{c}}(\mathbf{f}^{\mathbf{u}}) = \dot{\mathbf{c}} - \mathbf{D}^T \text{diag}(\mathbf{t}')\mathbf{D}\mathbf{f}^{\mathbf{u}}$, where $\mathbf{c}$ and $\dot{\mathbf{c}}$ represent the path cost and marginal path cost at the given aggregate link flow.

From a computational perspective, the form of fleet marginal cost that depends on the user path flow is attractive because the user path flow vector will generally be sparse. By the equilibrium principle, we know that individual users may only take least cost paths, so we need only specify user path flows for paths that are least cost at SO, a path set that, in practice, we expect to be known before solving CFS and substantially smaller than the set of all simple paths.

Next is fact 1. Because aggregate link flow is fixed, it is also true that the set of least marginal cost paths $\tilde{P}$, the set of usable paths for SO users, is fixed. When aggregate flow is SO any path flow vector that satisfies both the demand conservation constraint and produces the SO link flow must only use paths in $\tilde{P}$. As a result, at CFS-SO, the fleet may use only paths in $\tilde{P}$. For CFS-UE where the aggregate link flow at mixed equilibrium is the aggregate UE link flow, a similar argument applies: The fleet may only use least cost paths at UE: P^*.

Combining these two facts, we can rewrite the CFS-SO and CFS-UE MPECs as a mixed integer program (MIP) in Programs (15) and (16).

$$\min_{\mathbf{f}^{\mathbf{u}} \geq 0, \mathbf{f}^{\mathbf{c}} \geq 0, \lambda, z} \sum_{p \in \tilde{P}} \mathbf{f}^{\mathbf{c}}_p \tag{15a}$$

$$\text{s.t.} \quad \tilde{\mathbf{c}}_r(\mathbf{f}^{\mathbf{u}}) \geq \lambda_w \quad \forall r \in P_w \ \forall w \in W \tag{15b}$$

$$\tilde{\mathbf{c}}_r(\mathbf{f}^{\mathbf{u}}) \leq \lambda_w + m_1 \cdot (1 - z_r) \quad \forall r \in \tilde{P}_w \ \forall w \in W \tag{15c}$$

$$\mathbf{f}^{\mathbf{c}}_r \leq m_2 \cdot z_r \quad \forall r \in \tilde{P}_w \ \forall w \in W \tag{15d}$$

$$\mathbf{D}^* \mathbf{f}^{\mathbf{u}} + \tilde{\mathbf{D}} \mathbf{f}^{\mathbf{c}} = \bar{\mathbf{x}}^{\mathbf{so}} \tag{15e}$$

$$\mathbf{M}^* \mathbf{f}^{\mathbf{u}} + \tilde{\mathbf{M}} \mathbf{f}^{\mathbf{c}} = \bar{\mathbf{q}}^{\mathbf{so}} \tag{15f}$$

$$z_k \in \{0, 1\} \quad \forall k \in \tilde{P}_w \ \forall w \in W \tag{15g}$$

Here, m_1 and m_2 are large constants, $\tilde{P}$ and P^* are the set of least marginal cost and least cost paths at SO, respectively, $\tilde{\mathbf{D}}$ and $\mathbf{D}^*$ are the link path incidence matrices restricted to paths in $\tilde{P}$ and P^*, respectively, and $\tilde{\mathbf{M}}$ and $\mathbf{M}^*$ are the OD path incidence matrices restricted to paths in $\tilde{P}$ and P^*, respectively. The binary variable z_k encodes whether the path k is use by the fleet (one) or not (zero). The first two constraints ensure that the fleet flow is fleet optimal: The fleet marginal cost on all used paths is equal and minimal. MPECs may always be written and solved as mixed integer programs; this is not often done in practice because the number of integer variables scales with the number of paths in the network, and the cost function is typically nonlinear. In our problem, the cost function is linear, and the number of integer variables is substantially smaller than the number of paths.

The CFS-UE MIP is constructed analogously, notably as a maximization and using the set of least cost paths at UE where the CFS-SO program uses least marginal cost paths at SO.

$$\max_{\mathbf{f}^{\mathbf{u}} \geq 0, \mathbf{f}^{\mathbf{c}} \geq 0, \lambda, z} \sum_{p \in P^*} \mathbf{f}^{\mathbf{c}}_p \tag{16a}$$

$$\text{s.t.} \quad \tilde{\mathbf{c}}_r(\mathbf{f}^{\mathbf{u}}) \geq \lambda_w \quad \forall r \in P_w \ \forall w \in W \tag{16b}$$

$$\tilde{\mathbf{c}}_r(\mathbf{f}^{\mathbf{u}}) \leq \lambda_w + m_1 \cdot (1 - z_r) \quad \forall r \in P^*_w \ \forall w \in W \tag{16c}$$

$$\mathbf{f}^{\mathbf{c}}_r \leq m_2 \cdot z_r \quad \forall r \in P^*_w \ \forall w \in W \tag{16d}$$

$$\mathbf{D}^* (\mathbf{f}^{\mathbf{u}} + \mathbf{f}^{\mathbf{c}}) = \bar{\mathbf{x}}^{\mathbf{so}} \tag{16e}$$

$$\mathbf{M}^* (\mathbf{f}^{\mathbf{u}} + \mathbf{f}^{\mathbf{c}}) = \bar{\mathbf{q}}^{\mathbf{so}} \tag{16f}$$

$$z_k \in \{0, 1\} \quad \forall k \in P^*_w \ \forall w \in W \tag{16g}$$

Mathematical programs for CFS-SO and CFS-UE with multiple fleets are discussed in Online Appendix G.

5.3. Bounds of Critical Fleet Size

Although the MIP formulations of CFS-SO and CFS-UE offer substantial computational efficiencies relative to MIP formulations of general MPECs, they can still present challenges on large networks. For such cases, we present linear program relaxations of the CFS-SO and CFS-UE MIPs. The LP relaxation of the CFS-SO MIP is given by Program (17) and the LP relaxation of the CFS-UE MIP is given by Program (18).

$$\min_{\mathbf{f^u}\geq 0,\,\mathbf{f^c}\geq 0,\,\lambda} \sum_{p\in\tilde{P}} \mathbf{f}^{\mathbf{c}}_p \tag{17a}$$

$$\text{s.t.} \quad \tilde{\mathbf{c}}_r(\mathbf{f^u}) = \lambda_w \quad \forall r \in \tilde{P}_w \quad \forall w \in W \tag{17b}$$

$$\tilde{\mathbf{c}}_r(\mathbf{f^u}) \geq \lambda_w \quad \forall r \in P_w \quad \forall w \in W \tag{17c}$$

$$\mathbf{D}^*\mathbf{f^u} + \tilde{\mathbf{D}}\mathbf{f^c} = \bar{\mathbf{x}}^{\mathbf{so}} \tag{17d}$$

$$\mathbf{M}^*\mathbf{f^u} + \tilde{\mathbf{M}}\mathbf{f^c} = \bar{\mathbf{q}} \tag{17e}$$

$$\max_{\mathbf{f^u}\geq 0,\,\mathbf{f^c}\geq 0,\,\lambda} \sum_{p\in\tilde{P}} \mathbf{f}^{\mathbf{c}}_p \tag{18a}$$

$$\text{s.t.} \quad \tilde{\mathbf{c}}_r(\mathbf{f^u}) = \lambda_w \quad \forall r \in P^*_w \quad \forall w \in W \tag{18b}$$

$$\tilde{\mathbf{c}}_r(\mathbf{f^u}) \geq \lambda_w \quad \forall r \in P_w \quad \forall w \in W \tag{18c}$$

$$\mathbf{D}^*(\mathbf{f^u} + \mathbf{f^c}) = \bar{\mathbf{x}}^{\mathbf{ue}} \tag{18d}$$

$$\mathbf{M}^*(\mathbf{f^u} + \mathbf{f^c}) = \bar{\mathbf{q}} \tag{18e}$$

Formulating the critical fleet size in this way allows us to relate it directly to the minimum control ratio (MCR) introduced by Chen et al. (2020) and Sharon et al. (2018). In particular, because fleet vehicles must use a subset of the SO paths, CFS-SO is bounded below by the MCR, the proof of which is given in Online Appendix E.

Proposition 5 (Bounding Critical Fleet Size via Linear Programs). *The heuristic linear programs for CFS-SO and CFS-UE provide the following bounds:*

1. *Linear Program* (17) *provides an upper bound of CFS-SO.*
2. *Linear Program* (18) *provides a lower bound of CFS-UE.*

The proof is given in Online Appendix F.

It is worth noting that the solution returned by the linear program will typically have paths that are unused by the fleet but remain least fleet marginal cost, known in the literature as a *degenerate* equilibrium. Degeneracy occurs in the LP because the set of paths over which fleet marginal cost must be equal and minimal is fixed. Allowing the fleet to abandon unused paths, thereby releasing them from the constraint that their fleet marginal cost is minimal, could allow for a smaller fleet. As a result, it may be possible to refine a bound returned by the LP by iteratively removing unused paths and resolving the LP. Our preliminary work into such a procedure suggests that only modest gains can be achieved without a careful heuristic to select the paths to abandon each iteration. Broadly, the path abandonment procedure can be viewed as a member of the family of MIP heuristics. It is a very simple heuristic that always preserves feasibility by recognizing the structure of the problem but does not form an overall scheme that will find the global optimum. In contrast, MIP methods and heuristics available via state-of-the-art solvers do not recognize the structure of the problem but will progress toward the global optimum. A better option which we leave for future work would be to leverage path abandonment within the MIP solver.

5.4. Issue of an Unknown Path Set

In both the MIP formulation and their LP relaxations, it is assumed that we have access to the full path set, in addition to the least cost and least marginal cost path sets at UE or SO. The latter two sets are not difficult to obtain (see Online Appendix H). However, the full set of simple paths may not be feasible to enumerate. With only a subset of the paths, we cannot guarantee that the fleet is using paths that are minimal fleet marginal cost. To circumvent this difficulty, a column-generating approach is presented in Algorithm 1. Each iteration, the program is solved with an approximate path set $\hat{P}$, and least fleet marginal paths are computed at the candidate solution. If the least fleet marginal cost paths are not in the set of known paths, the solution assigned fleet flow to paths that are not in the fleet-usable path set. These paths are then added to the known path set, and the updated linear program is resolved. On resolving the linear program, the fleet marginal cost of these paths are constrained to be no smaller than λ_w, which is the least fleet marginal cost over all paths connecting the OD pair.

Algorithm 1 (Critical Fleet Size Solution Algorithm with Column Generation)

procedure CRITICALFLEETSIZE($\mathcal{G}, \bar{\mathbf{q}}$)
  $\bar{\mathbf{x}}^{\mathbf{so}} \leftarrow$ SystemOptimalLinkFlow($\mathcal{G}, \bar{\mathbf{q}}$)
  $P^*, \tilde{P} \leftarrow$ LeastCostPaths($\mathcal{G}, \bar{\mathbf{x}}^{\mathbf{so}}$), LeastMarginal CostPaths ($\mathcal{G}, \bar{\mathbf{x}}^{\mathbf{so}}$)
  $\hat{P} \leftarrow P^* \bigcup \tilde{P}$ ▷ Initialize the *known* path set
  while true **do**
    $\mathbf{f^u}, \mathbf{f^c}, \lambda \leftarrow$ Solve Program (15), (14), (17), or (18) using $\hat{P}$ as the path set P
    $P_k \leftarrow$ LeastFleetMarginalPaths($\mathcal{G}, \bar{\mathbf{x}}, \mathbf{x^c}$)
    if $P_k \cap \tilde{P} \neq \emptyset$ **then** ▷ A new path was discovered.
      $\hat{P} \leftarrow \hat{P} \bigcup (P_k \cap \tilde{P})$
    else
      return $\mathbf{f^u}, \mathbf{f^c}$
    end if
  end while
end procedure

5.5. Issue of Approximate SO Flow

Real-world networks present two practical challenges to Algorithm 1, using either the MIP or LP version. First, the programs scale linearly in the number of paths and OD pairs. Even moderately sized real-world networks will produce mixed integer programs too large for state-of-the-art solvers to solve in a reasonable amount of time. The MIP may simply be run with a fixed computational budget, or their LP relaxations may be used instead.

A second problem facing real-world networks has to do with the accuracy of the SO link flow. In both the heuristic LPs and the MIPs, if we were to fix the user path flow to zero, the program should recover an SO path flow as the fleet path flow variable. However, when we have access only to an approximation of the SO link flow, this guarantee no longer holds. In fact, the opposite is true: Because the approximate link flow is not exactly SO, there must be flow on paths that are not least marginal cost. As a result, when directly implemented with an approximate SO link flow, Programs (17) and (15) may be infeasible.

To overcome this second challenge, we use a modified version of Program (17) given by Program (19). In this modified program, the set of least marginal cost paths (and least cost paths for individual users) is constructed as the set of ϵ-least (marginal) cost paths: A path is least (marginal) cost if it is less than or equal to $(1+\epsilon)$ times the least marginal cost path for its OD pair. Correspondingly, instead of treating the link flow recovery as a constraint, it is treated as a penalty term in the objective weighted by $\beta > 0$. Larger penalty weight represent preference for accurately approximating the target aggregate link flow. This same modification is applied to CFS-SO LP and MIP, CFS-UE LP, and MIP and to the MCR.

$$\min_{\mathbf{f^u}\geq 0,\mathbf{f^c}\geq 0,\lambda} \sum_{p\in P^*} \mathbf{f}^{\mathbf{c}}_p + \beta \|\mathbf{D}^*\mathbf{f^u} + \tilde{\mathbf{D}}\mathbf{f^c} - \bar{\mathbf{x}}^{\mathbf{so}}\|_2 \tag{19a}$$

$$\text{s.t.} \quad \tilde{\mathbf{c}}_r(\mathbf{f^u}) \leq \lambda_w(1+\epsilon) \;\; \forall r \in \tilde{P}_w \; \forall w \in W \tag{19b}$$

$$\tilde{\mathbf{c}}_r(\mathbf{f^u}) \geq \lambda_w \;\; \forall r \in P_w \; \forall w \in W \tag{19c}$$

$$\mathbf{M}^*\mathbf{f^u} + \tilde{\mathbf{M}}\mathbf{f^c} = \bar{\mathbf{q}}^{\mathbf{so}} \tag{19d}$$

Generally speaking, the more accurate the SO solution, the larger the penalty weight β and the smaller the path cost tolerance ϵ should be. Intuitively, a high-quality SO solution should produce a link flow that is close to the true SO link flow and path costs that are close to the true path costs. Our mixed equilibrium then should not be allowed to stray too far from the aggregate link flow, and the usable paths should be near least (marginal) cost. However, exactly which value is best and how they relate to common convergence criteria are not immediately clear. To best compare our results to existing work, we set our hyperparameters as the most stringent values (lowest ϵ, highest β) to closely reproduce the Sioux Falls minimum control ratio as reported in Chen et al. (2020). This heuristic resulted in $\beta = 10$ and $\epsilon = 5\text{e}-4$. For the Pittsburgh network, $\beta = 0.1$ and $\epsilon = 1\text{e}-8$; the least ϵ and largest β to bring its MCR close to range reported for similar network in Chen et al. (2020). The choice of hyperparameters also affects the solving time of the programs. Smaller values of ϵ result in smaller usable path sets reducing the number of variables and significantly reducing solve time.

Here too, network size presents some difficulty. From the perspective of Program (19), a high-quality SO approximation will assign a large fraction of flow to paths that are close to least marginal cost. Common equilibrium convergence criteria, such as relative gap or average excess cost (Boyles, Lownes, and Unnikrishnan 2021), do not capture this notion, making it difficult to know a priori whether an approximation is good enough or how to set β and ϵ to compensate. This fact is not unique to our problem setting and is a challenge to any analysis that takes an equilibrium traffic assignment as input.

Taken together, these modifications allow us to use approximate SO (UE) link flow to compute an upper (lower) bound on CFS-SO (CFS-UE).

6. Experiments

Our experiments were performed on three networks. The Braess network (Sheffi 1985), shown in Figure 2, served to validate our methodology and code. The Sioux Falls network (Stabler, Bar-Gera, and Sall 2021) with 76 links and 528 OD pairs serves as a simple small example and as a point of comparison with previous work. The Pittsburgh network with 5,449 links and 4,881 OD pairs serves as a real-world example network. The Pittsburgh network contains the core urban areas of Pittsburgh, PA, extracted from the Southwestern Pennsylvania Commission's (SPC) road network (www.spcregion.org).

In our experiments, we use the CVXPY modeling language (Diamond and Boyd 2016, Agrawal et al. 2018) with GUROBI v9.5.0 as the backend solver. All experiments were performed on an Ubuntu 20.04 Linux machine with 16 GB RAM and an Intel i7-7700HQ processor. We report the critical fleet size results in Table 4 and devote the remainder of the section to our findings on these networks. The MIPs for CFS-SO and CFS-UE on the Sioux Falls network were run for 24 and 2 hours respectively; we also report the relative gap of the solution reported by the solver.

From Chen et al. (2020), we know that the minimum control ratio for the Braess network is one, so we know that the CFS-SO must also be one, a finding reproduced by our methodology. CFS-UE, on the other hand, is achieved by a fleet that uses two of the three paths exclusively, leaving the remaining path for the users. Because the CFS-UE LP, unlike the MIP, cannot abandon paths, it provides a very loose bound.

Table 4. Comparison of Critical Fleet Size on Experiment Networks

		CFS-SO			CFS-UE		
Network	MCR	LP	MIP	(Relative gap)	LP	MIP	(Relative gap)
Braess	100%	100%	100%	(0%)	0.34%	66.67%	(0%)
Sioux Falls	14.12%	36.53%	33.03%	(25.28%)	46.12%	78.85%	(15.48%)
Pittsburgh	30.69%	83.04%	—	—	56.92%	—	—

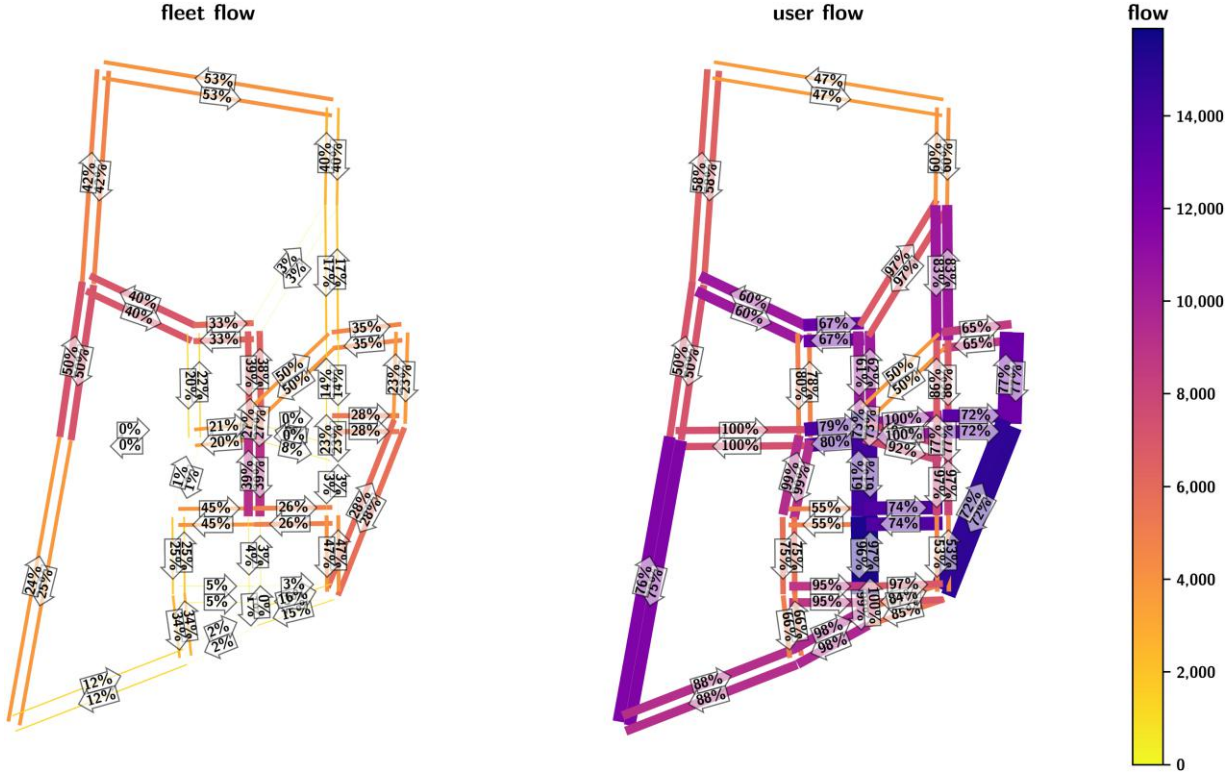
6.1. Link Flow Patterns at CFS-SO

The two immediate observations are that CFS-SO is highly variable across the two networks and is substantially higher than the MCR. In the Pittsburgh network, at most 83% of vehicles must be in the fleet to achieve SO traffic flow compared with 33% in the Sioux Falls network. However, that the presence of a self-interested fleet can give rise to SO flow on real-world networks without controlling all the vehicles at all is remarkable in and of itself.

Geographically, CFS-SO on both Sioux Falls and Pittsburgh produces fleet presence across the network. We show the user and fleet link flows at CFS-SO for Sioux Falls in Figure 4 and for Pittsburgh in Figure 5. At least part of the reason for this is that fleet presence on one OD pair influences the fleet marginal cost on many others, particularly in dense networks such as these where paths for different OD pairs are highly intertwined. As a result, fleet presence in one area of the network tends to induce fleet presence on at least some of the OD pairs that share links with it, leading to a cascading effect over the entire network. This effect may be contrasted with the minimum control ratio that treats OD pairs in isolation resulting in areas of the network with no fleet presence whatsoever.

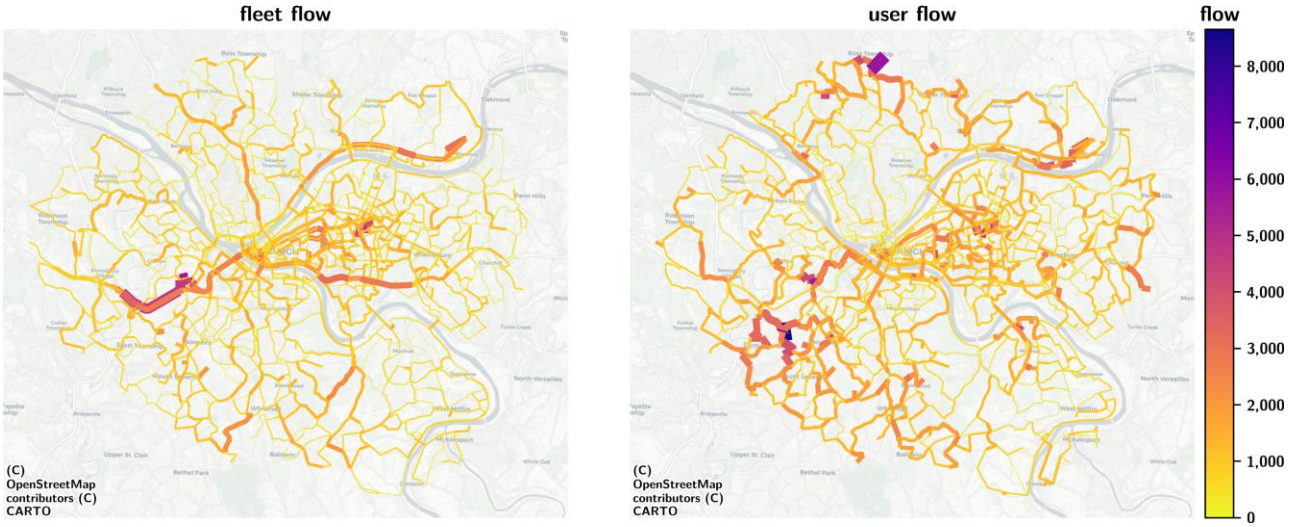
Nevertheless, fleet vehicles in Pittsburgh appear to accumulate along major highways and arterial roadways, suggesting that larger, more important roadways have an out-sized role to play in reducing total system travel cost with fleets. Fleets are most clearly present on the major highways leading into and out of downtown Pittsburgh, near the center of Figure 5 and at the confluence of the three rivers. In contrast, user flow appears clustered in several relatively isolated geographic regions

Figure 4. (Color online) Fleet and User Link Flows at the CFS-SO Mixed Equilibrium on the Sioux Falls Network



Notes. Color and width indicate the amount of flow of each class on each link. Arrows show the direction of link flow, and annotations denote the percentage of flow belong to each class.

Figure 5. (Color online) Fleet and User Link Flows at the CFS-SO Mixed Equilibrium on the Pittsburgh Network



Note. Color and width indicate the amount of flow of each class on each link.

throughout the network, but notably along the periphery and within Pittsburgh's densely populated East End, just east of Downtown.

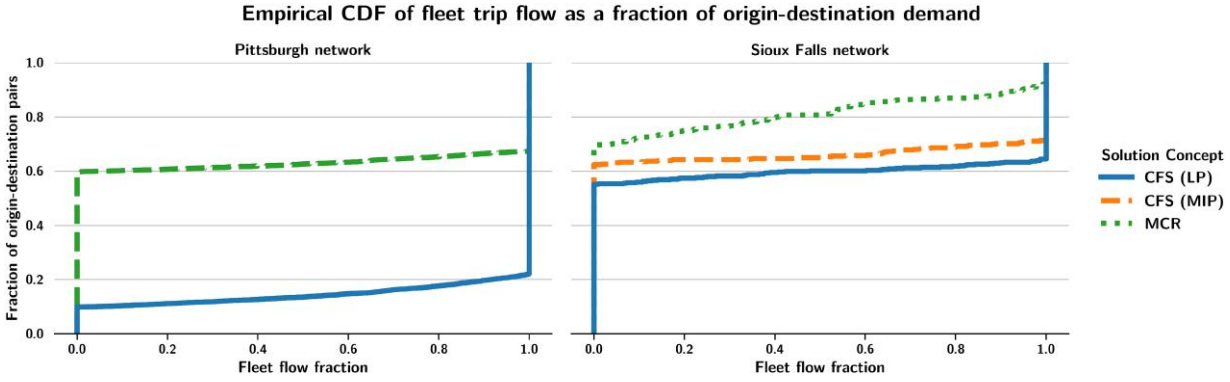
6.2. Fleet OD Demand Patterns at CFS-SO

Although fleets permeate the links of the network, they concentrate on paths and OD pairs on the network. On both the Sioux Falls and Pittsburgh networks, the mixed equilibrium at CFS-SO divides most OD pairs between being exclusively fleet and exclusively user. In analyzing the parallel network in Section 5.1, we noted that it is intuitive to imagine how CFS-SO would result in user or fleet exclusive OD pairs and less so to imagine an OD pairs with a mix of individual users and fleets. Indeed, in both the Sioux Falls and Pittsburgh networks, about 95% and 85% of OD pairs, respectively, are either exclusively user or exclusively fleet, as shown in Figure 6.

The requirement that fleet marginal cost be equalized across the paths used by the fleet between each OD pair and, specifically, that this equalizing requires fleet flow is a key driver of the gap between MCR and CFS-SO. In the MCR setting, the fleet cost is simply the system marginal cost, which, along with link cost, is a constant. As a result, fleet flow can be assigned to those least marginal cost paths independently. By contrast, in the CFS setting, changes to fleet flow on one path must be balanced by fleet flow to equalize fleet marginal cost across used paths on each OD pair.

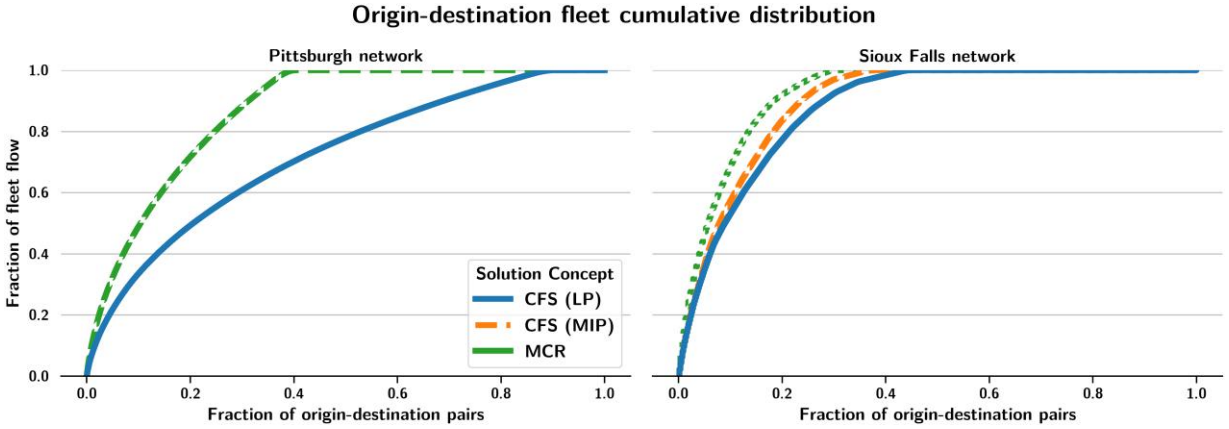
If OD pairs tend to be exclusive to either individual users or fleets, then transportation planners or fleet coordinators may focus their attention on a subset of OD pairs accounting for an outsized proportion of the fleet. Figure 7 shows that for both the Sioux Falls and Pittsburgh networks, a minority of the OD pairs account for a majority of the fleet. In Sioux Falls, just 10% of the OD pairs house half of the fleet; for Pittsburgh, it takes 20% of OD pairs. Remarkably, another 15% of OD pairs in Pittsburgh do not require any fleet presence whatsoever.

Figure 6. (Color online) CFS-SO and MCR Both Result in User- and Fleet-Exclusive OD Pairs



Note. The empirical cumulative distribution of the fraction of demand assigned to the fleet is shown for each solution concept and network.

Figure 7. (Color online) Cumulative Fraction of Fleet Flow over OD Pairs

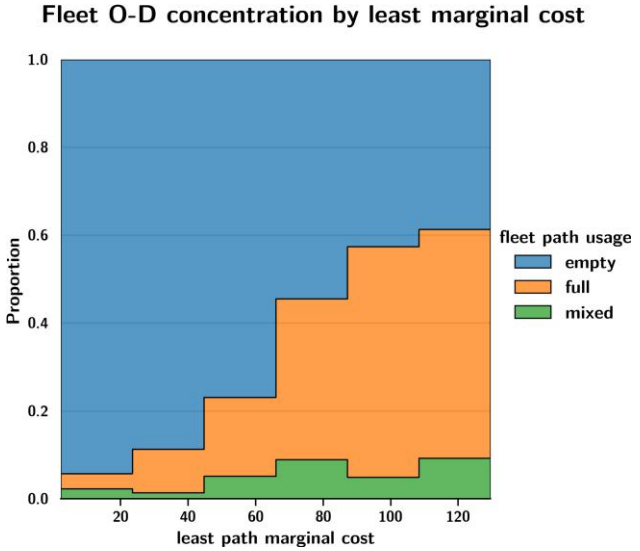


Notes. Fleet flow is concentrated with respect to OD pairs: In Sioux Falls, nearly 70% of the fleet comes from just 20% of OD pairs. In the Pittsburgh network, roughly half of the fleet operates on the only 20% of OD pairs.

6.3. Characteristics of Paths and OD Pairs with Fleet Flow at CFS-SO

In both the Pittsburgh and Sioux Falls networks, OD pairs with larger minimum marginal path cost generally have greater shares of fleet flow as shown in Figure 8. This conforms with the intuition that the fleet focuses on trips where congestion (and thus marginal cost) is greatest to achieve SO. Despite this apparent trend, it is still difficult to distinguish between fleet-exclusive OD pairs from user-exclusive ones a priori or characterize what drives the fleet penetration on OD pairs. For example, despite the strong trend in Figure 8, marginal cost has little to no predictive power: Among OD pairs with the largest marginal cost, only 60% have fleet flow on them. Instead, we derive a metric, the path independence factor, from the optimality conditions of the CFS-SO LP that is capable of classifying fleet-exclusive paths with greater precision than simple criteria.

Figure 8. (Color online) Proportion of Fleet-Exclusive (Full), User-Exclusive (Empty), and Mixed OD Pairs by the Minimum Marginal Cost over Paths for Each OD Pair at CFS-SO (MIP) in the Sioux Falls Network

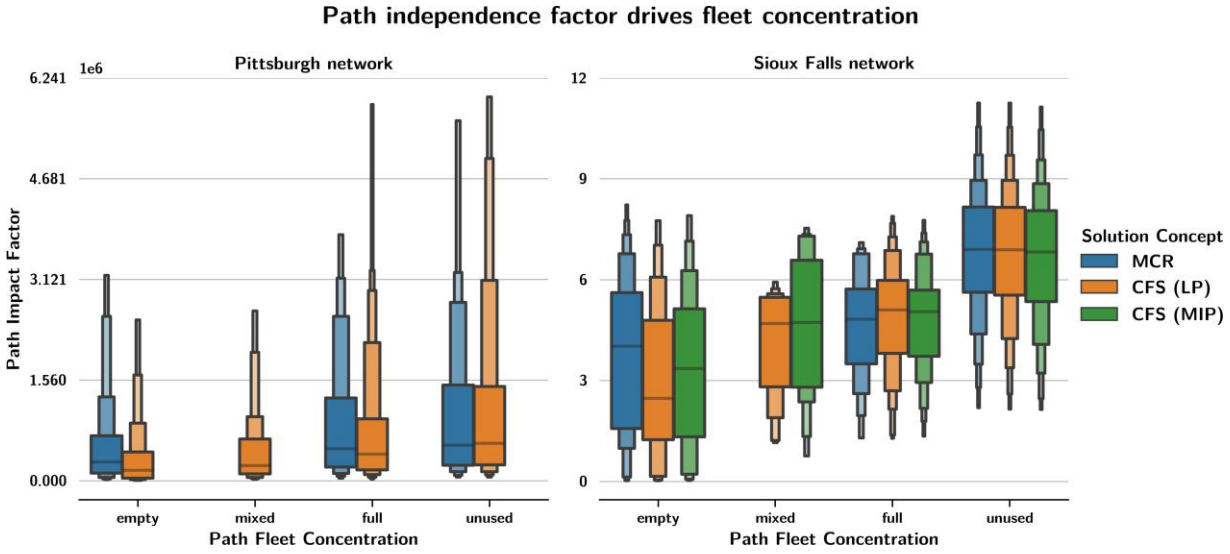


The path independence factor measures how sensitive fleet flow on a path could be to flow on other paths: how independent the cost of a path is from the flow on other paths. Fleet flow on a path that is highly intertwined with many other paths or has a large travel cost derivative has a greater capacity to increase fleet marginal cost than a path that is not. Following this intuition, we can compute path independence as the sum of the link cost derivatives on path r weighted by number of fleet usable paths that do not include the link:

$$\begin{aligned}\text{PathIndependenceFactor}(r) &= \sum_{p\in\tilde{P}\setminus\{r\}}\sum_{a\in A}\delta_{ap}\mathbf{t}'_a(1-\delta_{ar})\\ &= \sum_{a\in\text{links}(r)}\mathbf{t}'_a(|\tilde{P}|-N_a(\tilde{P})),\end{aligned} \tag{20}$$

where $N_a(\tilde{P})$ is the number of paths in $\tilde{P}$ containing link a and $\mathbf{t}'_a$ refers to the derivative of the link cost function of link a evaluated at the aggregate link flow. This term is derived from the part of the fleet marginal cost at SO flow that depends on fleet flow. A related form appears in the first-order optimality conditions for Program (17) (see Online Appendix I). The path independence factor is nonnegative. It is zero for paths where every link is on every path in $\tilde{P}$. Paths with larger path independence have fewer high-cost interactions with paths in $\tilde{P}$. Empirically, paths with larger values of path independence tend to have more fleet flow on them, as shown in Figure 9. Paths

Figure 9. (Color online) Paths with a Higher Independence Factor (Defined by (20)) Have More Fleet, Up to a Point



Notes. The paths with the highest independence factor will tend to be unused completely. Here, "empty" means at most 5% flow on the path is composed of fleet vehicles, "full" means at least 95% flow on the path is fleet flow, and "mixed" is the remaining paths.

that are completely unused tend to have the highest path independence factor; this is largely due to paths that have few interactions with fleet usable paths because they themselves are not fleet usable.

It is important to note that one may compute the path independence without having solved for CFS-SO; one only needs the set of least marginal cost paths at SO. As a first pass of testing the possibility that path independence might be used as a proxy for CFS-SO, we trained an explainable boosting machine (EBM) (Lou et al. 2013, Nori et al. 2019) on the Sioux Falls path flows to predict whether the flow on a path that is usable by the fleet (i.e., it is least marginal cost) will be exclusively fleet or not at CFS-SO. Using just the path independence factor and whether the path may also be used by individual users (i.e., it is also least cost), the EBM was able to achieve an area under the receiver operating characteristic curve (ROC AUC) of 75%. In other words, the model will rank a random fleet exclusive path higher than a random nonfleet exclusive path 75% of the time. As Figure 9 demonstrates, path independence is on wildly different scales on different networks so we cannot not expect any such model to generalize directly. Rather the model serves as a measure of the extent to which path independence is able to characterize this particular CFS-SO solution. The question of whether path independence can be generalized as a heuristic is left for future work.

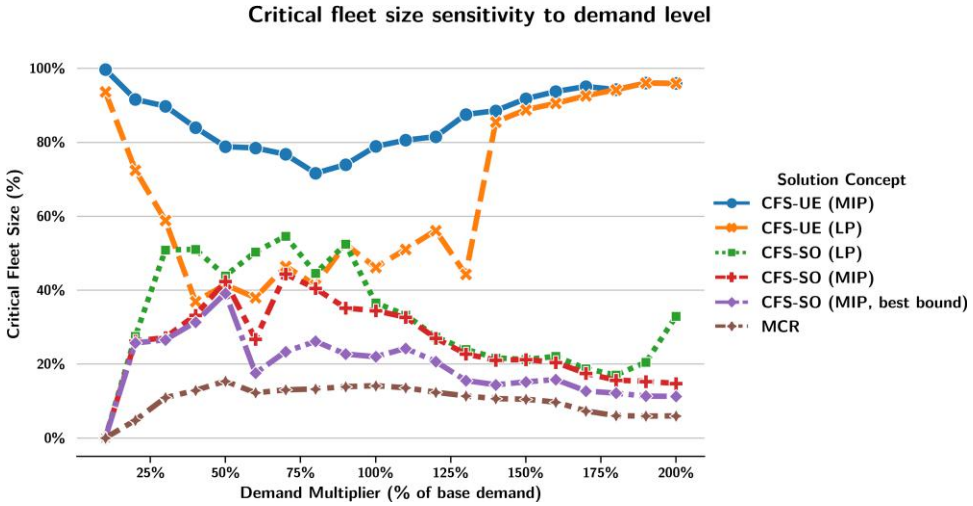
6.4. Sensitivity of CFS-SO and CFS-UE to Aggregate Demand

For the Sioux Falls network, not only does the linear program offer a comparable value of CFS-SO, the solution itself tracks closely with the solution returned by the CFS-SO MIP. This is most evident in Figures 6 and 7, where the curves generated by the MIP tracks closely enough with the curves generated by the LP so that none of our conclusions would be altered by taking one solution instead of the other. To examine the quality of the linear program upper bound in a more general sense, we solve MCR, CFS-SO (exact and upper bound), and CFS-UE (exact and lower bound) on varied demand levels on the Sioux Falls network. Each MIP was terminated at an absolute gap of 1% of the total travel demand or at a time limit of two hours, whichever came first. We compute an approximate lower bound for CFS-SO, based on the best objective lower bound provided by the solver and on the assumption that the critical fleet size is an equal percentage of the objective at the incumbent solution and at the best lower bound. Explicitly, the bound is computed using Equation (21), where CFS is the critical fleet size at the best integer solution returned by the solver. The results of the demand sensitivity analysis are summarized in Figure 10.

$$\text{CFS Lower Bound} = \text{CFS} \cdot \frac{\text{Objective Best Bound}}{\text{Objective Best Integer Solution}} \tag{21}$$

CFS-SO and MCR show similar trends with respect to the demand multiplier, although CFS is more sensitive to changes in demand level. Both CFS-SO and MCR are small for very low demand levels, increasing with demand before decreasing at higher demand levels. In the very-low demand regime, we expect CFS to be small or zero because we expect the network to be almost completely uncongested so that UE and SO are close.

Figure 10. (Color online) CFS-UE, CFS-UE Lower Bound, CFS-SO Upper Bound, CFS-SO MIP Solution with Approximate Lower Bound, and MCR as a Function of the Demand Multiplier That Scales Demand Volume on Each OD Pair



Encouragingly, past a certain point, both MCR and CFS-SO decrease, meaning that at high demands a smaller fraction of the total demand is needed for the system to benefit. Interestingly a similar pattern emerges for CFS-UE: past a certain demand level (roughly 80% of base demand) CFS-UE increases in demand level, implying that the network can handle high levels of optimized fleet demand without impacting UE traffic patterns at all.

In the Sioux Falls network, the linear program provides an upper bound within 6% of the CFS-SO found by the mixed integer program for 14 of the 20 demand levels. For large demand levels, between 1.0× and 1.9× the base demand, the linear program provides a reasonable bound at a small fraction of the computational cost. It is in the low demand regime (0.2×–0.9×) where the upper bound performs particularly poorly, overestimating the CFS-SO found by the mixed integer program by up to 24%. However, the low demand regime is precisely where the linear program's computational advantages are least useful. For each demand level less than or equal to 50%, the MIP was solved to a relative gap of at most 8% within two hours. In these cases, the upper bound is loose because the MIP was able to find high quality solutions. In contrast, no demand level above 50% was solved to a relative gap smaller than 22% within the two-hour time limit. The LP lower bound for CFS-UE, however, is reasonably tight only for very low and very high demand regimes and at worst underestimates CFS-UE by around 45%. The solutions returned by the CFS-UE MIP within the two-hour time limit, however, are of comparatively better quality than the solutions returned by the CFS-SO MIP within the same time limit. Over the demand levels we test, the CFS-UE solutions have an average relative gap of 3.6% compared with 24.5% for the CFS-SO solutions. Overall, this indicates that the CFS-UE MIP is somehow less challenging for the solver than the CFS-SO MIP. One possible reason for this is the substantially smaller path set used in the CFS-UE MIP compared with the CFS-SO MIP.

In general, is it difficult to justify the tightness of the LP bounds for either CFS-UE or CFS-SO. From Figure 10, it is clear that on the Sioux Falls network, the LP generally provides a tighter bound for CFS-SO than it does CFS-UE. It is not immediately clear why. Part of the reason may have to do with the fact that, as we have observed, the CFS-UE MIP is apparently easier to solve than the CFS-SO MIP, resulting in CFS-UE MIP solutions that are further from the LP bound than the CFS-SO MIP solutions. Another difficulty is that the LP equalizes the fleet marginal cost over a potentially much larger set of paths than the set the true CFS-SO fleet would use. It is not clear how to estimate a priori how much this affects the amount of fleet flow required. An option for practitioners who need both a bound and a measure of tightness is to run the MIP for a fixed computational budget.

7. Discussion

In both Pittsburgh and Sioux Falls networks, a nontrivial mixture of self-interested behavior induces SO traffic flow. Should CFS-SO be achieved, travelers would be on average are better off with the presence of the fleet than they would have been if they had all routed themselves independently. In contrast, the MCR, which also achieves SO traffic flow, may not make fleets better off. This important distinction highlights the advantage of explicitly modeling fleet behavior in traffic analysis and demonstrates the necessity of doing so. In this sense, of the many ways to partition SO as a mixed equilibrium, CFS-SO offers a particularly compelling case for its realism: no external subsidy is required.

Are fleets better off? Collectively, at both CFS-SO and MCR, individual users have lower average travel cost than the fleet simply because individual users use only least cost paths. One interesting effect of CFS-SO relative to MCR is that, although both achieve aggregate SO, they do so by partitioning demand between the fleet and individuals in very different ways as demonstrated by Table 5. First, the average travel cost, computed as the flow-weighted average path travel cost, is much higher for the fleet than for the user both at MCR and CFS-SO. This is partly because individual users are only using least cost paths, but it is also because, demonstrated in Figure 8, that fleets accumulate on OD pairs with large marginal cost. When the link cost function is monotonic increasing and convex, we should generally expect a positive correlation between cost and marginal cost ($t(x) > t(y) \Rightarrow t'(x) > t'(y)$). As a result, because fleet volume skews toward high marginal cost OD pairs, and will not necessarily use least cost paths, the average fleet cost is much higher than the average user cost.

Second, the average travel cost of both the fleet *and* the individual users decreases from MCR to CFS-SO, although the aggregate average cost is the same. This counter-intuitive result is possible only because the fleet and users have completely different OD demand between MCR and CFS-SO. In shifting from MCR to CFS-SO, the fleet needs to acquire flow on lower cost paths; it achieves this by taking paths from users that are lower cost than the average fleet path cost, but higher cost than the average user path cost, resulting in a reduction in both averages.

Third, the travel cost discount that the fleet receives relative to UE is greater at CFS-SO than at MCR. Moreover, at MCR on the Sioux Falls network, the fleet travel time *increases* relative to UE. We can compare the fleet and individual average travel cost under CFS-SO and MCR with the average travel cost of a set of vehicles with the same OD demand under aggregate UE using a measure we term the "coordination discount":

$$\text{CoordinationDiscount}(\mathbf{f}) = \frac{\langle \mathbf{c}, \mathbf{f} \rangle}{\langle \lambda_{\mathbf{ue}}, \mathbf{Mf} \rangle} - 1, \tag{22}$$

where $\mathbf{f}$ is the path flow of interest (i.e., the fleet, user, or aggregate path flow) under the solution concept of interest (i.e., MCR, CFS-SO, or aggregate SO), $\mathbf{c}$ is the path cost at $\mathbf{f}$, and $\lambda_{\mathbf{ue}}$ is the OD cost under aggregate UE. When $\mathbf{f}$ is an aggregate SO path flow, the coordination discount is termed the SO discount and is related to the price of anarchy as $1/\text{PoA} - 1$. Positive values of the coordination discount indicate that the flow in question is *worse off* under its current routing than they would have been at aggregate UE; negative values indicate that the flow in question is better off using its current routing. In Table 5, we can see that at MCR in Sioux Falls the fleet total cost *increases* 3.44% relative to UE even though the *system* total cost *decreases* by 3.82%. This is an example of the system optimality paradox discussed in Section 4.3. In contrast, a fleet at CFS-SO, *reduces* its total cost by 4.32% relative to UE. In Pittsburgh, the fleet is only slightly better off under MCR than under UE but could increase their savings to nearly match the system SO discount by optimizing their fleet.

7.1. CFS-SO as Optimization-by-Proxy

Viewed a different way, our results demonstrate that SO traffic flow is attainable by optimizing over a subset of vehicles rather than over all vehicles. In optimizing the entire network, the traffic manager should know, at the very least, the total system travel time, a measurement over all vehicular flow in the network. In practice, this measurement is simply not feasible to take. What a nontrivial critical fleet size indicates is that the measurement need only be taken and the optimization need only be performed on a subset of the vehicles. If this subset of vehicles is easy to measure (e.g., they are all using the same navigation software or information platform), then all that is additionally necessary are travel time estimates on the links of the network, which are substantially easier to measure than untracked vehicles.

However, how large could we reasonably expect such a fleet to be? According to a 2015 study, more than 90% of smartphone users use their smartphones for directions (Anderson 2016), and roughly 73% of Americans in 2020 owned smartphones (O'Dea 2020). Finally, 84% of those who use smartphones for directions use a Google-owned navigation app (Ceci 2021). By a simple back-of-the-envelope calculation, nearly 66% of drivers use some form of smartphone navigation app, and 55% of drivers use a Google-owned navigation app. By contrast, it is estimated that ride-sourcing vehicles make up

Table 5. Coordination Discount and Average Cost by Flow Class, Solution Concept, and Network

			Coordination discount (%)		Average cost	
Network	Solution concept	SO discount (%)	Fleet	User	Fleet	User
Sioux Falls	CFS-SO (LP)	−3.82	−4.32	−3.30	27.64	15.52
	MCR		**+3.44**	−5.81	32.64	17.86
Pittsburgh	CFS-SO (LP)	−1.19	−1.19	−1.28	110,188.88	47,590.55
	MCR		−0.21	−2.34	135,722.77	83,179.67

Note. Bold indicates the outcome highlights.

at most 14% of vehicle miles traveled among the largest urban areas in the United States (Balding et al. 2019). Although still less than the CFS-SO on the Pittsburgh network, an optimized fleet of Google Maps users is large enough to achieve SO on many networks, Sioux Falls included, if distributed in the right way.

It's worth noting that CFS-SO finds the smallest total fleet size capable of achieving SO with no constraints on its OD demand. Specifically in drawing an analogy to Google Maps, it may be of interest to find CFS-SO that balances the total fleet size with how reasonable the resulting OD fleet demand pattern appears, for example, how uniformly it is distributed across OD pairs. We leave this as an area for future work.

In general, however, the scale of the empirical CFS-SO results would seem to preclude the most obvious source of an optimized fleet; in the case of Sioux Falls for example, it is unlikely that a UPS or FedEx fleet would account for more than a third of traffic volume, but connectivity enables the coordination of vehicles on a much larger scale. FedEx can optimize because it either owns vehicles and uses CDL (commercial driver's license) drivers or contracts with freight operators. Both avenues are capital intensive and accordingly limit the size of the fleet. In contrast, Uber and Lyft have far lower driver acquisition costs because they do not own the vehicle nor employ the driver (who needs only a standard driver's license) directly. In effect, CFS-SO, and in many networks MCR as well, could *only* be achieved via the scale of vehicle coordination enabled by app-based services.

Finally, CFS-SO and CFS-UE capture two very different realities for fleet operators. When fleets on the network induce the same traffic flow as UE would, we are right to ask whether the fleet should bother optimizing itself at all. If mixed equilibrium is the same as UE, then the fleet has not gained anything by optimizing. In fact, if there is a cost to coordinating behavior, then the service is strictly worse off in coordinating the fleet. In contrast, when mixed equilibrium is SO, the potential benefits of optimization are clear. In fact, in both the Sioux Falls and Pittsburgh networks, the fleet vehicles are strictly better off optimizing themselves while at the same time driving the system to its optimal traffic flow. In contrast, they are worse off under the MCR.

7.2. Applications of CFS-SO and CFS-UE

There are four principal applications we see for critical fleet size analysis; it offers

1. Regulators a new way to inform congestion management policy,

2. Transportation planners an important metric to evaluate road network improvements,

3. Service providers, such as ride-sourcing companies, navigation services, and car manufacturers increasingly interested in mobility more broadly, a better understanding of the network effects of their fleet coordination, specifically in identifying areas of the network where their efforts could align with system level goals, and

4. The public a tool to understand how individual users might benefit from increased fleet presence on their roadways.

In the search for regulatory strategies for ad hoc fleets, including ride-sourcing services, CFS-SO may provide insights into corridors where fleets would be particularly helpful in reducing system-level congestion. Conversely, CFS-UE may provide useful insights into corridors where ad hoc fleets would not be expected to meaningfully change travel time or cost. Critical fleet size analysis can therefore inform more targeted congestion management schemes by recognizing where the interests of traffic managers and fleet coordinators align, where they do not align, and where fleets will not be expected to alter traffic patterns.

CFS-SO is also an important metric in evaluating a set of alternative road network improvements. A road network improvement that reduces CFS-SO could be leveraged by a fleet to improve system cost, potentially to SO levels. All else equal, a road improvement that reduces CFS-SO more will pay dividends in the future as it makes it easier for a coordinated fleet to reach the threshold required to bring the system to optimality, without any external subsidy or intervention on the part of the transportation planner.

Services that can coordinate their fleets should ask themselves whether and how to optimize their fleet as it is often in their interest to guide the fleet toward a service-level objective. CFS-SO offers service providers a window into how well their service may align with network level objectives. In the case of green routes in Google Maps, this is an existential question. By complement, CFS-UE offers service providers insight into where optimizing their fleet may not be worth it at all.

8. Conclusions and Future Work

As technological advances in transportation continue to permeate our road networks, it becomes easier for services to coordinate vehicle behavior toward a service-level objective. In this work, we study if and how self-interested services that optimize their fleet can reduce system cost on a road network. We demonstrate that self-interested fleets are capable of both paradoxically increasing system cost over UE and reducing system cost up to achieving SO network flow. We provide a mathematical program with equilibrium constraints to solve for the smallest fleet that would induce SO and the largest fleet that would induce UE in mixed traffic with UE users. We present efficient solution methodologies and apply them to the Sioux Falls and Pittsburgh networks finding that 33% and at most 83% of vehicles, respectively, must join the fleet for the network to reach

SO. Moreover, these vehicles are better off than they would have been had they either routed themselves independently or participated in a MCR-like scheme, meaning that neither subsidies nor tolls would be required to compensate the fleet. At CFS-SO on the Pittsburgh network, fleet vehicles are most present along highways and major arterials, whereas individual users tend to accumulate on shorter trips outside of the central business district. The "path impact" metric, derived from the KKT conditions of the CFS linear program, is found to drive the concentration of fleet vehicles on paths. In other words, we have found two examples of networks on which a nontrivial mixture of self-interested behavior can induce minimum system cost on the network, without the need for external subsidy. Critical fleet size offers regulators a new way to inform congestion management policy, transportation planners a new metric to evaluate network improvements, and fleet operators/coordinators a better understanding of benefits of their fleet optimization efforts at the level of both their own platform and the entire network.

A key avenue of future work concerns the conditions under which a fleet will reduce system cost on a network. Although the condition given by Equation (3) yields some theoretical insight, it is not of practical use. A better practical condition would depend only on the network itself and, potentially, the UE or SO link flow. Whether such a condition exists and if it is easier to compute than mixed equilibrium is an area for future work. Additionally, the example of the fleet optimality paradox is contrived. Whether the paradox can be demonstrated in real world networks remains to be seen. If the paradox can be found in real world networks, these examples may inform better conditions or heuristics to characterize networks that may suffer from the presence of fleets.

The use of critical fleet size in transportation planning exposes an interesting avenue of future work in characterizing the kinds of improvements that tend to reduce critical fleet size in real-world networks. Such work would expand our understanding of what drives critical fleet size and, more broadly, what drives system cost reduction when optimized fleets are present on the network. It may be of interest for public agencies to regulate a fleet coordinator toward a deployment pattern inline with CFS-SO. One would expect the fleet-optimal operation, in conjunction with incentives, may lead to SO flow, or considerably mitigate congestion.

As formulated, CFS enforces no constraints on the OD fleet demand. This is intentional, as in this work, we are focused on finding strictly the smallest optimized fleet capable of inducing SO. There are larger fleets that would also achieve SO, but as our analysis shows, adding fleet to the CFS-SO demand pattern may break SO. In practice then, a CFS OD demand pattern that adheres to some prior notion of fleet demand might be more useful. Adding a penalty term to the CFS mathematical programs would be one way to achieve this and we leave such studies for future work. For example, it may be more reasonable to find a critical fleet size that balances total fleet size with close to uniform OD demand. One could achieve this by constructing a measure of nonuniformity as the penalty term in the objective, that is by replacing the objective in Program 17 with, for example,

$$\min_{\mathbf{f}^{\mathbf{u}}\geq 0,\,\mathbf{f}^{\mathbf{c}}\geq 0,\,\lambda} \sum_{p\in\bar{P}} \mathbf{f}^{\mathbf{c}}_p + \gamma \left\| \mathbf{q}^{\mathbf{c}} - \frac{1}{|W|}\sum_{w\in W} \mathbf{q}^{\mathbf{c}}_w \right\|. \tag{23}$$

Other regularization is of course possible. A penalty term of the form $\|\mathbf{q}^{\mathbf{c}} - \hat{\mathbf{q}}\|$, for example, would encourage the demand pattern at CFS-SO to match some prior demand pattern.

In this paper, CFS is solved with the presence of only a single fleet. In reality, there could exist many such fleets. CFS-SO with multiple fleets finds the smallest total number of fleet vehicles such that the mixed equilibrium is SO. In Online Appendix G, our mathematical programs are expanded with additional constraints to ensure that *each* fleet meets FO criteria and that aggregate demand and link flows are conserved. Experiments on real networks and whether CFS increases or decreases when multiple fleets are present is left for future work. We hypothesize, however, that because multiple fleets will compete with each other, CFS-SO will be minimized when only one fleet is on the network. Refuting or proving this hypothesis is also left for future work.

In commenting on congestion, critical fleet size considers route-choice behavior alone. Our framework elides many congestion-influencing behaviors that have been attributed to ad hoc fleets in the literature. A complete treatment of the congestion effects of such fleets should not only account for the route choices but also the behaviors associated with that particular service. Of course, not every service provider wishes to optimize total fleet travel cost, which this particular formulation of critical fleet size assumes. An important area of future research is then to adapt the methods in this work to other fleet-level objectives a service may have, such as total emissions in the case of Google Maps' green routes.

The relationship between the two notions of critical fleet size on networks with a nonunitary price of anarchy is not clear. Characterizing the relationship either empirically through further experiments on real-world networks or theoretically would expand our understanding of both critical fleet size as a quantity of interest and how networks in general should be expected to respond to the presence of optimized fleets.

We study CFS in the context of static equilibrium. This was a deliberate choice as we leveraged the simplicity of the static setting to isolate fleet behavior as the sole reason for the increase in total system cost in the FO

paradox and as the sole reason that a mixture of fleets and individuals could achieve SO. CFS is also an interesting quantity in the dynamic setting and is an area of future work we are eager to explore.

Finally, solving critical fleet size, or its bound, in large networks is computationally challenging. Not only does the problem require a high-quality SO link flow solution, but the number of variables and constraints grows quickly with network size. There is no immediately obvious way of reducing the problem complexity, but a more clever row and column generating scheme and tighter integration with solver software may work well in practice to solve the smallest possible critical fleet size program. Finding ways to improve the computational efficiency of critical fleet size algorithms will make it a more easily applicable algorithm.

Acknowledgments

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein.

References

Agrawal A, Verschueren R, Diamond S, Boyd S (2018) A rewriting system for convex optimization problems. *J. Control Decisions* 5(1):42–60.

Alcántara AM (2021) Google Maps to add a greenest route to its driving directions. *Wall Street Journal*, Accessed April 23, 2021, https://www.wsj.com/articles/google-maps-to-add-a-greenest-route-to-its-driving-directions-11619197255.

Anderson M (2016) More Americans using smartphones for getting directions, streaming TV. https://www.pewresearch.org/fact-tank/2016/01/29/us-smartphone-use/.

Balding M, Whinery T, Leshner E, Womeldorff E (2019) Estimated percent of total driving by Lyft and Uber. Technical report, Fehr & Peers, Washington DC.

Beckmann M, McGuire CB, Winsten CB (1956) *Studies in the Economics of Transportation* (Yale University Press, New Haven, CT).

Boyles SD, Lownes NE, Unnikrishnan A (2021) *Transportation Network Analysis*, vol 1. https://sboyles.github.io/blubook.html.

Ceci L (2021) Most popular mapping apps in the United States as of April 2018, by reach. Technical report, Statista.

Chen Z, Lin X, Yin Y, Li M (2020) Path controlling of automated vehicles for system optimum on transportation networks with heterogeneous traffic stream. *Transporation Res., Part C Emerging Tech.* 110:312–329.

Cominetti R, Correa JR, Stier-Moses NE (2009) The impact of oligopolistic competition in networks. *Oper. Res.* 57(6):1421–1437.

Dafermos SC (1972) The traffic assignment problem for multiclass-user transportation networks. *Transportation Sci.* 6(1):73–87.

Delle Site P (2021) Pricing of connected and autonomous vehicles in mixed-traffic networks. *Transporation Res. Rec.* 2675(5):178–192.

Diamond S, Boyd S (2016) CVXPY: A Python-embedded modeling language for convex optimization. *J. Machine Learn. Res.* 17(83):1–5.

Erhardt GD, Roy S, Cooper D, Sana B, Chen M, Castiglione J (2019) Do transportation network companies decrease or increase congestion? *Sci. Advice* 5(5):eaau2670.

Fagnant DJ, Kockelman KM (2014) The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transporation Res., Part C Emerging Tech.* 40:1–13.

Hall JD, Palsson C, Price J (2018) Is Uber a substitute or complement for public transit? *J. Urban Econom.* 108:36–50.

Harker PT (1988) Multiple equilibrium behaviors on networks. *Transportation Sci.* 22(1):39–46.

Hestenes MR (1969) Multiplier and gradient methods. *J. Optim. Theory Appl.* 4(5):303–320.

Lau J (2020) Google Maps 101: How AI helps predict traffic and determine routes. https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/.

Lou Y, Caruana R, Gehrke J, Hooker G (2013) Accurate intelligible models with pairwise interactions. *Proc. 19th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining* (ACM, New York), 623–631.

Mansourianfar MH, Gu Z, Waller ST, Saberi M (2021) Joint routing and pricing control in congested mixed autonomy networks. *Transporation Res., Part C Emerging Tech.* 131:103338.

Mehr N, Horowitz R (2019) How will the presence of autonomous vehicles affect the equilibrium state of traffic networks? *IEEE Trans. Control Network Systems* 7(1):96–105.

Nori H, Jenkins S, Koch P, Caruana R (2019) InterpretML: A unified framework for machine learning interpretability. Preprint, submitted September 19, https://arxiv.org/abs/1909.09223.

O'Dea S (2020) Smartphone penetration rate as share of the population in the United States from 2010 to 2021. Technical report, Statista.

Roughgarden T (2005) *Selfish Routing and the Price of Anarchy* (MIT Press, Cambridge, MA).

Sharon G, Albert M, Rambha T, Boyles S, Stone P (2018) Traffic optimization for a mixture of self-interested and compliant agents. *Proc. 32nd AAAI Conf. on Artificial Intelligence*.

Sheffi Y (1985) *Urban Transportation Networks*, vol. 6. (Prentice-Hall, Englewood Cliffs, NJ).

Stabler B, Bar-Gera H, Sall E (2021) Transportation networks for research. Accessed November 10, 2021, https://github.com/bstabler/TransportationNetworks.

Tobin RL, Friesz TL (1988) Sensitivity analysis for equilibrium network flow. *Transportation Sci.* 22(4):242–250.

Uber Technologies (2022) How does Uber match riders with drivers? https://www.uber.com/us/en/marketplace/matching/.

Wang J, Zheng Y, Xu Q, Wang J, Li K (2020) Controllability analysis and optimal control of mixed traffic flow with human-driven and autonomous vehicles. *IEEE Trans. Intelligent Transportation Systems.* 22(12):7445–7459.

Ward JW, Michalek JJ, Azevedo IL, Samaras C, Ferreira P (2019) Effects of on-demand ridesourcing on vehicle ownership, fuel consumption, vehicle miles traveled, and emissions per capita in US states. *Transporation Res., Part C Emerging Tech.* 108:289–301.

Wardrop JG (1952) Some theoretical aspects of road traffic research. *Proc. Inst. Civil Engrg.* 1(3):325–362.

Wright S, Nocedal J, Wright SJ (1999) Numerical optimization. *Springer Sci.* 35(67–68):7.

Yang H, Zhang X, Meng Q (2007) Stackelberg games and multiple equilibrium behaviors on networks. *Transportation Res. Part B: Methodological* 41(8):841–861.

Zhang K, Nie YM (2018) Mitigating the impact of selfish routing: An optimal-ratio control scheme (ORCS) inspired by autonomous driving. *Transporation Res., Part C Emerging Tech.* 87:75–90.