Empty-car Routing in Ridesharing Systems

Anton Braverman

School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853, ab2329@cornell.edu

J.G. Dai

School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853, jim.dai@cornell.edu

Xin Liu

School of Electrical, Computer and Energy Engineering Arizona State University Tempe, Arizona 85287, xliu272@asu.edu

Lei Ying

School of Electrical, Computer and Energy Engineering Arizona State University Tempe, Arizona 85287, lei.ying.2@asu.edu

This paper considers a closed queueing network model of ridesharing systems such as Didi Chuxing, Lyft, and Uber. We focus on empty-car routing, a mechanism by which we control car flow in the network to optimize system-wide utility functions, e.g. the availability of empty cars when a passenger arrives. We establish both process-level and steady-state convergence of the queueing network to a fluid limit in a large market regime where demand for rides and supply of cars tend to infinity, and use this limit to study a fluid-based optimization problem. We prove that the optimal network utility obtained from the fluid-based optimization is an upper bound on the utility in the finite car system for any routing policy, both static and dynamic, under which the closed queueing network has a stationary distribution. This upper bound is achieved asymptotically under the fluid-based optimal routing policy. Simulation results with real-world data released by Didi Chuxing demonstrate the benefit of using the fluid-based optimal routing policy compared to various other policies.

Key words: ridesharing, fluid limit, closed queueing network, BCMP network, car routing

1. Introduction

This paper studies the modelling and control of ridesharing systems such as Didi Chuxing, Lyft and Uber. We consider a system with $r > 0$ regions and $N > 0$ cars. The regions can be interpreted as geographic regions in a city and cars drive around between regions transporting passengers. At time $t = 0$, all cars start off idling empty in some region, waiting for a passenger. Passengers arrive to region i according to a Poisson process with rate $N\lambda_i > 0$, and arrivals to different regions are

independent. When a passenger arrives to region i, if there is an empty car available there, then the passenger occupies that car and travels to region j with probability P_{ij}. If no empty car is available, the passenger abandons the system and finds an alternative form of transportation to her destination. We allow $P_{ii} > 0$ to represent trips within a region. Travel times from region i to j have mean $1/\mu_{ij}$ and are assumed to be i.i.d. exponential random variables, although this assumption is not essential (see Remark 3 in Section 2.1). Once the passenger arrives at region j, the car becomes empty. The empty car can either stay in region j with probability Q_{jj} (it becomes available to take new passengers immediately), or with probability Q_{jk}, relocate without a passenger to a different region k and wait for a passenger there. The time spent driving empty from j to k is identical in distribution to that of driving with a customer. In general, the routing matrix (also called the routing policy) $Q = (Q_{ij})$ is allowed to be *state-dependent*, i.e. Q may depend on the current distribution of cars across the regions. In this paper, Q will be a decision variable.

We model this ridesharing network with a closed queueing network consisting of both single-server and infinite-server stations, where cars are "jobs" moving through the queueing network. Cars waiting in a region for a passenger are modeled with a single-server station. The buffer content of the station corresponds to the number of cars waiting, and passenger ride requests correspond to service completions at the station. Thus, the service times at the single-server station are the interarrival times of passengers to the region, although there is no physical "server" at the station. The infinite-server stations are used to model car travel between regions. When the routing policy Q is static, i.e. not state dependent, our queueing network belongs to a class of networks called BCMP networks Baskett et al. (1975). The precise formulation of the model can be found in Section 2.

Because of the proliferation of ridesharing and bikesharing services, modelling and control of these systems have become important research topics over the last few years Adelman (2007), George and Xia (2011), Waserhole and Jost (2013, 2016), Banerjee et al. (2016), Iglesias et al. (2016), Pavone et al. (2012), Zhang and Pavone (2016), Ozkan and Ward (2016), Yang et al. (2016),

Bimpikis et al. (2016). Our paper focuses on empty-car routing as a mechanism to improve the efficiency of the system. To illustrate the effect of this mechanism, consider the two-region example in Figure 1. Passengers arrive at region 1 to go to region 2 according to a Poisson process with rate 800 passengers/unit time, and arrive at region 2 to go to region 1 with rate 400 passenger/unit time. After dropping off a passenger at region 2, a driver stays at region 2 with probability Q_{22}, or drives empty to region 1 with probability Q_{21}. The probabilities Q_{11} and Q_{12} are defined analogously. We define the *availability* at region i to be the long-run fraction of time that there is at least one empty car at the region available. Since Poisson arrivals see time averages (PASTA property), this is also the probability that a passenger's request for a ride originating from region i will get fulfilled.
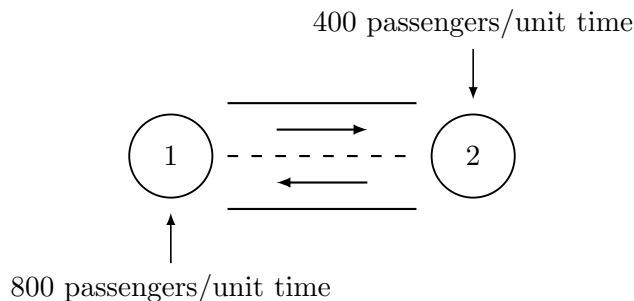

Figure 1 A two-region example

A number of existing models consider *one-way* vehicle sharing systems, in which a vehicle can only be moved from one region to another when carrying a passenger George and Xia (2011), Waserhole and Jost (2016), Banerjee et al. (2016). This is a realistic assumption for bikesharing systems, where a bicycle cannot move autonomously from one region to another, and only moves when a passenger rides it. In such a case, the performance of the system is largely determined by the passengers' arrival rates and destination probabilities. In our example, a one-way system would correspond to $Q_{12} = Q_{21} = 0$, and a bike taken from 1 to 2 will only return to 1 if it is brought there by a passenger. Since, on average, region 1 sees twice as many passengers as region 2, the availability of bikes at region 1 will always be at most 50%, regardless of the number of bikes in the system. That is, region 1 will lose at least half of its passengers to alternative modes

of transportation. This inefficiency due to passenger imbalance has been well recognized in the literature. Proposed solutions include demand throttling via pricing Waserhole and Jost (2016), Banerjee et al. (2016), or periodic bike rebalancing using trucks Chemla et al. (2013), Henderson et al. (2016).

Empty-car routing is an appropriate mechanism for commercial ridesharing systems, where drivers often wander around to find passengers, and it is not surprising that a good routing policy can increase the efficiency of the system. Returning to our example, we assume that there are a total of 1200 cars in the system, and that the mean travel time in either direction is one unit. Table 1 compares the availability in each region under several different routing policies. In particular, we see that the policy with $Q_{21} = 1/3$ is preferable to the policy with no empty-car routing $(Q_{12} = Q_{21} = 0)$. The question we answer in this paper is the following. Given a utility function measuring performance in the system (average availability for example), how should one choose a routing policy to maximize said utility.

Empty-car	Availability	
routing policy	Region 1	Region 2
$Q_{12} = 0, \; Q_{21} = 1/3$	73.19%	97.59%
$Q_{12} = 0, \; Q_{21} = 1/2$	74.64%	74.64%
$Q_{12} = Q_{21} = 0$	50%	100%

Table 1   Availabilities under several empty-car routing policies with 1200 cars in the system computed using the MVA algorithm.

Recall that with a static routing policy Q, our queueing network is a BCMP network. The stationary distribution of a BCMP network has a product form, but the normalization constant is expensive to compute because the state space of the network is too large. However, algorithms such as mean value analysis (MVA) Reiser and Lavenberg (1980) or approximate mean value analysis (AMVA) Suri and Sahu (2007) can bypass computing the normalization constant, and directly

compute performance metrics of interest, e.g. mean queue sizes. In other words, given N, λ, μ, P, and a static Q, steady-state performance analysis of the system can be done efficiently. However, the problem of optimizing some performance metric over Q is difficult. Even in the 2-region case, it can be verified numerically that

$$\max_{Q} \quad \lambda_1 \times (\text{region 1 availability}) + \lambda_2 \times (\text{region 2 availability}) \tag{1}$$

is a non-convex optimization problem. The reason for this is that availabilities have a non-linear dependence on Q. When we tried solving (1), MATLAB's built-in solver failed to converge to a solution. We also tried using NEOS Czyzyk et al. (1998), which is a collection of more sophisticated optimization tools. However, even the solvers in NEOS could not reliably solve the problem.

The main contribution of this paper is provide comprehensive solution to the empty-car routing problem, which is both theoretically grounded and efficient to implement. Namely, we show that as the number of cars and the passenger arrival rates tend to infinity, i.e. $N \to \infty$, the optimal solution of (1) converges to the optimal solution of a fluid-based optimization problem that can be solved by solving a linear program. Our results also hold for much broader class of utility functions; cf. Remark 5 in Section 2.2. We also show in Theorem 2 that the performance under the optimal static routing policy coming from the fluid-based optimization problem is an upper bound on the performance under any *state-dependent* routing policy. Furthermore, this upper bound is asymptotically tight. For any stochastic control problem of realistic size, the true optimal policy is rarely known. Thus, any upper bound, particularly a tight bound on the performance is valuable to develop good policies for finite-sized systems.

The typical asymptotic regime considered for a closed BCMP network is one where the number of jobs in the network goes to infinity, but the service rates at each station remain fixed. In the context of our ridesharing network, this would correspond to a regime where the number of cars N increases to infinity, while both passenger request rates and travel times remain fixed. A lot is known about the asymptotic behavior of BCMP networks in this asymptotic regime, see for instance (George 2012, Section 4.1) and the references within. Most importantly, the limiting network always has at

least one region with an infinite number of empty cars, i.e. a region where availability equals one. For this reason, we refer to this as the *infinite supply* regime. The asymptotic regime considered in this paper has both the number of cars N, and the passenger arrival rates $N\lambda$ going to infinity together. We refer to this regime as the *large market* regime. The infinite supply and large market regimes are qualitatively different. The latter is not guaranteed to have a region where availability converges to 100%; we discuss the implications of this further in Section 1.1.

In practice, the large market regime is more realistic than the infinite supply regime. For starters, it is natural that the supply of drivers in a city increases with demand for rides. Furthermore, the large market regime does not impose any restrictions on supply-demand imbalance in a city. The parameter λ_i is the rate of arriving passengers per car to region i, and gives our model the flexibility to distinguish between cases when there is an oversupply, undersupply, or critical level of supply of cars with respect to passenger demand. The latter two cases are more common during morning or afternoon rush hours, and are precisely the cases where an effective choice of routing matters the most.

Our main results are stated in the steady-state setting, assuming the system parameters stay constant. However, in practice it is very common for parameters to depend on the time of day, e.g. passenger arrivals spike during rush hours. In Section 3.2, we leverage our fluid-based optimization problem to suggest a time-dependent lookahead policy that anticipates future parameter changes and routes cars accordingly. We present several numerical examples with time-varying parameters where this anticipative behavior yields significant performance benefits over the typical approach of dividing time into smaller intervals and assuming constant parameters on each interval. We remark that the purpose of this lookahead heuristic is to demonstrate that the fluid-based optimization can guide the design of high performance empty-car routing policies in practical ridesharing systems where some of the modeling assumptions made in this paper may not hold. However, this lookahead is not meant to the "optimal" algorithm for these settings.

Before moving on to the literature review, we wish to say a few things about the routing mechanism in this model. Our model uses a centralized mechanism for routing, i.e. the ridesharing

company generates routing decisions according to the routing matrix Q, and cars then have to obey that decision. This mechanism is perfectly fine for systems with autonomous vehicles, which are already experimented with by Uber Chafkin (2016). However, a model with centralized control is still useful even when human drivers are free to make their own decisions. For instance, a centralized control mechanism is a best-case performance benchmark against which one can compare decentralized mechanisms. A centralized mechanism can also be used to quantify the "price of anarchy", i.e. the difference in revenue between centralized control and the case when drivers are free to make their own decisions. Part of this cost can then be used to incentivize drivers to obey routing instructions, e.g. by subsidizing fuel costs for driving empty.

1.1. Related Literature

BCMP networks are natural choice for modelling ridesharing systems. The closest papers to us are Zhang and Pavone (2016), Iglesias et al. (2016), where the authors also consider the supply repositioning problem. BCMP networks have also been used to study fleet-sizing George and Xia (2011) and pricing Waserhole and Jost (2016) problems in one-way vehicle systems, where no resource repositioning takes place.

In Zhang and Pavone (2016) and Iglesias et al. (2016), authors use an empty-car routing mechanism to rebalance their ridesharing networks. However, each paper focuses on a single optimization problem, whereas our technical approach is more robust and allows us to consider a large class of optimization problems. Like us, Iglesias et al. (2016) also faces a non-linear optimization problem. To deal with it, the authors pass to the *infinite supply* regime where they obtain a simpler optimization problem that they solve and apply as a heuristic to the original problem from the finite system. The connection between the finite system and the limiting system in the infinite supply regime received rigorous treatment in Banerjee et al. (2016), where the authors establish bounds on the gap between the optimal values of the optimization problems of the finite and infinite supply systems.

The aforementioned papers Banerjee et al. (2016), Zhang and Pavone (2016), Iglesias et al. (2016) all share the same feature that the optimization problems of those papers equalize availabilities across all regions. This is either enforced via an explicit constraint Iglesias et al. (2016), (Zhang and Pavone 2016, Equation 10), or arises implicitly in the optimal solution of the approximating optimization problem in Banerjee et al. (2016). The rationale behind enforcing the equal availability constraint is that in the infinite supply regime, at least one region achieves 100% availability. Equalizing availabilities among regions then ensures 100% availability everywhere. Therefore, the solutions proposed in those papers rely on an abundance of vehicle supply in the system, and are unsuitable for problems where demand is comparable to, or even exceeds, supply. Indeed, in our 2-region example in Figure 1, the only way to equalize availabilities is to use the routing policy $Q_{12} = 0$, and $Q_{21} = 1/2$, and we see in Table 1 that this hurts system performance.

After submission of this paper, a newer version of Banerjee et al. (2016) appeared, namely Banerjee et al. (2017). In both versions, the authors study approximations of optimization problems arising in finite sized ridesharing systems. The authors directly analyze the stationary distribution of BCMP networks to establish asymptotically tight guarantees for their approximating optimization problems. In Banerjee et al. (2016), the focus is on pricing as a control mechanism in the infinite supply regime. Motivated by the asymptotic regime and control mechanism considered in this paper, the new version Banerjee et al. (2017) proves that in our Theorem 2, the ratio between the optimal value of the optimization problem for the finite-sized system and the optimal value of the fluid-based optimization goes to one at a rate of $1 - 1/\sqrt{N}$; see appendix D.1 there.

Fluid models have been used by Pavone et al. (2012), Waserhole and Jost (2013) to study ridesharing networks. Those fluid models are different from the fluid model in this paper. Furthermore, they are only used as heuristics, and are not shown to be connected to an underlying stochastic system. In contrast, our fluid model is proven to be the limit of the queue length process of our BCMP network, and all fluid optimization problems considered are rigorously proven to be the limits of their stochastic counterparts.

We also wish to highlight some papers that either consider problems closely related to ridesharing, or address some of the other aspects of the ridesharing problem not focused on in this paper. In Adelman (2007), the author considers the problem of managing a network of shipping containers. That paper uses a combination of optimization and approximate dynamic programming techniques to determine a policy to accept/reject requests for containers. Both Ma et al. (2013), Santos and Xavier (2015) study a carpooling problem of having multiple riders with different destinations share the same car. In Ozkan and Ward (2016), the authors adopt a matching approach to the setting where a passenger requesting a ride from a region with no available cars is willing to wait a little bit for a driver to arrive from a nearby region. They study the problem of matching drivers to passengers when the two may have two different initial locations. In Bimpikis et al. (2016), the authors consider the issue of pricing rides in a ridesharing network. Also related to the ridesharing problem is Yang et al. (2016), where the authors study a mean field equilibrium of a system where agents explore and compete for resources that are both time-varying and location-dependent in nature.

Outside of the ridesharing setting, fluid models are a widely used tool in the study of closed queueing networks; we refer the reader to Anselmi et al. (2013) for a recent discussion of the literature. For the type of network considered in this paper, i.e. a closed network with single-server and infinite-server stations, process level convergence to the fluid model was established in Krichagina (1992). The technical novelty of our paper lies in Theorem 4, which characterizes the limiting behavior of the fluid model. A related paper is Anselmi et al. (2013), which characterizes the limiting behavior of the fluid model corresponding to a multiclass, closed queueing network consisting entirely of single-server stations, with no infinite-server stations. The results in Anselmi et al. (2013) were established using a relative entropy based Lyapunov function. In this paper, we use the L_1 distance from the equilibrium as our Lyapunov function.

1.2. Contributions

The following is a summary of the main contributions of this paper.

- Fixing a static empty-car routing policy $Q = (Q_{ij})$, we consider a fluid model associated with a closed queueing network composed of single and infinite server stations. We establish process level convergence of the scaled queue length process in our closed queueing network to a fluid limit. The fluid model's equilibrium set is explicitly characterized, and we show that the fluid model converges to this equilibrium set from any initial starting condition. We then elevate the process level convergence result to convergence of steady-state distributions, cf. Section 4. One consequence of our result is an answer to the following question: what is the minimum number of cars in the system needed to achieve 100% availability (asymptotically as $N \to \infty$) everywhere? This discussion can be found in Appendix EC.4.

- To find an optimal static empty-car routing policy Q^*, we formulate a fluid-based optimization problem that is able to accommodate a broad class of utility functions. The latter can depend on availabilities at different regions, and fractions of both empty or occupied cars on different roads. Then Q^* can be solved efficiently by solving a related problem with only linear constraints, cf. Lemma 2.

- We prove in Theorem 2 that as the number of car grows to infinity, the routing policy Q^* from the fluid-based optimization is asymptotically optimal among all state dependent routing policies. For any MDP or stochastic control problem of realistic size, the true optimal policy is rarely known. Thus, any upper bound, particularly a tight bound is valuable to developing good policies.

The rest of the paper is structured as follows. In Section 2, we formulate the fluid-based optimization problem and state our main results, Theorems 1 and 2. In Section 3, we describe the numerical study performed using real-world data from Didi Chuxing, China's largest ridesharing company. Section 4 is devoted to studying the fluid model of the ridesharing network, and establishing the machinery needed to prove our main results. Section 5 concludes.

1.3. Notation

For a function $f : \mathbb{R} \to \mathbb{R}^n$, we use $\dot{f}(t)$ to denote the derivative of $f(t)$ when the derivative exists. For any integer $n > 0$, we use $\mathbb{D}^n$ to denote the space of all cadlag functions $x : \mathbb{R}_+ \to \mathbb{R}^n$, i.e.

functions that are right-continuous on $[0, \infty)$ with left limits on $(0, \infty)$. We define

$$\mathbb{D}_0^n = \{x \in \mathbb{D}^n : x(0) = 0\},$$

$$\mathbb{D}_1^n = \left\{x \in \mathbb{D}^n : x(0) \in [0, 1]^n \text{ and } \sum_{i=1}^n x_i(0) = 1\right\},$$

$$\mathbb{D}_{0+}^n = \{x \in \mathbb{D}^n : x(0) \geq 0\}.$$

For any $x \in \mathbb{D}^n$ and any $T > 0$, we define

$$\|x\|_T = \max_{1 \leq i \leq n} \sup_{0 \leq t \leq T} |x_i(t)| = \sup_{0 \leq t \leq T} \max_{1 \leq i \leq n} |x_i(t)|. \tag{2}$$

We let $C^n \subset \mathbb{D}^n$ be the subspace of continuous functions $x : \mathbb{R}_+ \to \mathbb{R}^n$, and define C_1^n analogously to $\mathbb{D}_1^n$. For any $x \in \mathbb{D}^n$, we write $\int_0^t x(s)ds$ to denote a vector in $\mathbb{R}^n$ whose ith component is $\int_0^t x_i(s)ds$. For a vector $a \in \mathbb{R}^n$, we use $|\cdot|$ to denote the max-norm, i.e. $|a| = \max_{1 \leq i \leq n} |a_i|$. For a set $A \subset \mathbb{Z}$, we write $|A|$ to denote the number of elements contained by this set. For a collection of random variables $Y, \{X_n\}_{n=1}^\infty$, we write $X_n \Rightarrow Y$ to denote weak convergence of X_n to Y (as $n \to \infty$).

2. The Ridesharing Optimization Problem

In this section we formally introduce the sequence of ridesharing networks discussed in the introduction. We then introduce the fluid-based optimization and state our main results, Theorems 1 and 2. We then show in Lemma 2 that the fluid-based optimization can be solved efficiently by solving a related optimization problem with linear constraints.

In our model, there are N cars serving r regions in a city. For any time $t \geq 0$, let $E_{ij}^{(N)}(t)$ be the number of empty cars en route from region i to region $j \neq i$, and let $E_{ii}^{(N)}(t)$ be the number of empty cars that are waiting in region i for a new passenger. Similarly, let $F_{ij}^{(N)}(t)$ be the number of full cars driving from region i to j (by full car we mean a car with a passenger). We allow $F_{ii}^{(N)}(t)$ to be non-zero, because a passenger's destination can be located in the same region as he was picked up. Let $E^{(N)}(t)$ and $F^{(N)}(t)$ be the $r \times r$ matrices whose (i, j)th elements are $E_{ij}^{(N)}(t)$ and $F_{ij}^{(N)}(t)$, respectively.

$$E^{(N)} = \{E^{(N)}(t) \in \mathbb{Z}_+^{r \times r}, \ t \geq 0\} \quad \text{and} \quad F^{(N)} = \{F^{(N)}(t) \in \mathbb{Z}_+^{r \times r}, \ t \geq 0\},$$

$$\bar{E}^{(N)} = \left\{ \frac{1}{N} E^{(N)}(t) \in \mathbb{R}_+^{r \times r}, \; t \geq 0 \right\} \quad \text{and} \quad \bar{F}^{(N)} = \left\{ \frac{1}{N} F^{(N)}(t) \in \mathbb{R}_+^{r \times r}, \; t \geq 0 \right\},$$

$$\mathcal{T} = \left\{ (e, f) \in [0,1]^{r \times r} \times [0,1]^{r \times r} : \sum_{i=1}^r \sum_{j=1}^r (e_{ij} + f_{ij}) = 1 \right\}. \tag{3}$$

Recall the dynamics introduced in Section 1. We allow the empty-car routing probabilities to be state-dependent, i.e. the empty-car routing probability matrix at time $t \geq 0$ is

$$Q(\bar{E}^{(N)}(t), \bar{F}^{(N)}(t)) = \left(Q_{ij}(\bar{E}^{(N)}(t), \bar{F}^{(N)}(t)) \right).$$

The process $(E^{(N)}, F^{(N)})$ is then a continuous time Markov chain (CTMC), whose transition properties are listed in Table 2. Going forward, we focus on the fluid-scaled CTMC $(\bar{E}^{(N)}, \bar{F}^{(N)})$ and

Table 2   Markov Chain Transition Rates

Rate	Transition	Description
$N\lambda_i P_{ij} 1(E_{ii}^{(N)}(t) > 0)$	$E_{ii}^{(N)}(t) - 1,$ $F_{ij}^{(N)}(t) + 1$	Passenger arrives to region i, and starts ride from i to j.
$\mu_{ij} F_{ij}^{(N)}(t) Q_{jk}(\bar{E}^{(N)}(t), \bar{F}^{(N)}(t))$	$F_{ij}^{(N)}(t) - 1,$ $E_{jk}^{(N)}(t) + 1$	Passenger dropped off at region j. Car stays put if $k = j$, or starts driving empty to region k if $k \neq j$
$\mu_{ij} E_{ij}^{(N)}(t) 1(j \neq i)$	$E_{ij}^{(N)}(t) - 1,$ $E_{jj}^{(N)}(t) + 1$	Empty car arrives to region j. Stays there until next passenger.

assume it has a single recurrent class. Since the state space is finite, this implies that the CTMC is positive recurrent. Let $(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty)) \in \mathcal{T}$ be the random element having the stationary distribution of $(\bar{E}^{(N)}, \bar{F}^{(N)})$. Let

$$A_i^{(N)}(\infty) = \mathbb{P}(\bar{E}_{ii}^{(N)}(\infty) > 0)$$

be the availability at region i, and let $A^{(N)}(\infty)$ be the r-dimensional vector with entries $A_i^{(N)}(\infty)$.

Our model assumes that a routing decision is made, the car is committed to the region until a passenger arrives to take it away (as opposed to being able to hop around empty between regions).

This assumption is reasonable in a setting with a centralized router, where the decision maker has a global view of the system and will therefore 'get it right the first time' when making the routing decisions. The incentive for cars to keep jumping around the network is further reduced by the assumption that passengers will not linger in the system if they cannot receive service immediately. This passenger behavior is realistic in settings when comparable transportation modes are available, e.g. hailing a yellow cab in Manhattan, or taking public transportation.

REMARK 1. The process $(E^{(N)}, F^{(N)})$ can also be interpreted as the queue length process in a closed queueing network of r single server stations and $2r^2 - r$ infinite server stations, where cars are the "jobs" in the network. For $1 \leq i \leq r$, the process $E_{ii}^{(N)} = \{E_{ii}^{(N)}(t), \ t \geq 0\}$ corresponds to a single server station with service rate $N\lambda_i$, and

$$E_{ij}^{(N)} = \{E_{ij}^{(N)}(t), \ t \geq 0\}, \quad 1 \leq i \neq j \leq r,$$

$$F_{ij}^{(N)} = \{F_{ij}^{(N)}(t), \ t \geq 0\}, \quad 1 \leq i, j \leq r,$$

correspond to infinite server stations where the service rate of each server at station $E_{ij}^{(N)}$ or $F_{ij}^{(N)}$ is μ_{ij}. In the special case when Q is not state-dependent, this network belongs to the class of BCMP networks Baskett et al. (1975). Note that in our model, the service time at a single-server station can begin even before a job enters the station, e.g. the inter-arrival "timer" of the next passenger is counting down regardless of whether there is an available car in the region or not. However, the memoryless property of passenger inter-arrival times makes our model equivalent to one where service starts only when the station is non-empty.

We are now ready to introduce the fluid-based optimization problem, and state our main results.

2.1. Main Results

Recall the network primitives λ, μ, P. We now consider the fluid-based optimization problem to be fully specified from (4a) to (4g) below. In the optimization problem, $c_{ij} > 0$ are rewards for completing a ride from i to j. The variables in the optimization problem are q, $\bar{e}$, $\bar{f}$, $\bar{a}$, where

$q = (q_{ij})$ is an $r \times r$ matrix representing a static empty-car routing policy Q and $(\bar{e}, \bar{f}, \bar{a})$ is a point

in $\mathcal{T} \times [0,1]^r$, whose interpretation will be given after the equations (4a)-(4g):

$$\max_{q, \bar{e}, \bar{f}, \bar{a}} \sum_{i=1}^{r} \sum_{j=1}^{r} \bar{a}_i \lambda_i P_{ij} c_{ij} \tag{4a}$$

subject to $\lambda_i P_{ij} \bar{a}_i = \mu_{ij} \bar{f}_{ij}, \quad 1 \le i, j \le r,$ (4b: full car i-j Little's Law)

$$\mu_{ij} \bar{e}_{ij} = q_{ij} \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}, \quad 1 \le i, j \le r, \ j \ne i, \qquad \text{(4c: empty car i-j Little's Law)}$$

$$\lambda_i \bar{a}_i = \sum_{k=1, k \ne i}^{r} \mu_{ki} \bar{e}_{ki} + q_{ii} \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}, \quad 1 \le i \le r, \qquad \text{(4d: car flow balance region i)}$$

$(1 - \bar{a}_i) \bar{e}_{ii} = 0, \quad 1 \le i \le r,$ (4e: availability to idle car relation)

$(\bar{e}, \bar{f}) \in \mathcal{T},$ (4f: unit mass)

$$q_{ij} \ge 0, \quad \sum_{j=1}^{r} q_{ij} = 1, \quad 0 \le \bar{a}_i \le 1, \quad 1 \le i, j \le r \tag{4g}$$

To help guide intuition, one can think of $\bar{e}, \bar{f}$, and $\bar{a}$ as placeholders for $\mathbb{E}[\bar{E}^{(N)}(\infty)], \mathbb{E}[\bar{F}^{(N)}(\infty)]$

and $A^{(N)}(\infty)$, respectively. We can interpret $\bar{a}_i \lambda_i P_{ij}$ as the rate at which rides are initialized from

i to j, and since a ride from i to j has a reward of c_{ij}, the problem above aims to maximize revenue

generation. Our results actually hold for a much larger class of utility functions; cf. Remark 5 in

Section 2.2. The constraints in (4b) are simply Little's Laws for the number of occupied cars on

the road from i to j in equilibrium. That is, $\lambda_i P_{ij} \bar{a}_i$ is the rate at which rides are initialized, which

equals the mass of occupied cars on the road f_{ij} divided by the average travel time $1/\mu_{ij}$. Another

interpretation is that the inflow $\lambda_i P_{ij} \bar{a}_i$ into the infinite server station equals the outflow $\mu_{ij} \bar{f}_{ij}$.

Similarly, the constraints in (4c) are also Little's Laws for the number of empty cars travelling from

i to j; the rate at which empty cars start their journey is $q_{ij} \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}$. The constraints in (4d)

say that the total rate of outflow from region i, given by $\lambda_i \bar{a}_i$, must equal the total inflow into the

region, given by the right hand side of (4d). Constraints (4e) state that a shortage of availability

$(1 - \bar{a}_i > 0)$ is only possible when the fraction of cars at the region equals zero ($\bar{e}_{ii} = 0$), because

there are not enough cars to meet demand. Conversely, having a positive mass of cars at a region

$(\bar{e}_{ii} > 0)$ implies that all passenger requests are satisfied there $(1 - \bar{a}_i = 0)$. Additional intuition can

be gained once the fluid model is introduced and its equilibrium behavior discussed in Section 4, where the reason why (4a)–(4g) is called the fluid-based optimization problem becomes apparent. Finally, although this optimization problem is stated for static empty-car routing policies, the connection to state-dependent polices will be made in Theorem 2.

The following are our main results. The first establishes the connection between the fluid-based optimization problem and $(\bar{E}^{(N)}, \bar{F}^{(N)})$. The second shows that asymptotically, the optimal static policy from the fluid-based optimization outperforms all state dependent policies. The ingredients to prove Theorem 1 are developed in Section 4, and the proof is left to Appendix EC.5.3. Theorem 2 is proved in Appendix EC.5.4.

THEOREM 1. *Let $q, \bar{e}, \bar{f}, \bar{a}$ be a feasible solution to the optimization problem in (4a)–(4g). Set $Q = q$. Assume $P_{ij} > 0$ for all $1 \leq i, j \leq r$ and $q_{ii} > 0$ for all $1 \leq i \leq r$. Then*

$$\bar{F}^{(N)}(\infty) \Rightarrow \bar{f}, \tag{5}$$

$$\bar{E}_{ij}^{(N)}(\infty) \Rightarrow \bar{e}_{ij}, \quad 1 \leq i \neq j \leq r, \tag{6}$$

$$\bar{E}_{ii}^{(N)}(\infty) \Rightarrow 0, \quad \text{for } i \text{ such that } \bar{a}_i < 1, \tag{7}$$

$$\sum_{i:\bar{a}_i=1} \bar{E}_{ii}^{(N)}(\infty) \Rightarrow \sum_{i:\bar{a}_i=1} \bar{e}_{ii}, \tag{8}$$

and

$$\mathbb{P}(E_{ii}^{(N)}(\infty) > 0) \to \bar{a}_i, \quad 1 \leq i \leq r, \tag{9}$$

as $N \to \infty$.

REMARK 2. The assumptions that $P_{ij} > 0$ for all i, j and $q_{ii} > 0$ for all i are made to facilitate exposition in the proof of Theorem 4, which plays a central role in establishing Theorem 1. We expect that Theorem 4, and hence Theorem 1, holds under the simpler assumption that $(\bar{E}^{(N)}, \bar{F}^{(N)})$ has a single recurrent class. Assuming $q_{ii} > 0$ for all i is not very restrictive, as it simply means that a driver has a positive probability to stay in a region after dropping off a passenger there.

REMARK 3. The exponentially distributed travel time assumption is non-essential. We know that $\left(E^{(N)}, F^{(N)}\right)$ is a BCMP network, and in full generality, BCMP networks only require that the service time distributions in the infinite server stations have rational Laplace transforms Baskett et al. (1975). This class of distributions is dense in the set of all probability distributions on $(0, \infty)$ Asmussen (2003). Furthermore, stationary distribution of BCMP networks is known to depend only on the service rates at the stations, and not on the entire distribution. Hence, the results of Theorem 1 hold for travel time distributions with rational Laplace transforms.

THEOREM 2. *(a) Suppose $(\bar{E}^{(N)}, \bar{F}^{(N)})$ has a single recurrent class under P and Q, where Q can be a state-dependent empty-car routing policy. Let $(q^*, \bar{e}^*, \bar{f}^*, \bar{a}^*)$ be an optimal solution of the optimization problem in (4a)–(4g). Then*

$$\sum_{i=1}^{r} \sum_{j=1}^{r} A_i^{(N)}(\infty) \lambda_i P_{ij} c_{ij} < \sum_{i=1}^{r} \sum_{j=1}^{r} \bar{a}_i^* \lambda_i P_{ij} c_{ij}, \quad N > 0.$$

(b) Let $(\bar{E}^{(N)*}, \bar{F}^{(N)*})$ denote the CTMC under the static routing policy q^*. If $P_{ij} > 0$ for all $1 \leq i, j \leq r$ and $q_{ii}^* > 0$ for all $1 \leq i \leq r$, then

$$\lim_{N \to \infty} \sum_{i=1}^{r} \sum_{j=1}^{r} A_i^{(N)*}(\infty) \lambda_i P_{ij} c_{ij} = \sum_{i=1}^{r} \sum_{j=1}^{r} \bar{a}_i^* \lambda_i P_{ij} c_{ij}.$$

REMARK 4. Part (a) of Theorem 2 states that the optimal value of the fluid-based optimization problem (4a)–(4g) is a strict upper bound on the expected system utility of the system with N cars under any state-dependent routing policy under which the CTMC has a single recurrent class. Part (b) states that the upper bound is asymptotically achievable under the static routing policy q^* if $P_{ij} > 0$ for all $1 \leq i, j \leq r$ and $q_{ii}^* > 0$ for all $1 \leq i \leq r$.

2.2. Efficient Solution of the Fluid-Based Optimization

Having established the relevance of the fluid-based optimization, we now discuss how to solve it efficiently. The main issue with solving (4a)–(4g) is the presence of bilinear constraints, e.g. (4c) and (4d) are bilinear in q and $\bar{f}$, and in (4e), the bilinearity is in $\bar{a}$ and $\bar{e}$. In this section we show that the problem can be transformed into one with only linear constraints. Our first step is the following lemma, which transforms the constraints in (4c) and (4d). It is proved in Appendix EC.5.1.

Lemma 1. *Consider the set of constraints*

$$\lambda_i P_{ij} \bar{a}_i = \mu_{ij} \bar{f}_{ij}, \quad 1 \le i, j \le r, \tag{10a}$$

$$\mu_{ij} \bar{e}_{ij} \le \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}, \quad 1 \le i, j \le r, \ j \ne i, \tag{10b}$$

$$\sum_{k=1, k \ne i}^{r} \mu_{ki} \bar{e}_{ki} \le \lambda_i \bar{a}_i \le \sum_{k=1, k \ne i}^{r} \mu_{ki} \bar{e}_{ki} + \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}, \quad 1 \le i \le r, \tag{10c}$$

$$\lambda_i \bar{a}_i + \sum_{j=1, j \ne i}^{r} \mu_{ij} \bar{e}_{ij} = \sum_{k=1, k \ne i}^{r} \mu_{ki} \bar{e}_{ki} + \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}, \quad 1 \le i \le r, \tag{10d}$$

$$(\bar{e}, \bar{f}) \in \mathcal{T}, \tag{10e}$$

$$0 \le \bar{a}_i \le 1, \quad 1 \le i \le r, \tag{10f}$$

$$(1 - \bar{a}_i) \bar{e}_{ii} = 0, \quad 1 \le i \le r. \tag{10g}$$

If $(\bar{e}, \bar{f}, \bar{a})$ and q satisfy (4b)–(4g), then $(\bar{e}, \bar{f}, \bar{a})$ satisfy (10a)–(10g). Conversely, suppose $(\bar{e}, \bar{f}, \bar{a})$ satisfy (10a)–(10g) and let

$$q_{ij} = \frac{\mu_{ij} \bar{e}_{ij}}{\sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}}, \quad 1 \le i \ne j \le r, \quad q_{ii} = \frac{\lambda_i \bar{a}_i - \sum_{k=1, k \ne i}^{r} \mu_{ki} \bar{e}_{ki}}{\sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}} \quad 1 \le i \le r. \tag{11}$$

Then $(\bar{e}, \bar{f}, \bar{a})$ and q satisfy (4b)–(4g).

With the help of this lemma, the fluid-based optimization problem can be rewritten as

$$\max_{\bar{e}, \bar{f}, \bar{a}} \sum_{i=1}^{r} \sum_{j=1}^{r} \bar{a}_i \lambda_i P_{ij} c_{ij} \tag{12}$$

$$\text{subject to: } (10a) - (10g). \tag{13}$$

Observe that (10a)–(10f) are all linear constraints, and that only (10g) is bilinear. The following result says that we can safely ignore (10g). It is proved in Appendix EC.5.2.

Lemma 2. *Consider the relaxed optimization problem*

$$\max_{\bar{e}, \bar{f}, \bar{a}} \sum_{i=1}^{r} \sum_{j=1}^{r} \bar{a}_i \lambda_i P_{ij} c_{ij} \tag{14}$$

$$\text{subject to: } (10a) - (10f), \tag{15}$$

and let $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$ be an optimal solution.

1. *If $\bar{a}_i^* < 1$ for all $1 \leq i \leq r$, then $\bar{e}_{ii}^* = 0$ for all $1 \leq i \leq r$, implying $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$ also satisfies* (10g).

2. *Otherwise, choose any i' such that $\bar{a}_{i'}^* = 1$, define $\tilde{e}$ by letting $\tilde{e}_{ij} = \bar{e}_{ij}^*$ for all $1 \leq i \neq j \leq r$,*

$$\tilde{e}_{i'i'} = \sum_{i=1}^{r} \bar{e}_{ii}^*, \quad and \quad \tilde{e}_{ii} = 0, \quad i \neq i'.$$

Then $(\tilde{e}, \bar{f}^*, \bar{a}^*)$ is a feasible solution that yields the same objective value as $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$, and also satisfies (10g).

Solving the fluid-based optimization can be broken down into the following procedure. We first solve (14)–(15), which has $(4r^2 + 2r + 1)$ linear constraints and $2r^2 + r$ variables, and can be solved efficiently using any standard linear program solver. We then modify the solution using Lemma 2 (if needed) so that it also satisfies (10g), and recover the optimal routing policy using (11).

REMARK 5. So far, we only considered the utility function defined in (4a). In fact, both Lemma 2 and Theorem 2 can be extended to hold for any function $U(\bar{e}, \bar{f}, \bar{a})$ that is

(a) nondecreasing in $\bar{a}_i$ for all i,

(b) nondecreasing in $\bar{f}_{ij}$ for all i and j,

(c) nonincreasing in $\bar{e}_{ij}$ for all $i \neq j$,

(d) independent of $\bar{e}_{ii}$ for all i, and

(e) concave in $(\bar{e}, \bar{f})$.

Lemma 2 can be extended because its proof relies only on the objective value in (14) satisfying conditions (a)–(d). Theorem 2 can be extended to say that

$$\mathbb{E}\big[U(A^{(N)}(\infty), E^{(N)}(\infty), F^{(N)}(\infty)\big] \leq U\big(A^{(N)}(\infty), \mathbb{E}\big[E^{(N)}(\infty)\big], \mathbb{E}\big[F^{(N)}(\infty)\big]\big) \leq U(\bar{a}^*, \bar{e}^*, \bar{f}^*),$$

where the first inequality follows from Jensen's inequality and (e), and the second inequality is proved by repeating the proof of Theorem 2.

3. A Numerical Study

This section is devoted to a numerical study of our empty-car routing policy. To ground the study, we use a data set obtained from a data challenge by the Didi Research Institute (DRI 2016). Using

the data set, we extract a nine-region network and the associated realistic parameters N, μ, λ, and P. For more details about the dataset and how we obtained our parameters, see Appendix EC.3.1.

Figure 2 shows order fulfillment levels for each of the nine regions during the 5PM-6PM evening rush hour, plotted over 21 days. Each point represents the total number of orders received, and orders fulfilled during that one hour window. We can see that three of the nine regions, regions 13, 47, and 50, had significant supply shortages in most of the 21 days, while most orders in the remaining six regions were fulfilled. Our data did not permit us to deduce the surplus of drivers in those six regions, but these figures illustrate the significance of a good empty-car routing policy.
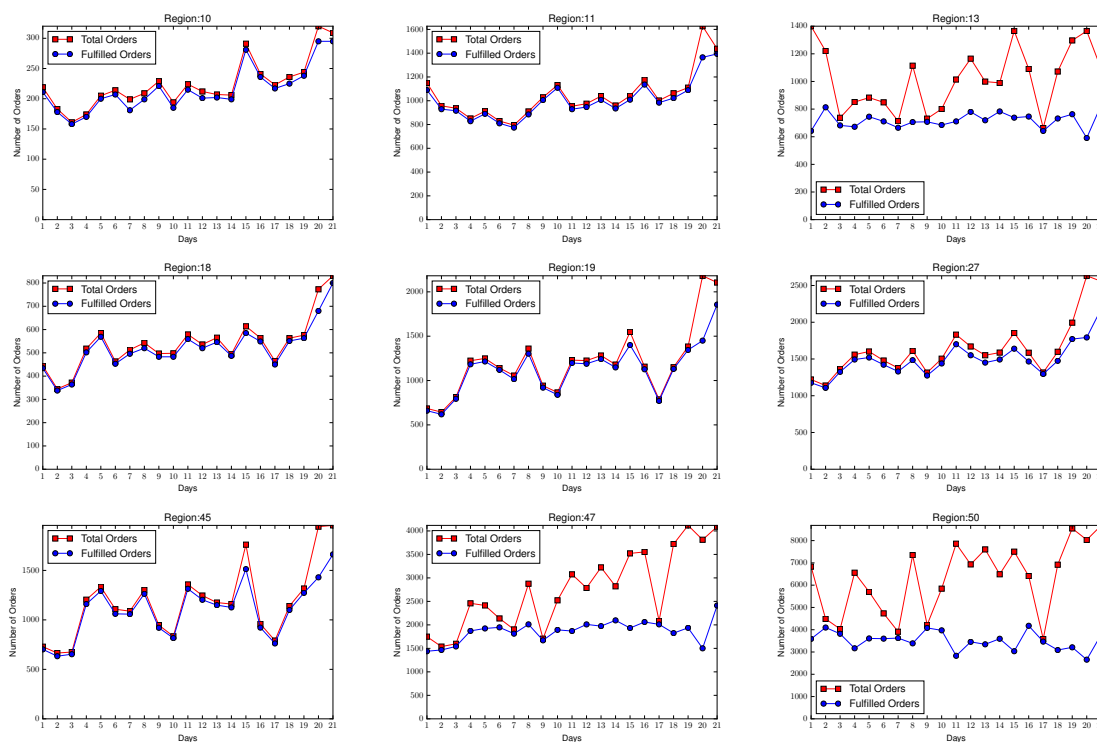

Figure 2   The gaps between the number of passenger orders and the number of fulfilled orders of the nine regions

The rest of the section is structured as follows. We first compare the fluid-based optimal routing policy to a state-dependent routing policy in Section 3.1. We then use the fluid-based optimization to propose a lookahead policy to deal with cases where the system parameters are not constant over time in Section 3.2. Lastly, we perform some numerical robustness tests in Section 3.3 to see how the fluid-based policy performs when there is estimation noise present in the parameters.

20

Braverman, Dai, Liu and Ying: *Empty-car Routing in Ridesharing Systems*
Article submitted to *Operations Research*; manuscript no. (Please, provide the manuscript number!)

3.1. Performance Comparison with Dynamic Routing Polices

In Theorem 2, we showed that the expected utility under any state-dependent routing policy for the finite-sized system is upper bounded by the optimal utility of the fluid-based optimization. However, the question remains open whether a state-dependent routing policy can outperform the static routing policy Q^* for a finite-sized system. Choosing $c_{ij} = 1/\sum_{i=k}^{r} \lambda_k$ in (4a), we consider the utility function

$$U(\bar{e}, \bar{f}, \bar{a}) = U(\bar{a}) = \frac{\sum_{i=1}^{r} \bar{a}_i \lambda_i}{\sum_{i=1}^{r} \lambda_i}. \tag{16}$$

This can be thought of as the probability that a passenger requesting a ride at *any* region is fulfilled. It is impossible to consider all possible dynamic routing policies, so we focus on the following two intuitive heuristics.

Join-the-Least-Congested-Region with Threshold η (JLCR-η): When a car drops off a passenger at region i at time t, the driver stays at region i if

$$(1-\eta)\frac{\sum_k E_{ki}^{(N)}(t)}{\lambda_i} \leq \min_{j=1, j \neq i} \frac{\sum_k E_{kj}^{(N)}(t)}{\lambda_j}. \tag{17}$$

Otherwise, the driver drives empty to region j^*, where

$$j^* \in \arg \min_{j=1, j \neq i} \frac{\sum_k E_{kj}^{(N)}(t)}{\lambda_j}. \tag{18}$$

Ties are broken uniformly at random. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

To understand the JLCR-η policy, we note that $\sum_k E_{ki}^{(N)}(t) = E_{ii}^{(N)}(t) + \sum_{k \neq j} E_{ki}^{(N)}(t)$ is the number of empty cars both currently waiting and en-route to region i. Therefore, $\frac{\sum_k E_{ki}^{(N)}(t)}{\lambda_i}$ is a measure of congestion, in terms of empty cars, at region i. When $\eta = 0$, the policy routes empty cars to the least congested region. However, such a policy can be wasteful if congestion levels among regions are similar, because it takes time for a car to go from one region to another. We therefore introduce the threshold η such that a driver drives empty from i to j only if the difference in congestion levels surpasses $\eta\frac{\sum_k E_{ki}^{(N)}(t)}{\lambda_i}$. We test the policy on our 9-region network with parameters as in (EC.62)-(EC.64). With η ranging from 0 to 1, we find that JLCR-η performs best when η

is around 0.5. In addition to JLCR, we also consider the following policy where drivers aim to minimize the time until their next pickup.

Shortest-Wait (SW): When a car drops off a passenger at region i at time t, the driver stays at region i if

$$\frac{E_{ii}^{(N)}(t)}{N\lambda_i} \leq \min_{j \neq i} \frac{1}{\mu_{ij}} + \frac{\left(E_{jj}^{(N)}(t) + \frac{1}{\mu_{ij}}\sum_{k \neq j}\mu_{kj}E_{kj}^{(N)}(t) - \frac{N\lambda_j}{\mu_{ij}}\right)^+}{N\lambda_j},$$

Otherwise, the driver drives empty to region that minimizes the right hand side above. Ties are broken uniformly at random. $\qquad\square$

The intuition behind the SW policy is that drivers want to minimize the time to get their next passenger. Since passengers arrive to region i every $1/N\lambda_i$ time units (on average), then $E_{ii}^{(N)}(t)/N\lambda_i$ is a proxy (we have not assumed any priority scheme for choosing how to allocate passengers between multiple cars in the same region) for the amount of time it will take until the driver gets a passenger if he stays in region i. If the driver chooses to go to region j, then the time taken until he gets a passenger is the sum of the travel time $1/\mu_{ij}$ and the time spent idling in region j. We use $\left(E_{jj}^{(N)}(t) + \frac{1}{\mu_{ij}}\sum_{k \neq j}\mu_{kj}E_{kj}^{(N)}(t) - \frac{N\lambda_j}{\mu_{ij}}\right)^+/N\lambda_j$ as a proxy for the latter quantity: $E_{jj}^{(N)}(t)$ is the number of cars idling in region j at the decision point, $\frac{1}{\mu_{ij}}\sum_{k \neq j}\mu_{kj}E_{kj}^{(N)}(t)$ estimates the number of cars to arrive to j, and $\frac{N\lambda_j}{\mu_{ij}}$ estimates the number of cars that will leave the region due to passenger arrivals by the time the driver makes it from i to j.

Figure 3 compares the static routing policy under Q^* to SW and JLCR-η with different values of η in the 9-region network. In particular, we included JLCR policies with

- $\eta = 0$: Under JLCR-0, an empty car always goes to the least congested region.

- $\eta = 1$: Under JLCR-1, after a car drops off a passenger, it always stays at its current region.

- $\eta = 0.5$: JLCR-0.5 maximizes system-wide availability among all JLCR-η when $N = 2,000$.

A few remarks are in order. The figure confirms that static routing with Q^* outperforms both the SW and JLCR-η family of policies. However, a typical quality of state-dependent policies is robustness to system parameters. In our case, computing Q^* requires knowledge of λ, μ, and

P, whereas a JLCR-0.5 only requires knowledge of λ. Robustness to parameters is a particularly important quality when one only has noisy observations of the true parameters, or when the true parameters change over time. The objective of this paper is not to pursue optimal state-dependent policies. Rather, it is to establish an initial, rigorous theoretical foundation for the study of ridesharing networks. Our optimal static policy can then be used as a benchmark against which one compares the performance of other routing policies. As a case in point, had Figure 3 not included the performance of static-routing under Q^*, it would have been impossible to say whether JLCR-η^* was a good policy or not.
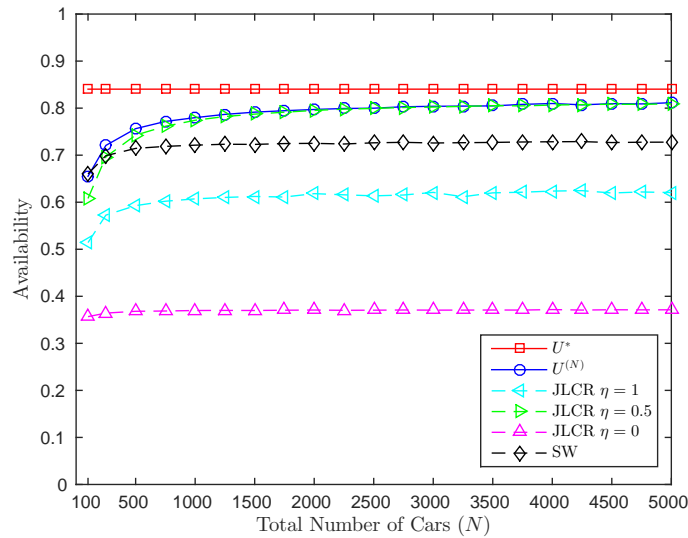

Figure 3   Performance comparison between the static routing, JLCR and SW policies in the nine-region network. In the plot, U^* is the optimal utility (16) from the fluid-based optimization, and $U^{(N)}$ is the steady-state utility in the finite sized system under optimal static routing.

3.2. Time-varying Parameters and the Lookahead Heuristic

The effectiveness of the fluid-based empty-car routing policy rests on the assumptions that a) parameters remain constant over time, and b) that the system has reached equilibrium. In practice a) is violated, e.g. when a rush hour starts. A common solution is to divide the day into periods of time where parameters are assumed constant, and treat the system as if it were in steady-state

during each of those periods Green et al. (2007). This approach can yield good results provided our time windows are long enough for the system to converge to equilibrium in each of them.

The Didi dataset we use suggests it is reasonable to assume constant parameters over time windows 1-2 hours in length; see Figure 4. However, numerical experiments suggest that for certain
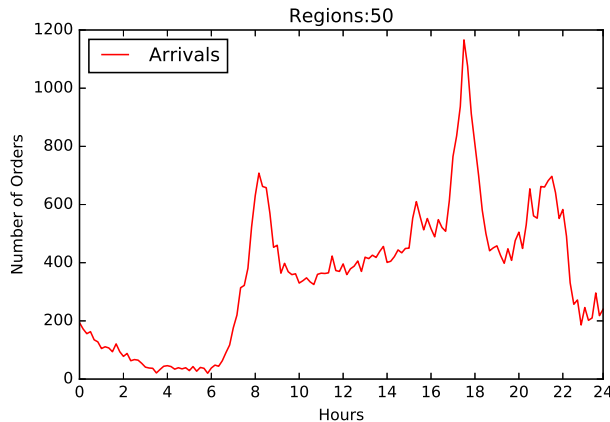

Figure 4   Ride requests on January 5, 2016 in region 50 of the Didi dataset. The x-axis shows the time of day, with 0 being midnight and 12 being noon.

choices of parameters and initial conditions, convergence of our system to equilibrium can occur on timescales on the order of 10 hours. With the rate at which parameters vary, and the slow convergence to equilibrium, the system never really reaches steady-state. To address this, we now propose a lookahead policy that is grounded in the fluid-based optimization problem. The purpose is to demonstrate that our fluid model framework, even though developed in a stationary environment, can potentially be used for other purposes.

The T-lookahead policy: Recall the per-ride rewards c_{ij} from (4a), and suppose that they can depend on λ, μ, and P, i.e. $c_{ij} = c_{ij}(\lambda, \mu, P)$ (like in (16)). Given time-varying parameter $\left\{ \lambda(t), \mu(t), P(t) \right\}_{t \geq 0}$, the per-ride rewards c_{ij} also depend on time; for simplicity, let us write $c_{ij}(t)$ instead of $c_{ij}(\lambda(t), \mu(t), P(t))$. Instead of using the time-independent routing matrix q^* from the fluid-based optimization problem (4a)–(4g), at time $t \geq 0$ we use the routing matrix $q^*(t)$ that solves

$$\max_{q, \bar{e}, \bar{f}, \bar{a}} \frac{1}{T} \int_0^T \sum_{j=1}^r \bar{a}_i \lambda_i(t+u) P_{ij}(t+u) c_{ij}(t+u) du$$

24

Braverman, Dai, Liu and Ying: *Empty-car Routing in Ridesharing Systems*
Article submitted to *Operations Research*; manuscript no. (Please, provide the manuscript number!)

$$\text{subject to} \quad \frac{1}{T}\int_0^T \lambda_i(t+u)P_{ij}(s+u)\bar{a}_i du = \frac{1}{T}\int_0^T \mu_{ij}(s+u)\bar{f}_{ij}du, \quad 1 \le i,j \le r,$$

$$\frac{1}{T}\int_0^T \mu_{ij}(t+u)du\bar{e}_{ij} = q_{ij}\sum_{k=1}^r \frac{1}{T}\int_0^T \mu_{ki}(t+u)\bar{f}_{ki}du, \quad 1 \le i,j \le r, \ j \ne i,$$

$$\frac{1}{T}\int_0^T \lambda_i(t+u)\bar{a}_i du$$

$$= \sum_{k=1,k\ne i}^r \frac{1}{T}\int_0^T \mu_{ki}(t+u)\bar{e}_{ki}du + q_{ii}\sum_{k=1}^r \frac{1}{T}\int_0^T \mu_{ki}(t+u)\bar{f}_{ki}du, \quad 1 \le i \le r,$$

$$(1-\bar{a}_i)\bar{e}_{ii} = 0, \quad 1 \le i \le r,$$

$$(\bar{e},\bar{f}) \in \mathcal{T}, \quad q_{ij} \ge 0, \quad \sum_{j=1}^r q_{ij} = 1, \quad 0 \le \bar{a}_i \le 1, \quad 1 \le i,j \le r.$$

The above problem is solved in the exact same way as the fluid-based optimization. In other words, the T-lookahead policy is a time-varying routing policy that uses paramater averages over the time window $[t,t+T]$ to make a decision at time t. In theory, $q^*(t)$ can be computed in real time at any decision point. In our numerical results, we discretize time into Δ-spaced intervals and use $q^*(k\Delta)$ as the routing decision for all times $t \in [k\Delta,(k+1)\Delta]$, where k is some non-negative integer.

We now consider some example networks with time-varying parameters, and compare the T-lookahead policy to both the standard fluid-based routing policy, and the state-dependent JLCR and SW policies from Section 3.1. Since ridesharing trip lengths are typically on the order of 10-30 minutes, we choose T to equal 30 and 45 minutes in our examples. This corresponds to looking ahead a few trips into the future.

3.2.1. The 5-Region City We consider a simplified model of a city that consists of 5 regions: a downtown area, a midtown area, and three suburban areas. We take a typical evening, from 5pm-11pm, and divide it into three 2-hour slots each having different parameters λ, μ, P to represent different traffic patterns. The details of this 5-region network are provided in Appendix EC.3.2.

Table 3 displays the results of a simulation that compares the system-wide availability introduced in (16) under several routing policies: 'standard fluid', JLCR-0.5, SW, 30 minute lookahead, and 45 minute lookahead. The 'standard fluid' policy treats the system as if it were in equilibrium during each of the 2-hour time slots. It solves three separate fluid-based optimization problems, one for

each of the three 2-hour time slots, and uses the resulting routing policy in the appropriate time slot. The JLCR-0.5 and SW policies are used because they adapt quickly to changes in parameters due to their state-dependent nature.

At 5pm, the system is initialized by assuming all cars are idle and distributing them across regions proportionally with the expected demand in the region, i.e. region i gets $\lambda_i / \sum_j \lambda_j$ of the cars, where λ are the 5pm arrival rates. The results are consistent with what we would expect, and in our experiments, the performance of the lookahead policy was consistent across different choices of initial system configurations (e.g. starting system according to fluid equilibrium). The standard fluid policy performs very well during the first two hours because the system was initialized with a favorable initial condition, and parameters stay constant during that time. However, the performance of this policy degrades as soon as the parameter change happens at 7pm. The explanation for this is simple: operating under the 5-7pm parameters for the first two hours puts the system in a state that is very far from the 7-9pm fluid equilibrium, and performance suffers as a result. Unlike the standard fluid policy, the lookahead policies anticipate parameter changes and prepare for them by pre-positioning the cars in the system into a more favorable state. For example, by sacrificing 4% in performance during 6-7pm, the 45 minute lookahead policy increases performance during 7-8pm to 86% from the 67% of the standard fluid policy. Factoring in the different ride request rates during the above two hours, the 4% performance drop corresponds to $0.04 * N \sum_i \lambda_i^{6-7\text{pm}} \approx 61$ customers lost, while the 19% increase during 7-8pm means an extra 433 customers served.

3.2.2. The 9-Region Didi Network Here we present some numerical results based on the 9-region network detailed in Appendix EC.3.1. In our first experiment, we consider a 4-hour period where passenger arrivals are constant for the first two hours, and then change abruptly in the last two hours. Namely, for the first two hours we let our arrival rates equal 0.3λ, where λ is as in (EC.64), and for the last two hours we use $0.85\lambda^p$, where

$$\lambda^p = \Big(0.0131 \ 0.0624 \ 0.1178 \ 0.0870 \ 0.0652 \ 0.0381 \ 0.0762 \ 0.2751 \ 0.1438 \Big)$$

26

Braverman, Dai, Liu and Ying: *Empty-car Routing in Ridesharing Systems*
Article submitted to *Operations Research*; manuscript no. (Please, provide the manuscript number!)

	5-6pm	6-7pm	7-8pm	8-9pm	9-10pm	10-11pm	Total 5-11pm
Standard Fluid	0.93	0.82	0.67	0.66	0.76	0.72	0.75
JLCR-0.5	0.91	0.73	0.74	0.75	0.97	0.66	0.79
SW	0.92	0.77	0.74	0.75	0.94	0.78	0.81
T-Lookahead, $T = 0.5$	0.93	0.82	0.82	0.8	0.96	0.71	0.84
T-Lookahead, $T = 0.75$	0.93	0.78	0.86	0.76	0.93	0.76	0.83

Table 3   Comparing the performance of the lookahead policy ($N = 1000$). The time unit for T is hours and the

time-discretization Δ is chosen to be one minute. Based on the utility function in (16), the values displayed are the

fraction of fulfilled ride requests. We used constant travel times, but the results remain qualitatively unchanged if

the travel times are exponentially distributed.

is a permutation of λ. With this experiment, we want to create a case where a) the passenger

request rate in the last two hours is significantly higher than the first two, and b) where the sources

of passenger requests change significantly. Similar to the 5-region city example, we initialize cars in

our system in proportion to $\lambda_i / \sum_j \lambda_j$, and we present our results in Table 4. The results again show

that the performance of the standard fluid policy suffers right after the change in parameters and

that the lookahead policy corrects for this. In this example, using the longer 45 minute lookahead

window does not degrade performance during hour 2 when compared to the 30 minute window (or

even the standard fluid). This happens because there is an excess supply of cars during the first

two hours, e.g. availabilities are almost 100%. This excess allows one to act suboptimally in hour

2 to preposition cars for hour 3 without hurting performance.

3.3. Robustness

The optimal fluid-based routing policy depends on perfect knowledge of system parameters λ, P, μ.

In practice, one uses estimates of these parameters, and the estimates contain noise. In this section

we provide several numerical examples to get a sense of the robustness of our optimal fluid-based

routing policy. The setup is the following: given a set of true parameters λ, μ, P and the utility

function in (16), we compute the optimal routing policy Q^*. Let $\sigma > 0$, and let η_{ij} and ξ_{ij} be

	Hour 1	Hour 2	Hour 3	Hour 4	4-hour Total
Standard Fluid	0.995	0.988	0.705	0.759	0.8
JLCR-0.5	1	1	0.715	0.717	0.79
SW	1	1	0.67	0.641	0.745
T-Lookahead, $T = 0.5$	0.995	0.991	0.759	0.771	0.824
T-Lookahead, $T = 0.75$	0.995	0.993	0.788	0.779	0.838

Table 4   Comparing the performance of the lookahead policy ($N = 2000$). The time unit for T is hours and the time-discretization Δ is chosen to be one minute. Based on the utility function in (16), the values displayed are the fraction of fulfilled ride requests. We used constant travel times, but the results remain qualitatively unchanged if the travel times are exponentially distributed.

i.i.d. random variables that equal 1 or -1 with equal probability. To simulate noisy estimates of passenger request rates along a route $\lambda_i P_{ij}$ and mean travel times $1/\mu_{ij}$, we use

$$\lambda_i P_{ij}(1 + \sigma \eta_{ij}), \quad \text{and} \quad (1/\mu_{ij})(1 + \sigma \xi_{ij}), \quad 1 \leq i, j \leq r,$$

respectively. We use four sets of parameters: the first set corresponds to the 9-region network (EC.62)–(EC.64), and the other three sets correspond to the three parameters sets used with the 5-region model in Section 3.2.1. Given a true parameter set and a realization of the η_{ij} and ξ_{ij}'s, we compute a routing matrix $\hat{Q}^*$ based on the fluid optimization with the noisy parameters. To obtain a measure of suboptimality, we then evaluate (16) with the true parameters λ, μ, P but with routing matrix $\hat{Q}^*$. We treat the objective value under $\hat{Q}^*$ as a random variable (it depends on η_{ij} and ξ_{ij}), and report its mean and standard deviation in Table 5. We see that 5% and 10% estimation errors lead to policies that are approximately 5% and 10% suboptimal, respectively. Outside these four examples, the sensitivity of the optimal solution will depend on the choice of parameters. However, the table gives us an idea of the magnitude of the suboptimality due to estimation error.

In the next section we introduce the fluid model together with the tools needed to prove Theorems 1 and 2.

Parameter Set	Optimal Fluid	$\hat{Q}^*$ Performance $\sigma = 0.05$ (mean, std. dev.)	Relative Suboptimality $(\sigma = 0.05)$	$\hat{Q}^*$ Performance $(\sigma = 0.1)$ (mean, std. dev.)	Relative Suboptimality $(\sigma = 0.1)$
5-region, 5-7pm	0.91	(0.87,0.024)	0.044	(0.82,0.044)	0.099
5-region, 7-9pm	0.92	(0.88,0.015)	0.043	(0.84,0.028)	0.087
5-region, 9-11pm	0.92	(0.86,0.031)	0.065	(0.81,0.053)	0.120
9-region	0.8403	(0.818868, 0.01503)	0.021478	(0.800520, 0.029091)	0.039826

Table 5   The third and fifth columns report the performance of the routing policy obtained from noisy

parameters, $\hat{Q}^*$, under the true parameters λ, μ, P. We ran 1000 replications, where in each replication we generated

a value for η_{ij} and ξ_{ij}; $2r^2$ random variables per replication in total.

4. The Fluid Model

In this section we study the fluid model of the ridesharing network in Section 2 and prepare all the

ingredients needed to prove Theorem 1. We start by introducing the fluid model for static empty-car

routing matrices Q, and establish process-level convergence in Theorem 3. In Section 4.1, we then

characterize the fluid model's set of equilibria. This equilibrium behavior motivates the fluid-based

optimization in Section 2. Theorem 1 then follows from a standard argument involving process-level

convergence and the convergence of the fluid model to its equilibrium, and we therefore relegate it

to Appendix EC.5.3.

Recall the primitive parameters λ, μ, P, and assume a static empty-car routing matrix Q is given.

Recall the set $\mathcal{T}$ defined in (3). Let

$$I_i^{(N)}(t) = \int_0^t 1\big(E_{ii}^{(N)}(s) = 0\big)ds$$

be the cumulative idle time of the single-server station corresponding to $E_{ii}^{(N)}$. The following is a

process-level convergence result for $(\bar{E}^{(N)}, \bar{F}^{(N)})$, and is proved in Appendix EC.1.

THEOREM 3. *Assume* $\big(\bar{E}^{(N)}(0), \bar{F}^{(N)}(0)\big) \Rightarrow (e(0), f(0)) \in \mathcal{T}$ *as* $N \to \infty$. *There exists a unique solu-*

tion $(e, f) : \mathbb{R}_+ \to \mathcal{T}$ *and* $u : \mathbb{R}_+ \to \mathbb{R}_+^r$ *to the dynamical system*

$$f_{ij}(t) = f_{ij}(0) + \lambda_i P_{ij}\big(t - u_i(t)\big) - \mu_{ij} \int_0^t f_{ij}(s)ds, \qquad\qquad 1 \le i, j \le r, \qquad (19)$$

$$e_{ij}(t) = e_{ij}(0) - \mu_{ij} \int_0^t e_{ij}(s)ds + Q_{ij} \sum_{k=1}^r \mu_{ki} \int_0^t f_{ki}(s)ds, \qquad 1 \le i \ne j \le r, \qquad (20)$$

$$e_{ii}(t) = e_{ii}(0) - \lambda_i\big(t - u_i(t)\big)$$
$$+ \sum_{\substack{j=1 \\ j \ne i}}^r \mu_{ji} \int_0^t e_{ji}(s)ds + Q_{ii} \sum_{j=1}^r \mu_{ji} \int_0^t f_{ji}(s)ds, \qquad 1 \le i \le r, \qquad (21)$$

$$u(t) \text{ is non-decreasing with } u(0) = 0, \text{ and } \int_0^\infty e_{ii}(s)du_i(s) = 0 \qquad \text{for all } 1 \le i \le r. \qquad (22)$$

Furthermore, for all $T \ge 0$,

$$\lim_{N \to \infty} \|\big(\bar{E}^{(N)}, \bar{F}^{(N)}, I^{(N)}\big) - (e, f, u)\|_T = 0 \qquad (23)$$

almost surely.

REMARK 6. Theorem 1 assumes that $P_{ij} > 0$ for all i, j and $Q_{ii} > 0$ for all i. This assumption is not used in the proof of Theorem 3, but appears in Section 4.1.

We refer to equations (19)-(22) as the fluid model of the ridesharing network, and to $\big(e, f, u\big)$ as the fluid analog of $\big(\bar{E}^{(N)}, \bar{F}^{(N)}, I^{(N)}\big)$. Note that $P_{ij} = 0$ implies $f_{ij}(t) \equiv 0$, and $Q_{ij} = 0$ for $i \ne j$ implies $e_{ij}(t) \equiv 0$. It will come in handy later on to know that, $e(t)$, $f(t)$, and $u(t)$ are Lipschitz continuous. To see why, observe that for any $\epsilon > 0$, (23) says that we can choose N large enough such that

$$|u(t) - u(s)| \le |u(t) - I^{(N)}(t)| + |I^{(N)}(t) - I^{(N)}(s)| + |I^{(N)}(s) - u(s)|$$
$$\le \epsilon + |t - s| + \epsilon,$$

where in the second inequality we used the definition of $I^{(N)}$ to get

$$|I^{(N)}(t) - I^{(N)}(s)| \le |t - s|, \quad 0 \le s, t < \infty.$$

Hence,

$$|u(t) - u(s)| \le |t - s|, \quad 0 \le s, t < \infty. \qquad (24)$$

Combining (19)–(21) with (24) and the fact that $\big(e(t), f(t)\big)$ is bounded (because it is in $\mathcal{T}$), we deduce that both $e(t)$ and $f(t)$ are also Lipschitz-continuous. The next section considers the equilibrium behavior of the fluid model.

30

Braverman, Dai, Liu and Ying: *Empty-car Routing in Ridesharing Systems*
Article submitted to *Operations Research*; manuscript no. (Please, provide the manuscript number!)

4.1. Equilibrium Points

We begin with an informal discussion on how to characterize the equilibrium of the fluid model. In equilibrium, we expect $\big(\dot{e}(t), \dot{f}(t)\big)$ to equal zero. Taking derivatives in (19)–(21) and setting the left hand sides to zero gives us

$$0 = \lambda_i P_{ij}\big(1 - \dot{u}_i(t)\big) - \mu_{ij} f_{ij}(t), \qquad\qquad 1 \le i,j \le r, \qquad (25)$$

$$0 = -\mu_{ij} e_{ij}(t) + Q_{ij} \sum_{k=1}^{r} \mu_{ki} f_{ki}(t), \qquad\qquad 1 \le i \neq j \le r, \qquad (26)$$

$$0 = -\lambda_i\big(1 - \dot{u}_i(t)\big) + \sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} e_{ji}(t) + Q_{ii} \sum_{j=1}^{r} \mu_{ji} f_{ji}(t), \qquad\qquad 1 \le i \le r. \qquad (27)$$

The derivative $\dot{u}(t)$ above exists for almost every t because $u(t)$ is Lipschitz continuous by (24). Next, by adding up all the terms in (19)–(21) we can see that the total amount of fluid in the system always equals one, or

$$\big(e(t), f(t)\big) \in \mathcal{T}. \qquad (28)$$

Lastly, we combine Lipschitz continuity of $u(t)$ together with (22) to see that

$$\int_0^\infty e_{ii}(s) du_i(s) = \int_0^\infty e_{ii}(s) \dot{u}_i(s) ds = 0.$$

Since $e_{ii}(t) \ge 0$ and $\dot{u}_i(t) \ge 0$ (because $u(t)$ is an increasing function), for all $t \ge 0$ where $\dot{u}_i(t)$ exists,

$$e_{ii}(t)\dot{u}_i(t) = 0. \qquad (29)$$

Therefore, we expect the equilibrium of the fluid model (if it exists) to satisfy (25)–(29). The following lemma addresses the issue of existence of an equilibrium. In the lemma, we use $\bar{a}_i \in [0,1]$ as a placeholder for $1 - \dot{u}_i(t)$ to represent the equilibrium server utilization at the station corresponding to e_{ii}. We let $\bar{a}$ be the r-dimensional vector whose components are $\bar{a}_i$. We can now see that (25)–(29) are exactly the constraints in the fluid-based optimization problem (4b)–(4g).

LEMMA 3. *Assume $P_{ij} > 0$ and $Q_{ii} > 0$ for all $i, j = 1, \ldots, r$. The system of equations*

$$\lambda_i P_{ij} \bar{a}_i = \mu_{ij} \bar{f}_{ij}, \qquad\qquad 1 \le i, j, \le r, \tag{30a}$$

$$\mu_{ij} \bar{e}_{ij} = Q_{ij} \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}, \qquad\qquad 1 \le i \le r, \ j \ne i, \tag{30b}$$

$$\lambda_i \bar{a}_i = \sum_{\substack{k=1 \\ k \ne i}}^{r} \mu_{ki} \bar{e}_{ki} + Q_{ii} \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki}, \qquad\qquad 1 \le i \le r, \tag{30c}$$

$$(1 - \bar{a}_i) \bar{e}_{ii} = 0, \quad 1 \le i \le r, \tag{30d}$$

$$(\bar{e}, \bar{f}) \in \mathcal{T}, \quad \bar{a} \in [0,1]^r \tag{30e}$$

has at least one solution $(\bar{e}, \bar{f}, \bar{a})$. Multiple solutions may exist, but the only difference between solutions are in the components $\bar{e}_{ii}$ for i such that $\bar{a}_i = 1$. In other words, the components $\bar{a}, \bar{f}, \bar{e}_{ij}$ for $i \ne j$, and $\bar{e}_{ii}$ for i such that $\bar{a}_i < 1$, are identical across all solutions.

Define

$$\mathcal{E} = \left\{ (\bar{e}, \bar{f}) \in \mathcal{T} : \exists \ \bar{a} \in [0,1]^r \text{ such that } (\bar{e}, \bar{f}, \bar{a}) \text{ solves } (30\text{a})-(30\text{d}) \right\}. \tag{31}$$

Lemma 3 implies that we can associate a unique $\bar{a}$ to the set $\mathcal{E}$. Furthermore, the lemma implies that the quantity $\bar{m}$, defined as

$$\bar{m} = \sum_{i : \bar{a}_i = 1} \bar{e}_{ii} = 1 - \sum_{i=1}^{r} \sum_{j=1}^{r} \bar{f}_{ij} - \sum_{i=1}^{r} \sum_{\substack{j=1 \\ j \ne i}}^{r} \bar{e}_{ij} \tag{32}$$

is unique. In light of the following theorem, we refer to $\mathcal{E}$ as the equilibrium set. The theorem is proved in Appendix EC.2 and is the final ingredient to prove Theorem 1.

THEOREM 4. *Let $\big(e(t), f(t), u(t)\big)$ be the unique solution to (19)–(22) with initial condition $(e(0), f(0)) \in \mathcal{T}$. Assume $P_{ij} > 0$ and $Q_{ii} > 0$ for all $i, j = 1, \ldots, r$. Then for any $\epsilon > 0$, there exists a $T > 0$ such that*

$$\inf_{x \in \mathcal{E}} \big| \big(e(t), f(t)\big) - x \big| < \epsilon, \quad t \ge T.$$

REMARK 7. In Lemma 3 and Theorem 4, we make use of the assumption that $P_{ij} > 0$ for all i, j and $Q_{ii} > 0$ for all i. We expect that the result holds if the CTMC has a single recurrent class, but our extra assumption greatly facilitates exposition. This is especially true in the proof of Theorem 4, which is already rather cumbersome.

5. Conclusions

This paper considered empty-car routing in a ridesharing network under a regime where both supply and demand for cars tend to infinity, and provided a comprehensive analysis of the design of an optimal empty-car routing policy based on asymptotic fluid analysis. A numerical study using real-world ridesharing data confirmed the effectiveness of our solution. Our paper is only a first step to understand ridesharing networks, and poses some interesting problems for future research directions:

Decentralized routing: Our routing policy is a centralized solution that assumes the ridesharing platform has full control over its empty cars, which provides a best-case benchmark. One future research topic is the design of decentralized incentive mechanisms for achieving the routing probabilities of the centralized solution. One can use the centralized routing policy as a benchmark to quantify the efficiency of a decentralized mechanism.

Time-varying parameters: Our T-lookahead policy is just one heuristic for dealing with time varying parameters, but it would be very interesting to be able to say something rigorous about time-varying policies. Along this line, studying the transient control problem would also be of interest (given the long time it takes the fluid to converge to equilibrium from certain initial conditions). Even studying the fluid transient control problem is non-trivial, because the fluid model is a non-linear dynamical system.

Robust routing policies: We witnessed in Figure 3 that certain state-dependent policies can attain performance levels close to those of the optimal static policy, yet have the benefit of not requiring explicit knowledge of system parameters. Now that a benchmark has been established, can we find state-dependent policies that are provably asymptotically optimal, yet rely as little as possible on primitive system parameters?

Appendix

EC.1. Process Level Convergence for Closed Networks of Single and Infinite Server Stations

In this section we consider closed queueing networks with exponential service times that consist of both single-server and infinite-server stations. We prove that as both the number of jobs in the network, and the service rates at the single server stations increase, the appropriately scaled queue length process converges to a fluid limit. The ridesharing model in Section 2 is a special type of such networks, meaning that Theorem 3 will be a special case of the results here. We remark that process-level convergence for the class of networks considered here is a consequence of the results in Krichagina (1992). However, in that paper the limiting process is defined as the solution to a differential inclusion, and our proof technique is sufficiently different to merit a separate write-up.

We consider a closed queueing network with $N > 0$ jobs and J stations, consisting of both single-server and infinite-server stations. We let the stations be indexed by the set $\mathcal{J} = \{1, \ldots, J\}$. Let $\mathcal{I} \subset \mathcal{J}$ and $\mathcal{S} \subset \mathcal{J}$ be the *non-empty* index sets corresponding to the infinite server stations, and single server stations, respectively. To describe the network dynamics, we introduce the following primitives. Let $Q^{(N)}(0) \in \mathbb{Z}_+^J$ with $\sum_{i \in \mathcal{J}} Q_i^{(N)}(0) = N$ be the vector representing the initial job distribution in the network. To keep track of service completions at each stations, we let

$$S_i = \{S_i(t), t \geq 0\}, \quad i \in \mathcal{J}$$

be a collection of unit-rate Poisson processes with S_i independent of S_j for $i \neq j$. Let $\lambda, \mu \in \mathbb{R}_+^J$ be two vectors with $\lambda_i = 0$ for $i \in \mathcal{I}$ and $\mu_i = 0$ for $i \in \mathcal{S}$, which we will use to represent the service rates at different stations. We assume that the service rate of each server at station $i \in \mathcal{J}$ is

$$N\lambda_i, \quad i \in \mathcal{S}$$

$$\mu_i, \quad i \in \mathcal{I}.$$

Let

$$\left\{ \left(\Phi_{i1}(n), \dots, \Phi_{iJ}(n) \right) \in \mathbb{Z}_+^J, \ n \in \mathbb{Z}_+ \right\}, \quad i \in \mathcal{J}$$

be a collection of routing processes defined as follows. For each $n \in \mathbb{Z}_+$ and each $i \in \mathcal{J}$, the vector

$$\left(\Phi_{i1}(n), \dots, \Phi_{iJ}(n) \right) = \sum_{m=1}^n \phi_i(m),$$

where $\left\{ \phi_i(m) \in \{0,1\}^J \right\}_{m=1}^\infty$ is a sequence of i.i.d. random variables with

$$\mathbb{P}(\phi_i(1) = e^{(j)}) = R_{ij}, \quad i,j \in \mathcal{J}.$$

Furthermore, the sequences $\left\{ \phi_i(m) \right\}_{m=1}^\infty$ and $\left\{ \phi_j(m) \right\}_{m=1}^\infty$ are assumed to be independent for $i \neq j$. Let R be the routing probability matrix whose i,jth entry is R_{ij}, and observe that it is a column stochastic matrix.

Using an inductive argument similar to the proof of (Chen and Mandelbaum 1991, Theorem 2.1), one can show that there exists a unique process

$$Q^{(N)} = \{ Q^{(N)}(t) = (Q_1^{(N)}(t), \dots, Q_J^{(N)}(t)), t \geq 0 \}$$

satisfying

$$
\begin{aligned}
Q_i^{(N)}(t) = Q_i^{(N)}(0) - S_i\left(N\lambda_i T_i^{(N)}(t) \right) + \sum_{j \in \mathcal{S}} \Phi_{ji}\left(S_j\left(N\lambda_j T_j^{(N)}(t) \right) \right) \\
+ \sum_{k \in \mathcal{I}} \Phi_{ki}\left(S_k\left(\mu_k \int_0^t Q_k^{(N)}(s)ds \right) \right), \quad i \in \mathcal{S},
\end{aligned}
\tag{EC.1}
$$

$$
\begin{aligned}
Q_i^{(N)}(t) = Q_i^{(N)}(0) - S_i\left(\mu_i \int_0^t Q_i^{(N)}(s)ds \right) + \sum_{j \in \mathcal{S}} \Phi_{ji}\left(S_j\left(N\lambda_j T_j^{(N)}(t) \right) \right) \\
+ \sum_{k \in \mathcal{I}} \Phi_{ki}\left(S_k\left(\mu_k \int_0^t Q_k^{(N)}(s)ds \right) \right), \quad i \in \mathcal{I},
\end{aligned}
\tag{EC.2}
$$

where

$$T_i^{(N)} = \left\{ T_i^{(N)}(t) = \int_0^t 1(Q_i^{(N)}(s) > 0)ds \right\}, \quad i \in \mathcal{S},$$

is the cumulative busy time process of the server at each single server station. At any time $t \geq 0$, $Q_i^{(N)}(t)$ is the job count at station $i \in \mathcal{J}$. It is a straightforward exercise to verify that $Q^{(N)}$ satisfies

the Markov property and is therefore a CTMC. Furthermore, the CTMC $(E^{(N)}, F^{(N)})$ introduced in Section 2 is a special case of $Q^{(N)}$.

To write (EC.1)–(EC.2) in a form that is more convenient for analysis, for any $t \geq 0$ let us define

$$\widehat{S}_i(t) = S_i(t) - t, \quad i \in \mathcal{J},$$

$$\widehat{\Phi}_{ij}(n) = \Phi_{ij}(n) - R_{ij}n, \quad i, j \in \mathcal{J}, \ n \in \mathbb{Z}_+,$$

and

$$\widehat{M}_i^{(N)}(t) = -\widehat{S}_i\big(N\lambda_i T_i^{(N)}(t)\big) + \sum_{j \in \mathcal{S}}\Big[\widehat{\Phi}_{ji}\big(S_j\big(N\lambda_j T_j^{(N)}(t)\big)\big) + R_{ji}\widehat{S}_j\big(N\lambda_j T_j^{(N)}(t)\big)\Big]$$

$$+ \sum_{k \in \mathcal{I}}\Big[\widehat{\Phi}_{ki}\Big(S_k\Big(\mu_k \int_0^t Q_k^{(N)}(s)ds\Big)\Big) + R_{ki}\widehat{S}_k\Big(\mu_k \int_0^t Q_k^{(N)}(s)ds\Big)\Big], \quad i \in \mathcal{S}, \quad \text{(EC.3)}$$

$$\widehat{M}_i^{(N)}(t) = -\widehat{S}_i\Big(\mu_i \int_0^t Q_i^{(N)}(s)ds\Big) + \sum_{j \in \mathcal{S}}\Big[\widehat{\Phi}_{ji}\big(S_j\big(N\lambda_j T_j^{(N)}(t)\big)\big) + R_{ji}\widehat{S}_j\big(N\lambda_j T_j^{(N)}(t)\big)\Big]$$

$$+ \sum_{k \in \mathcal{I}}\Big[\widehat{\Phi}_{ki}\Big(S_k\Big(\mu_k \int_0^t Q_k^{(N)}(s)ds\Big)\Big) + R_{ki}\widehat{S}_k\Big(\mu_k \int_0^t Q_k^{(N)}(s)ds\Big)\Big], \quad i \in \mathcal{I}, \quad \text{(EC.4)}$$

and let $\widehat{M}^{(N)}(t)$ be the vector whose components are $\widehat{M}_i^{(N)}(t)$. For $t \geq 0$, we also define

$$I_i^{(N)}(t) = 0 \text{ for } i \in \mathcal{I} \quad \text{and} \quad I_i^{(N)}(t) = t - T_i^{(N)}(t) \text{ for } i \in \mathcal{S}, \quad \text{(EC.5)}$$

and let $I^{(N)} = \{I^{(N)}(t) \in \mathbb{R}_+^J, t \geq 0\}$. Then for $i \in \mathcal{S}$, $I_i^{(N)}(t)$ represents the cumulative idle time up to time t. Setting

$$\bar{Q}^{(N)}(t) = \frac{1}{N}Q^{(N)}(t) \quad \text{and} \quad \bar{M}^{(N)}(t) = \frac{1}{N}\widehat{M}^{(N)}(t),$$

we from (EC.1)–(EC.4) that

$$\bar{Q}_i^{(N)}(t) = \bar{Q}_i^{(N)}(0) + \bar{M}_i(t) + \Big(\sum_{j \in \mathcal{S}} R_{ji}\lambda_j - \lambda_i\Big)t + \sum_{k \in \mathcal{I}} R_{ki}\mu_k \int_0^t \bar{Q}_k^{(N)}(s)ds$$

$$+ \lambda_i I_i^{(N)}(t) - \sum_{j \in \mathcal{S}} R_{ji}\lambda_j I_j^{(N)}(t), \quad i \in \mathcal{S}, \quad \text{(EC.6)}$$

$$\bar{Q}_i^{(N)}(t) = \bar{Q}_i^{(N)}(0) + \bar{M}_i(t) + \Big(\sum_{j \in \mathcal{S}} R_{ji}\lambda_j\Big)t - \mu_i \int_0^t \bar{Q}_i^{(N)}(s)ds$$

$$+ \sum_{k \in \mathcal{I}} R_{ki}\mu_k \int_0^t \bar{Q}_k^{(N)}(s)ds - \sum_{j \in \mathcal{S}} R_{ji}\lambda_j I_j^{(N)}(t), \quad i \in \mathcal{I}. \quad \text{(EC.7)}$$

In the next section, we describe the fluid model to which the process $\bar{Q}^{(N)}$ will converge to as $N \to \infty$.

EC.1.1. The Fluid Model

Recalling that $\mu_i = 0$ for $i \in \mathcal{S}$ and $\lambda_i = 0$ for $i \in \mathcal{I}$, we set

$$M = \operatorname{diag}(\mu) \quad \text{and} \quad \Lambda = \operatorname{diag}(\lambda). \tag{EC.8}$$

We also define the $J \times J$ matrix $\tilde{R}$ by setting

$$\tilde{R}_{ij} = R_{ij}, \quad i \in \mathcal{S},$$

$$\tilde{R}_{ij} = 0, \quad i \in \mathcal{I}.$$

That is $\tilde{R}$ is the matrix R with all rows corresponding to infinite server stations being set to zero. Since $\mathcal{I} \neq \emptyset$, the matrix $\tilde{R}^T$ is sub-stochastic. The following lemma is proved in Section EC.1.2.

LEMMA EC.1. *For each $x \in \mathbb{D}_1^J$, there exists a unique $(q, v) \in \mathbb{D}^{2J}$, with $q(t) \in \mathbb{R}_+^J$ and $v(t) \in \mathbb{R}_+^J$ for all $t \geq 0$, such that*

$$q(t) = x(t) - (I - R^T)M \int_0^t q(s)ds + (I - \tilde{R}^T)v(t) \tag{EC.9}$$

$$q(t) \geq 0, \quad t \geq 0, \tag{EC.10}$$

$$v(\cdot) \text{ is non-decreasing with } v(0) = 0, \tag{EC.11}$$

$$\int_0^\infty q_i(s)dv_i(s) = 0, \quad i \in \mathcal{J}. \tag{EC.12}$$

Furthermore, the map $\Upsilon : \mathbb{D}_1^J \to \mathbb{D}^{2J}$ given by $\Upsilon(x) = (q, v)$ is well-defined and is Lipschitz-continuous, in the sense that for any $x, \tilde{x} \in \mathbb{D}_1^J$, and any $T > 0$, there exists a constant c_Υ^T such that

$$\|\Upsilon(x) - \Upsilon(\tilde{x})\|_T \leq c_\Upsilon^T \|x - \tilde{x}\|_T, \tag{EC.13}$$

THEOREM EC.1. *Assume $\tilde{Q}^{(N)}(0) \to a$ as $N \to \infty$ for some $a \in [0, 1]^J$ with $\sum_{i=1}^J a_i = 1$. Let $e \in \mathbb{R}^J$ be the vector of ones, and let $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ be the identity map defined by $\gamma(t) = t$. Set*

$$(q, v) = \Upsilon\Big(a + \big((R^T - I)\Lambda e\big)\gamma\Big). \tag{EC.14}$$

Then for any $T > 0$,

$$\lim_{n \to \infty} \|\tilde{Q}^{(N)}(t) - q(t)\|_T = 0,$$

$$\lim_{n \to \infty} \|I^{(N)}(t) - v(t)\|_T = 0.$$

Proof of Theorem EC.1. From (EC.6)–(EC.7) we see that

$$\left(\bar{Q}^{(N)}, I^{(N)}\right) = \Upsilon\left(\bar{Q}^{(N)}(0) + \bar{M}^{(N)} + \left((R^T - I)\Lambda e\right)\gamma\right),$$

where $\bar{M}^{(N)} = \{\bar{M}^{(N)}(t) \in \mathbb{R}^J, t \geq 0\}$. Suppose we knew that for every $T \geq 0$,

$$\lim_{N \to \infty} \|\bar{M}^{(N)}\|_T = 0 \quad \text{almost surely.} \tag{EC.15}$$

Then the continuous mapping theorem Billingsley (1999), together with (EC.13) would imply Theorem EC.1. The proof of (EC.15) involves a standard argument using the functional strong law of large numbers (FSLLN), and is therefore omitted. For an example of such an argument, see the proof of (5.6) in Dai et al. (2010).

REMARK EC.1. Since $I_i^{(N)}(t) = 0$ for $i \in \mathcal{I}$ and $t \geq 0$, Theorem EC.1 implies

$$v_i(t) = 0, \quad i \in \mathcal{I}. \tag{EC.16}$$

Establishing (EC.16) by relying on convergence of $\left(\bar{Q}^{(N)}, I^{(N)}\right)$ to (q, v) may seem strange, because (EC.16) should be a standalone property of $\Upsilon\left(a + \left((R^T - I)\Lambda e\right)\gamma\right)$. Indeed, it is possible to establish (EC.16) using a direct argument that relies on Proposition 1 of Reiman (1984). However, we avoid using said argument because it is significantly longer.

We immediately see that Theorem 3 is a special case of Theorem EC.1. The rest of this section is devoted to proving Lemma EC.1.

EC.1.2. Proof of Lemma EC.1

In order to prove Lemma EC.1, we first need to introduce the Skorohod problem. Let $\tilde{Q}$ be a $J \times J$ column sub-stochastic matrix with non-negative entries. For any $x \in \mathbb{D}_{0+}^J$, let $(z, y) \in \mathbb{D}^{2J}$ with $z(t) \in \mathbb{R}_+^J$ and $y(t) \in \mathbb{R}_+^J$ for all $t \geq 0$ be the solution to

$$z = x + (I - \tilde{Q}^T)y, \tag{EC.17}$$

$$z \geq 0, \tag{EC.18}$$

$$y(\cdot) \text{ is non-decreasing and } y(0) = 0, \tag{EC.19}$$

$$\int_0^\infty z_i(s)dy_i(s) = 0, \quad 1 \leq i \leq J. \tag{EC.20}$$

Existence and uniqueness of (z, y) was proved in Harrison and Reiman (1981) when x is continuous, but the arguments there hold for $x \in \mathbb{D}_{0+}^J$ as well. We refer to (EC.17)–(EC.20) as the Skorohod problem associated with $(x, \tilde{Q})$, and write $\mathrm{SP}(x, \tilde{Q})$ for short. We refer to $(z, y) \in \mathbb{D}^{2J}$ as the solution to $\mathrm{SP}(x, \tilde{Q})$ if it satisfies (EC.17)–(EC.20). Furthermore, for any $x \in \mathbb{D}_{0+}^J$ we define the Skorohod map $\Psi^{\tilde{Q}} : \mathbb{D}_{0+}^J \to \mathbb{D}^{2J}$ by

$$\Psi^{\tilde{Q}}(x) = \left(\Psi_z^{\tilde{Q}}(x), \Psi_y^{\tilde{Q}}(x)\right) = (z, y),$$

where (z, y) is the solution of $\mathrm{SP}(x, \tilde{Q})$. From (EC.17) it is clear that

$$\Psi_z^{\tilde{Q}}(x) = x + (I - \tilde{Q}^T)\Psi_y^{\tilde{Q}}(x), \quad x \in \mathbb{D}^J, \ x(0) \geq 0. \tag{EC.21}$$

Both $\Psi_z^{\tilde{Q}}$ and $\Psi_y^{\tilde{Q}}$ are Lipschitz-continuous, in the sense that for any $T > 0$, there exists constants $c_z^T, c_y^T > 0$, which depend on $\tilde{Q}$, such that for any $x, \tilde{x} \in \mathbb{D}_{0+}^J$,

$$\|\Psi_y^{\tilde{Q}}(x) - \Psi_y^{\tilde{Q}}(\tilde{x})\|_T \leq c_y^T \|x - \tilde{x}\|_T,$$

$$\|\Psi_z^{\tilde{Q}}(x) - \Psi_z^{\tilde{Q}}(\tilde{x})\|_T \leq c_z^T \|x - \tilde{x}\|_T,$$

where $\|\cdot\|_T$ is defined in (2). This was established in (Harrison and Reiman 1981, p. 305) when both x and $\tilde{x}$ are continuous, but the argument used there holds for $x, \tilde{x} \in \mathbb{D}_{0+}^J$ as well. We are now ready to prove Lemma EC.1.

Proof of Lemma EC.1. Fix $x \in \mathbb{D}_1^J$ and omitting the superscript $\tilde{R}$, let (Ψ_z, Ψ_y) be the Skorohod map associated with $\text{SP}(x, \tilde{R})$. Define the integral operator $\alpha : \mathbb{D}_1^J \times \mathbb{D}^J \to \mathbb{D}^J$ by

$$\alpha(x, w)(t) = x(t) - (I - R^T)M \int_0^t w(s)ds, \tag{EC.22}$$

and consider the integral equation

$$q = \Psi_z(\alpha(x, q)) = \alpha(x, q) + (I - \tilde{R}^T)\Psi_y(\alpha(x, q)) \tag{EC.23}$$

$$q \in \mathbb{D}^J, \tag{EC.24}$$

where the second equality in (EC.23) follows from (EC.21). Provided (EC.23)–(EC.24) has a unique solution q, we can set $v = \Psi_y(\alpha(x, q))$ and observe that by definition of α and (Ψ_z, Ψ_y),

$$(q, v) = \Upsilon(x).$$

Hence, we now establish the existence and uniqueness of a solution q to (EC.23)–(EC.24). Construct a sequence $\{q^n \in \mathbb{D}^J\}_{n=0}^\infty$ by letting

$$q^0(t) \equiv x(0), \quad t \geq 0,$$

$$q^{n+1}(t) = \Psi_z(\alpha(x, q^n))(t), \quad t \geq 0.$$

Observe that $q^n(0) = x(0)$ for all $n \geq 0$. We first show that for any $T > 0$, this sequence is a Cauchy sequence in the Hilbert space $(\mathbb{D}^J[0, T], \|\cdot\|_T)$. Let $\bar{\mu} = \max_{i \in \mathcal{I}}\{\mu_j\}$ (remembering that $\mu_i = 0$ for $i \in \mathcal{S}$), and observe that

$$\|q^{n+1} - q^n\|_T \leq c_z^T \|\alpha(x, q^n) - \alpha(x, q^{n-1})\|_T$$

$$= c_z^T \max_{i \in \mathcal{J}} \sup_{0 \leq t \leq T} \left|\left(\alpha(x, q^n)\right)_i(t) - \left(\alpha(x, q^{n-1})\right)_i(t)\right|$$

$$\leq c_z^T \bar{\mu} J \max_{i \in \mathcal{I}} \left\{ \int_0^T \left|q_i^n(s) - q_i^{n-1}(s)\right| ds \right\}$$

$$\leq c_z^T \bar{\mu} J \int_0^T \max_{i \in \mathcal{I}} \left|q_i^n(s) - q_i^{n-1}(s)\right| ds$$

$$\leq \frac{(c_z^T \bar{\mu} J T)^n}{n!} \|q^1 - q^0\|_T. \tag{EC.25}$$

The first inequality follows from the Lipschitz property of Ψ_z, the second inequality is from the form of α, and the last inequality follows by recursion. From this point it is not hard to conclude (see for instance (11.22) of Mandelbaum et al. (1998)) that $\{q^n\}_{n=0}^\infty$ is a Cauchy sequence in $\left(|D^J[0,T], \|\cdot\|_T\right)$ for each $T > 0$. Therefore, q^n converges to some limit $q \in \mathbb{D}^J[0,T]$ that satisfies (EC.23). Since the choice of $T > 0$ was arbitrary, we have proved existence of a solution to (EC.23)–(EC.24). Uniqueness can be argued by taking two potential solutions q and $\tilde{q}$, and applying the chain of arguments in (EC.25) with $\|q - \tilde{q}\|_T$ on the left hand side there.

It remains to prove the Lipschitz-continuity of Υ. Fix any $x, \tilde{x} \in \mathbb{D}_1^J$, and set

$$(q, v) = \Upsilon(x) \quad \text{and} \quad (\tilde{q}, \tilde{v}) = \Upsilon(\tilde{x}).$$

Repeating the logic used to obtain (EC.25), we see that for any $n \geq 1$,

$$
\begin{aligned}
\|q - \tilde{q}\|_T &= \|\Psi_z\big(\alpha(x, q)\big) - \Psi_z\big(\alpha(\tilde{x}, \tilde{q})\big)\|_T \\
&\leq c_z^T \|x - \tilde{x}\|_T + c_z^T \bar{\mu} J \int_0^T \max_{i \in \mathcal{I}} |q_i(s) - \tilde{q}_i(s)| \, ds \\
&\leq c_z^T \|x - \tilde{x}\|_T \sum_{k=0}^{n-1} \frac{\left(c_z^T \bar{\mu} J T\right)^k}{k!} + \frac{\left(c_z^T \bar{\mu} J T\right)^n}{n!} \|q - \tilde{q}\|_T,
\end{aligned}
$$

where the last inequality follows by recursion. Choosing n large enough so that $\frac{\left(c_z^T \bar{\mu} J T\right)^n}{n!} < 1$, we conclude the existence of a constant $c_{\Upsilon,q}^T > 0$ such that

$$\|q - \tilde{q}\|_T \leq c_{\Upsilon,q}^T \|x - \tilde{x}\|_T. \tag{EC.26}$$

Similarly, we see that

$$
\begin{aligned}
\|v - \tilde{v}\|_T &= \|\Psi_y\big(\alpha(x, q)\big) - \Psi_y\big(\alpha(\tilde{x}, \tilde{q})\big)\|_T \\
&\leq c_y^T \|x - \tilde{x}\|_T + c_y^T \bar{\mu} J \int_0^T \max_{i \in \mathcal{I}} |q_i(s) - \tilde{q}_i(s)| \, ds \\
&\leq c_y^T \|x - \tilde{x}\|_T + c_y^T \bar{\mu} J T \|q - \tilde{q}\|_T,
\end{aligned}
$$

and by (EC.26), there exists a constant $c_{\Upsilon,v}^T > 0$ satisfying

$$\|v - \tilde{v}\|_T \leq c_{\Upsilon,v}^T \|x - \tilde{x}\|_T.$$

This establishes (EC.13) and concludes the proof the lemma.

EC.2. Proof of Theorem 4

This section is devoted to proving Theorem 4. For the remainder of this section, we fix an initial condition $\big(e(0), f(0)\big) \in \mathcal{T}$ and let $\big(e(t), f(t), u(t)\big)$ be the unique solution to the fluid model with this initial condition. Furthermore, we fix $(\bar{e}, \bar{f}) \in \mathcal{E}$ and let $\bar{a}$ and $\bar{m}$ be defined by (30a) and (32), respectively. For $(x, y) \in \mathcal{T}$, define the function $V : \mathcal{T} \to \mathbb{R}_+$ by

$$V(x,y) = \sum_{i=1}^{r}\sum_{j=1}^{r}\big|y_{ij} - \bar{f}_{ij}\big| + \sum_{i=1}^{r}\sum_{\substack{j=1 \\ i \neq j}}^{r}\big|x_{ij} - \bar{e}_{ij}\big| + \sum_{i:\bar{a}_i < 1} x_{ii} + \Big|\bar{m} - \sum_{i:\bar{a}_i = 1} x_{ii}\Big|. \tag{EC.27}$$

To prove Theorem 4, we will show that $V\big(e(t), f(t)\big)$ is a Lyapunov function. We know that $V\big(e(t), f(t)\big)$ is a Lipschitz-continuous function from $\mathbb{R}_+ \to \mathbb{R}_+$ because $V(\cdot), e(\cdot)$, and $f(\cdot)$ are all Lipschitz-continuous.

We say $t > 0$ is a *regular point* of $V\big(e(t), f(t)\big)$ if for all $i, j = 1, \ldots, r$, the functions $e_{ii}(t)$, $\big|f_{ij}(t) - \bar{f}_{ij}\big|$, $\big|e_{ij}(t) - \bar{e}_{ij}\big|$ for $i \neq j$, and $\big|\bar{m} - \sum_{i:\bar{a}_i = 1}^{r} e_{ii}(t)\big|$ are differentiable at t. Since these functions are all Lipschitz-continuous, then almost every point is a regular point. Furthermore, if t is a regular point, we claim that

$$f_{ij}(t) = \bar{f}_{ij} \quad \Rightarrow \quad \dot{f}_{ij}(t) = 0, \quad 1 \leq i, j \leq r \tag{EC.28}$$

$$e_{ii}(t) = 0 \quad \Rightarrow \quad \dot{e}_{ii}(t) = 0, \quad 1 \leq i \leq r, \tag{EC.29}$$

$$e_{ij}(t) = \bar{e}_{ij} \quad \Rightarrow \quad \dot{e}_{ij}(t) = 0, \quad 1 \leq i \neq j \leq r \tag{EC.30}$$

$$\sum_{i:\bar{a}_i = 1}^{r} e_{ii} = \bar{m} \quad \Rightarrow \quad \sum_{i:\bar{a}_i = 1}^{r} \dot{e}_{ii} = 0, \tag{EC.31}$$

$$\dot{u}_i(t) \text{ exists for all } 1 \leq i \leq r. \tag{EC.32}$$

Most of these properties are follow directly from the definition of a regular point. For instance, the existence of $\frac{d\big|f_{ij}(t) - \bar{f}_{ij}\big|}{dt}$ together with $f_{ij}(t) = \bar{f}_{ij}$ implies $\dot{f}_{ij}(t) = 0$; the same argument applies to (EC.29)–(EC.31). We know (EC.32) is true because $\dot{u}_i(t)$ must exist in order for $\dot{e}_{ii}(t)$ to exist; cf. (21).

LEMMA EC.2. *Assume* $P_{ij} > 0$ *and* $Q_{ii} > 0$ *for all* $i, j = 1, \ldots, r$. *If* $t > 0$ *is a regular point of* $V\big(e(t), f(t)\big)$, *then* $\dot{V}\big(e(t), f(t)\big) \leq 0$. *Furthermore, if* $\big(e(t), f(t)\big) \notin \mathcal{E}$, *then* $\dot{V}\big(e(t), f(t)\big) < 0$.

We postpone the proof of Lemma EC.2 to the end of this section, and first demonstrate how it is used to prove Theorem 4 via LaSalle's Invariance Principle (Khalil 2002, Theorem 4.4). To do so, we need to introduce a few definitions. A point $p \in \mathcal{T}$ is said to be a positive limit point of $\big(e(t), f(t)\big)$ if there exists a sequence $\{t_n\}_{n=1}^{\infty}$ with $t_n \to \infty$ such that $\big(e(t_n), f(t_n)\big) \to p$ as $n \to \infty$. A set $B \subset \mathcal{T}$ is said to be positively invariant if for any $s \geq 0$,

$$\big(e(s), f(s)\big) \in B \quad \Rightarrow \quad \big(e(t), f(t)\big) \in B, \quad \forall\, t \geq s.$$

The following lemma is a version of (Khalil 2002, Lemma 4.1) adapted to the setting of this paper. We prove it in Section EC.2.4.

LEMMA EC.3. *Let L^+ be the set of all positive limit points of $\big(e(t), f(t)\big)$. Then L^+ is a nonempty, compact, and positively invariant set. Moreover,*

$$\lim_{t \to \infty} \inf_{x \in L^+} \big\| \big(e(t), f(t)\big) - x \big\| = 0$$

We are now ready to prove Theorem 4. The following argument is repeated from the proof of (Khalil 2002, Theorem 4.4).

Proof of Theorem 4. We know that $V\big(e(t), f(t)\big)$ is a bounded function from $\mathbb{R}_+ \to \mathbb{R}_+$, because $V(\cdot)$ is continuous and $\big(e(t), f(t)\big)$ belong to the compact set $\mathcal{T}$ for all $t \geq 0$. Furthermore, Lemma EC.2 implies that it is a non-increasing function. Therefore, $\lim_{t \to \infty} V\big(e(t), f(t)\big)$ exists, and we denote it by ℓ. Recall by Lemma EC.3 that L^+, the set of positive limit points of $\big(e(t), f(t)\big)$, is not empty. For any point $p \in L^+$, there exists a sequence $\{t_n\}_{n=1}^{\infty}$ such that $\big(e(t_n), f(t_n)\big) \to p$. By continuity of $V(\cdot)$,

$$V(p) = \lim_{n \to \infty} V\big(e(t_n), f(t_n)\big) = \ell, \quad \forall p \in L^+.$$

Now suppose $\big(\tilde{e}(t), \tilde{f}(t)\big)$ is a solution to the fluid model with initial condition $\big(\tilde{e}(0), \tilde{f}(0)\big) \in L^+$. Since L^+ is positively invariant, $\tilde{e}(t), \tilde{f}(t) \in L^+$ for all $t \geq 0$, which implies that $V\big(\tilde{e}(t), \tilde{f}(t)\big) = \ell$, or

$$\dot{V}\big(\tilde{e}(t), \tilde{f}(t)\big) = 0.$$

Using Lemma EC.2, we conclude that $L^+ \subset \mathcal{E}$, which proves Theorem 4.

The rest of this section is devoted to proving Lemma EC.2. Fix a regular point $t > 0$. For notational simplicity, we omit the time index t when referring to $V\big(e(t), f(t)\big)$, $e_{ij}(t)$, $f_{ij}(t)$, $u_i(t)$, or their derivatives. From (19)–(21) we can see that

$$\sum_{i=1}^{r}\sum_{j=1}^{r}\dot{e}_{ij} + \sum_{i=1}^{r}\sum_{j=1}^{r}\dot{f}_{ij} = 0. \tag{EC.33}$$

Set

$$\hat{f}_{ij} = f_{ij} - \bar{f}_{ij}, \quad 1 \le i, j \le r,$$

$$\hat{e}_{ij} = e_{ij} - \bar{e}_{ij}, \quad 1 \le i \ne j \le r.$$

Recall that t is a regular point, meaning (EC.28)–(EC.32) hold. Therefore,

$$\begin{aligned}
\frac{1}{2}\dot{V}(e,f) &= \frac{1}{2}\sum_{i:\bar{a}_i<1}^{r}\dot{e}_{ii} + \frac{1}{2}\sum_{i=1}^{r}\sum_{\substack{j=1\\j\ne i}}^{r}\dot{e}_{ij}\mathbf{1}\big(e_{ij}>\bar{e}_{ij}\big) - \frac{1}{2}\sum_{i=1}^{r}\sum_{\substack{j=1\\j\ne i}}^{r}\dot{e}_{ij}\mathbf{1}\big(e_{ij}<\bar{e}_{ij}\big)\\
&\quad + \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r}\dot{f}_{ij}\mathbf{1}\big(f_{ij}>\bar{f}_{ij}\big) - \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r}\dot{f}_{ij}\mathbf{1}\big(f_{ij}<\bar{f}_{ij}\big)\\
&\quad + \mathbf{1}\Big(\sum_{i:\bar{a}_i=1}e_{ii}>\bar{m}\Big)\frac{1}{2}\sum_{i:\bar{a}_i=1}\dot{e}_{ii} - \mathbf{1}\Big(\sum_{i:\bar{a}_i=1}e_{ii}<\bar{m}\Big)\frac{1}{2}\sum_{i:\bar{a}_i=1}\dot{e}_{ii}\\
&= \sum_{i:\bar{a}_i<1}^{r}\dot{e}_{ii} + \sum_{i=1}^{r}\sum_{\substack{j=1\\j\ne i}}^{r}\dot{e}_{ij}\mathbf{1}\big(e_{ij}>\bar{e}_{ij}\big) + \sum_{i=1}^{r}\sum_{j=1}^{r}\dot{f}_{ij}\mathbf{1}\big(f_{ij}>\bar{f}_{ij}\big)\\
&\quad + \mathbf{1}\Big(\sum_{i:\bar{a}_i=1}e_{ii}>\bar{m}\Big)\sum_{i:\bar{a}_i=1}\dot{e}_{ii}, \tag{EC.34}
\end{aligned}$$

where the second equality follows from (EC.33). The expression in (EC.34) will soon become very bulky to work with, so before moving forward we first present an illustrative example to help the reader gain some intuition.

EC.2.1. An Illustrative Example

Suppose $r = 3$ and that (λ, μ, P, Q) are such that $\bar{a}_1 < 1$, $\bar{a}_2 = \bar{a}_3 = 1$, and $\bar{m} > 0$. For simplicity, we also assume that all entries of P and Q are strictly positive. We now compute $\frac{1}{2}\dot{V}(e,f)$ for several choices of $(e,f) \notin \mathcal{E}$, and show that it is always strictly negative. The list of cases we consider is by no means exhaustive, but is nevertheless helpful to develop intuition.

Case 1: Suppose $e_{ii} > 0$ for $i = 1, 2, 3$ and $e_{22} + e_{33} > \bar{m}$. Furthermore, suppose $\hat{f}_{ij} < 0$ for all i, j, and $\hat{e}_{ij} < 0$ for all $i \neq j$. Using (EC.34) and (27), we see that

$$\frac{1}{2}\dot{V}(e, f) = \dot{e}_{11} + \dot{e}_{22} + \dot{e}_{33} = \sum_{i=1}^{3}\left[-\lambda_i(1 - \dot{u}_i) + \sum_{\substack{j=1 \\ j \neq i}}^{3}\mu_{ji}e_{ji} + Q_{ii}\sum_{j=1}^{3}\mu_{ji}f_{ji}\right].$$

By (29), $\dot{u}_i = 0$ for $i = 1, 2, 3$ because $e_{ii} > 0$. Furthermore, using (30c) we see that

$$\sum_{\substack{j=1 \\ j \neq i}}^{3}\mu_{ji}e_{ji} + Q_{ii}\sum_{j=1}^{3}\mu_{ji}f_{ji} = \sum_{\substack{j=1 \\ j \neq i}}^{3}\mu_{ji}\bar{e}_{ji} + Q_{ii}\sum_{j=1}^{3}\mu_{ji}\bar{f}_{ji} + \sum_{\substack{j=1 \\ j \neq i}}^{3}\mu_{ji}\hat{e}_{ji} + Q_{ii}\sum_{j=1}^{3}\mu_{ji}\hat{f}_{ji}$$

$$= \lambda_i\bar{a}_i + \sum_{\substack{j=1 \\ j \neq i}}^{3}\mu_{ji}\hat{e}_{ji} + Q_{ii}\sum_{j=1}^{3}\mu_{ji}\hat{f}_{ji}, \quad 1 \leq i \leq 3.$$

Recalling that $\bar{a}_2 = \bar{a}_3 = 1$, we arrive at

$$\frac{1}{2}\dot{V}(e, f) = \lambda_1(1 - \bar{a}_1) + \sum_{i=1}^{3}\left[\sum_{\substack{j=1 \\ j \neq i}}^{3}\mu_{ji}\hat{e}_{ji} + Q_{ii}\sum_{j=1}^{3}\mu_{ji}\hat{f}_{ji}\right] < 0.$$

Case 2: Suppose $e_{11} = e_{22} = 0$, $e_{33} > \bar{m}$, $\hat{f}_{ij} = 0$ for all i, j, $\hat{e}_{12} < 0$ and $\hat{e}_{ij} = 0$ for all other i, j with $i \neq j$. In such a case, (EC.34) and (27) tell us that

$$\frac{1}{2}\dot{V}(e, f) = \dot{e}_{33} = -\lambda_3(1 - \dot{u}_3) + \sum_{\substack{j=1 \\ j \neq 3}}^{3}\mu_{j3}e_{j3} + Q_{33}\sum_{j=1}^{3}\mu_{j3}f_{j3}.$$

Now $e_{33} > 0$ and (29) implies that $\dot{u}_3 = 0$. Furthermore, we know $\hat{e}_{j3} = 0$ for $j = 1, 2$ and $\hat{f}_{j3} = 0$ for $j = 1, 2, 3$. Therefore, we can use (30c) to see that

$$\frac{1}{2}\dot{V}(e, f) = -\lambda_3 + \sum_{\substack{j=1 \\ j \neq 3}}^{3}\mu_{j3}\bar{e}_{j3} + Q_{33}\sum_{j=1}^{3}\mu_{j3}\bar{f}_{j3} = 0,$$

which appears to contradict what we set out to prove. However, it turns out that the time t corresponding to this configuration of (e, f) is not a regular point. If t were a regular point, then by (EC.28) we would have $\dot{\hat{f}}_{2k} = 0$ for all $k = 1, 2, 3$, because $\hat{f}_{2k} = 0$. However, by (25), the fact that $\hat{f}_{2k} = 0$, and (30a) we see that

$$\dot{f}_{2k} = \lambda_2 P_{2k}(1 - \dot{u}_2) - \mu_{2k}f_{2k} = \lambda_2(1 - \dot{u}_2) - \lambda_2 P_{2k}\bar{a}_2.$$

Since $e_{22} = 0$, (EC.29) forces $\dot{e}_{22} = 0$, which together with (27) implies that

$$\lambda_2 P_{2k}(1 - \dot{u}_2) = \sum_{\substack{j=1 \\ j \neq 2}}^{3} \mu_{j2} e_{j2} + Q_{22} \sum_{j=1}^{3} \mu_{j2} f_{j2} < \sum_{\substack{j=1 \\ j \neq 2}}^{3} \mu_{j2} \bar{e}_{j2} + Q_{22} \sum_{j=1}^{3} \mu_{j2} \bar{f}_{j2} = \lambda_2 P_{2k} \bar{a}_2,$$

where in the inequality we used that $\hat{f}_{j2} = \hat{e}_{32} = 0$, and $\hat{e}_{12} < 0$, and in the last equality we used

(30c). Therefore, we just showed that $\dot{f}_{2k} < 0$, which is a contradiction to the assumption that t is

a regular point.

Having worked through the example, we now present a general algebraic expansion of $\frac{1}{2} \dot{V}(e, f)$ in

Lemma EC.4, which is proved in Section EC.2.3. A line by line inspection of (EC.35) and (EC.36)

confirms that $\dot{V}(e, f) \leq 0$, but proving $\dot{V}(e, f) < 0$ requires more careful arguments presented in

Section EC.2.2.

LEMMA EC.4. *If* $\{i : \bar{a}_i = 1\} \neq \emptyset$ *and* $\sum_{i : \bar{a}_i = 1}^{r} e_{ii} \leq \bar{m}$, *then*

$$
\begin{aligned}
\frac{1}{2} \dot{V}(e, f) =\ & \sum_{i : \bar{a}_i < 1}^{r} \lambda_i (\bar{a}_i - 1) \Big(1 - \sum_{j=1}^{r} P_{ij} 1(\hat{f}_{ij} > 0) \Big) 1(\dot{u}_i = 0) \\
& - \sum_{i=1}^{r} \Big(1 - \sum_{j=1}^{r} P_{ij} 1(\hat{f}_{ij} > 0) \Big) \Big(\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \geq 0) + Q_{ii} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0) \Big) 1(\dot{u}_i > 0) \\
& + \sum_{i=1}^{r} \Big(\sum_{j=1}^{r} P_{ij} 1(\hat{f}_{ij} > 0) \Big) \Big(\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \leq 0) + Q_{ii} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \leq 0) \Big) 1(\dot{u}_i > 0) \\
& + \sum_{i : \bar{a}_i < 1}^{r} 1(\dot{u}_i = 0) \Big[\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \leq 0) + Q_{ii} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \leq 0) \Big] \\
& + \sum_{i=1}^{r} \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1(\hat{e}_{ij} > 0) \sum_{k=1}^{r} \mu_{ki} \hat{f}_{ki} 1(\hat{f}_{ki} \leq 0) \\
& - \sum_{i : \bar{a}_i < 1}^{r} \Big(1 - Q_{ii} - \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1(\hat{e}_{ij} > 0) \Big) \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0) \\
& - \sum_{i : \bar{a}_i = 1}^{r} 1(\dot{u}_i = 0) \sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \geq 0) \\
& - \sum_{i : \bar{a}_i = 1}^{r} \Big(1 - Q_{ii} 1(\dot{u}_i > 0) - \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1(\hat{e}_{ij} > 0) \Big) \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0).
\end{aligned}
\tag{EC.35}
$$

and if $\{i: \ \bar{a}_i = 1\} = \emptyset$, or $\{i: \ \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} > \bar{m}$, then

$$
\begin{aligned}
\frac{1}{2}\dot{V}(e,f) = & \sum_{i:\bar{a}_i<1}^{r} \lambda_i(\bar{a}_i - 1)\Big(1 - \sum_{j=1}^{r} P_{ij}1(\hat{f}_{ij} > 0)\Big)1(\dot{u}_i = 0) \\
& - \sum_{i=1}^{r}\Big(1 - \sum_{j=1}^{r} P_{ij}1(\hat{f}_{ij} > 0)\Big)\Big(\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji} \geq 0) + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji} \geq 0)\Big)1(\dot{u}_i > 0) \\
& + \sum_{i=1}^{r}\Big(\sum_{j=1}^{r} P_{ij}1(\hat{f}_{ij} > 0)\Big)\Big(\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji} \leq 0) + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji} \leq 0)\Big)1(\dot{u}_i > 0) \\
& + \sum_{i=1}^{r} 1(\dot{u}_i = 0)\Big[\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji} \leq 0) + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji} \leq 0)\Big] \\
& + \sum_{i=1}^{r}\sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij}1(\hat{e}_{ij} > 0)\sum_{k=1}^{r}\mu_{ki}\hat{f}_{ki}1(\hat{f}_{ki} \leq 0) \\
& - \sum_{i=1}^{r}\Big(1 - Q_{ii} - \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij}1(\hat{e}_{ij} > 0)\Big)\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji} \geq 0). && \text{(EC.36)}
\end{aligned}
$$

EC.2.2. Proof of Lemma EC.2

Proof of Lemma EC.2. For notational simplicity, we omit the argument t when referring to $V\big(e(\cdot), f(\cdot)\big), e(\cdot), f(\cdot), u(\cdot)$ and their derivatives. We first work in the case when $\{i : \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} \leq \bar{m}$, meaning that $\frac{1}{2}\dot{V}(e, f)$ is given by (EC.35). The following is a list of conditions that are necessary for $\dot{V}(e, f) = 0$, and are obtained by equating each line in (EC.35) to zero. These conditions are necessary because each line in (EC.35) is non-positive. Any condition on $\hat{e}_{ij}$ for some $i \neq j$ assumes $Q_{ij} > 0$, because if $Q_{ij} = 0$ then e_{ij} always equals zero.

1. Consider i such that $\bar{a}_i < 1$ and $\dot{u}_i = 0$. Setting line 1 of (EC.35) to zero requires $\hat{f}_{ij} > 0$ for all i, j. Line 4 requires $\hat{f}_{ji} \geq 0$ for all i, j and $\hat{e}_{ji} \geq 0$ for all $i \neq j$. Line 6 requires $\hat{e}_{ij} > 0$ for all $i \neq j$.

2. Consider i such that $\bar{a}_i < 1$ and $\dot{u}_i > 0$. One of the following three mutually exclusive sets of conditions must hold:

   a. To set line 2 to zero, we choose to enforce $\hat{f}_{ij} > 0$ for all i, j. Line 3 then requires $\hat{f}_{ji} \geq 0$ for all i, j and $\hat{e}_{ji} \geq 0$ for all $i \neq j$. Line 6 requires $\hat{e}_{ij} > 0$ for all $i \neq j$.

b. This time, to set line 2 to zero, we choose to enforce $\hat{f}_{ji} \leq 0$ for all i, j, and $\hat{e}_{ji} \leq 0$ for $i \neq j$. Line 2 then requires $\hat{f}_{ij} \leq 0$ for all i, j. Line 5 requires that if $\hat{f}_{ki} < 0$ for some k, then $\hat{e}_{ij} \leq 0$ for all $j \neq i$.

c. Suppose $f_{ij} > 0$ and $f_{ik} \leq 0$ for some $j, k = 1, \ldots, r$. The only way to make both lines 2 and 3 equal zero is to enforce $\hat{f}_{ji} = 0$ for all i, j and $\hat{e}_{ji} = 0$ for $i \neq j$.

3. Consider i such that $\bar{a}_i = 1$ and $\dot{u}_i = 0$. Since we assumed $Q_{ii} > 0$, setting line 8 to zero requires $\hat{f}_{ji} \leq 0$ for all i, j. Line 7 requires $\hat{e}_{ji} \leq 0$ for $i \neq j$, and line 5 requires that if $\hat{f}_{ki} < 0$ for some k, then $\hat{e}_{ij} \leq 0$ for all $j \neq i$.

4. Consider i such that $\bar{a}_i = 1$ and $\dot{u}_i > 0$. One of the following three mutually exclusive sets of conditions must hold:

a. To set line 2 to zero, we choose to enforce $\hat{f}_{ij} > 0$ for all i, j. Line 3 then requires $\hat{f}_{ji} \geq 0$ for all i, j and $\hat{e}_{ji} \geq 0$ for all $i \neq j$. Line 8 requires that $\hat{e}_{ij} > 0$ for all $i \neq j$.

b. This time, to set line 2 to zero, we choose to enforce $\hat{f}_{ji} \leq 0$ for all i, j, and $\hat{e}_{ji} \leq 0$ for $i \neq j$. Line 2 then requires $\hat{f}_{ij} \leq 0$ for all i, j. Line 5 requires that if $\hat{f}_{ki} < 0$ for some k, then $\hat{e}_{ij} \leq 0$ for all $j \neq i$.

c. Suppose $f_{ij} > 0$ and $f_{ik} \leq 0$ for some $j, k = 1, \ldots, r$. The only way to make both lines 2 and 3 equal zero is to enforce $\hat{f}_{ji} = 0$ for all i, j and $\hat{e}_{ji} = 0$ for $i \neq j$.

We now argue that there does not exist a configuration of e_{ij}'s and f_{ij}'s that satisfies conditions 1–4. If region i satisfies condition 1, we refer to it as a type 1 region. This convention is adopted for all other conditions as well. Recall our assumption that $P_{ij} > 0$ and $Q_{ii} > 0$ for all i, j. While we conjecture this lemma to hold even without these conditions, we impose them to prevent the following arguments from becoming even more involved.

First observe that a region i can never be of type 4a or 4c. For such a region, the fact that $\dot{u}_i > 0$, together with (29) and (EC.29) will imply that $\dot{e}_{ii} = 0$, or

$$\lambda_i(1 - \dot{u}_i) = \sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} e_{ji} + Q_{ii} \sum_{j=1}^{r} \mu_{ji} f_{ji}.$$

However, the conditions in 4a and 4c also imply that

$$\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}e_{ji}+Q_{ii}\sum_{j=1}^{r}\mu_{ji}f_{ji}\geq\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\bar{e}_{ji}+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\bar{f}_{ji}=\lambda_{i},$$

where the equality above follows from (30c) and the fact that $\bar{a}_i = 1$. This leads to a contradiction

because conditions 4a and 4c require that $\dot{u}_i > 0$.

Second, by our assumption that $\{i : \bar{a}_i = 1\} \neq \emptyset$, there must always be a region of either type 3

or type 4b. This implies that there are no type 1 or type 2a regions. If i were a type 1 or 2a region

and j were a type 3 or 4b region, then the conditions in 1 and 2a would require $\hat{f}_{ij} > 0$, but the

conditions in 3 and 4b would require that $\hat{f}_{ij} \leq 0$, which is a contradiction.

Third, we argue that there cannot be a type 2c region. Suppose i is a type 2c region. Then for

some region j, we would have $\hat{f}_{ij} > 0$. However, this region j could not belong to any of types 2b,

2c, 3, or 4b, causing a contradiction.

Lastly, we show why it cannot be that $\dot{V}(e, f) = 0$. We have shown that all regions must be of

types 2b, 3, or 4b, which means that $\hat{f}_{ji} \leq 0$ for all i, j and $\hat{e}_{ji} \leq 0$ for all $j \neq i$. We also assumed

that $\sum_{i:\bar{a}_i=1}^{r} e_{ii} \leq \bar{m}$. Since we assumed $(e, f) \notin \mathcal{E}$, at least one of the previous inequalities must be

strict. This implies that

$$\sum_{i=1}^{r}\sum_{j=1}^{r}f_{ji}+\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}e_{ji}+\sum_{i=1}^{r}e_{ii}=\sum_{i=1}^{r}\sum_{j=1}^{r}f_{ji}+\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}e_{ji}+\sum_{i:\bar{a}_i=1}e_{ii}$$

$$<\sum_{i=1}^{r}\sum_{j=1}^{r}\bar{f}_{ji}+\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}\bar{e}_{ji}+\bar{m}=1,$$

where in the first equality we used the fact that any region i with $\bar{a}_i < 1$ is of type 2b and must

satisfy $\dot{u}_i > 0$, which together with by (29) implies $e_{ii} = 0$. The result above implies that the total

mass in the system is strictly less than one, which is impossible because the total mass in the

system must always equal one. Hence, we have just shown that at any regular point, $\dot{V}(e, f) < 0$

in the case when $\{i : \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} \leq \bar{m}$.

We now assume that $\{i : \bar{a}_i = 1\} = \emptyset$, or $\{i : \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} > \bar{m}$. Just as before,

we assume that $\dot{V}(e, f) = 0$ and use (EC.36) to list the necessary conditions required for that to

happen. Observe that lines 1–3, and 5–6 of both (EC.35) and (EC.36) are identical. There is a slight difference in line 4 of both equations. In the former, the outside summation is over $i : \bar{a}_i < 1$, whereas in the latter the sum is over all i. Therefore, conditions 1, 2, and 4 must hold as before, and a region cannot be of type 4a or 4c. Only condition 3 changes slightly:

3′ Consider i such that $\bar{a}_i = 1$ and $\dot{u}_i = 0$. Setting line 4 of (EC.36) to zero requires $\hat{f}_{ji} \geq 0$ for all i, j, and $\hat{e}_{ji} \geq 0$ for all $i \neq j$. Line 6 requires that if $\hat{f}_{ki} > 0$ for some k, then $\hat{e}_{ij} > 0$ for all $j \neq i$.

Again, any condition on $\hat{e}_{ij}$ assumes $Q_{ij} > 0$, because otherwise e_{ij} always equals zero.

First, we show that there must be a region of type 2b, 2c or 4b. If that were not the case, then all regions would be of type 1, 2a or 3′. Recall our assumption that $(e, f) \notin \mathcal{E}$, $\{i : \bar{a}_i = 1\} = \emptyset$, or $\{i : \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} > \bar{m}$. If $\{i : \bar{a}_i = 1\} = \emptyset$, then all regions are of type 1 or 2a and

$$\sum_{i=1}^{r}\sum_{j=1}^{r} f_{ij} + \sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r} e_{ij} + \sum_{i=1}^{r} e_{ii} = \sum_{i=1}^{r}\sum_{j=1}^{r} f_{ij} + \sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r} e_{ji} + \sum_{i:\bar{a}_i<1}^{r} e_{ii}$$

$$> \sum_{i=1}^{r}\sum_{j=1}^{r} \bar{f}_{ij} + \sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r} \bar{e}_{ji}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{r} \bar{f}_{ij} + \sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r} \bar{e}_{ji} + \sum_{i:\bar{a}_i<1}^{r} \bar{e}_{ii} = 1.$$

The inequality holds because if region i is of type 1 or 2a, then $\hat{f}_{ij} > 0$ for all j and $\hat{e}_{ji} \geq 0$ for all $j \neq i$, and the second last equality holds because $\bar{e}_{ii} = 0$ for i such that $\bar{a}_i < 1$. The inequality above is a contradiction because the total fluid mass in the system must always equal one. By similar reasoning, if $\{i : \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} > \bar{m}$, then

$$\sum_{i=1}^{r}\sum_{j=1}^{r} f_{ij} + \sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r} e_{ij} + \sum_{i=1}^{r} e_{ii} > \sum_{i=1}^{r}\sum_{j=1}^{r} \bar{f}_{ij} + \sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r} \bar{e}_{ij} + \bar{m} = 1,$$

which is again a contradiction because fluid mass is conserved. Hence, it cannot be that all regions are of type 1, 2a or 3′. The fact that there is always a region of type 2b, 2c or 4b implies that there cannot be any type 1 or 2a regions. If j belonged to the former group and i to the latter, then the

definitions of type 1 or 2a would imply that $\hat{f}_{ij} > 0$, which would contradict the requirements in types 2b, 2c and 4b.

Second, we show that there must always be a region of type 3′. If there is a region of type 4b then it must be that $\{i : \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} > \bar{m}$. However, if i is a type 4b region, then $\bar{e}_{ii} = 0$ because $u_i > 0$. Hence, for $\sum_{i:\bar{a}_i=1}^{r} e_{ii} > \bar{m}$ to be true, the set $\{i : \bar{a}_i = 1\}$ must contain at least one type 3′ region. Now suppose there is no region of type 4b. We argue that it cannot be the case that all regions are exclusively of type 2b or exclusively of type 2c. The former case would imply that the total mass in the system is strictly less than one, and the latter case cannot happen by definition of 2c (i.e. if i were of type 2c then there would be some j with $\hat{f}_{ij} > 0$, but this j could not be of type 2b or 2c). The same reasoning implies that the regions cannot be a mixture of exclusively types 2b and 2c. Therefore, there must always exist a region of type 3′.

Lastly, we show why it cannot be that $\dot{V}(e,f) = 0$. Let i be a type 3′ region with $e_{ii} > 0$ (such a region must always exist because $\sum_{i:\bar{a}_i=1}^{r} e_{ii} > \bar{m}$). Since fluid mass is conserved, there must exist some k, ℓ, such that either $\hat{f}_{k\ell} < 0$, or $k \neq \ell$ and $\hat{e}_{k\ell} < 0$. Observe that ℓ cannot be a type 2c or 3′ region, so it must be a type 2b or 4b region. Furthermore, since i is of type 3′ and ℓ is of type 2b or 4b, it must be true that $\hat{f}_{\ell i} = 0$, and by (EC.28) this would imply that $\dot{f}_{\ell i} = 0$. We will now show that $\dot{f}_{\ell i}$ must also be strictly less than zero, leading to a contradiction. Using (19), (30a), and the fact that $f_{\ell i} = \bar{f}_{\ell i}$, it follows that

$$0 = \dot{f}_{\ell i} = \lambda_\ell P_{\ell i}(1 - \dot{u}_\ell) - \mu_{\ell i} f_{\ell i} = \lambda_\ell P_{\ell i}(1 - \dot{u}_\ell) - \lambda_\ell P_{\ell i}\bar{a}_\ell.$$

We know that $\dot{u}_\ell > 0$ by definition of a type 2b and 4b region. Hence, $e_{\ell\ell} = 0$ by (29), which in turn means that $\dot{e}_{\ell\ell} = 0$ by (EC.29), and therefore

$$\begin{aligned}
\lambda_\ell\big(1 - \dot{u}_\ell\big) &= \sum_{\substack{j=1 \\ j \neq \ell}}^{r} \mu_{j\ell} e_{j\ell} + Q_{\ell\ell} \sum_{j=1}^{r} \mu_{j\ell} f_{j\ell} \\
&= \lambda_\ell \bar{a}_\ell + \sum_{\substack{j=1 \\ j \neq \ell}}^{r} \mu_{j\ell} \hat{e}_{j\ell} + Q_{\ell\ell} \sum_{j=1}^{r} \mu_{j\ell} \hat{f}_{j\ell},
\end{aligned} \tag{EC.37}$$

where the first equality follows from differentiating (21) and setting the left hand side to zero, and the second equality follows from (30c). Combining (EC.37) with the form of $\dot{f}_{\ell i}$, we see that

$$\dot{f}_{\ell i} = P_{\ell i}\Big(\lambda_\ell \bar{a}_\ell + \sum_{\substack{j=1 \\ j\neq \ell}}^{r} \mu_{j\ell}\hat{e}_{j\ell} + Q_{\ell\ell}\sum_{j=1}^{r}\mu_{j\ell}\hat{f}_{j\ell}\Big) - \lambda_\ell P_{\ell i}\bar{a}_\ell = P_{\ell i}\Big(\sum_{\substack{j=1 \\ j\neq \ell}}^{r}\mu_{j\ell}\hat{e}_{j\ell} + Q_{\ell\ell}\sum_{j=1}^{r}\mu_{j\ell}\hat{f}_{j\ell}\Big).$$

Since ℓ is of type 2b or 4b, it follows that $\hat{f}_{j\ell} \leq 0$ for all j and $\hat{e}_{j\ell} \leq 0$ for all $j \neq \ell$. Furthermore, we know that either $\hat{f}_{k\ell} < 0$, or $k \neq \ell$ and $\hat{e}_{k\ell} < 0$, which implies that $\dot{f}_{\ell i} < 0$. However, this is a contradiction because $\dot{f}_{\ell i} = 0$ at regular points. Therefore, the necessary conditions required for $\dot{V}(e,f) = 0$ to hold cause a contradiction, and so $\dot{V}(e,f) < 0$. This concludes the proof of Lemma EC.2.

EC.2.3. Proof of Lemma EC.4

Proof of Lemma EC.4. We begin with the case that $\{i: \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} \leq \bar{m}$. Starting with (EC.34), and taking derivatives in (19)–(21), we know that

$$\frac{1}{2}\dot{V}(e,f) = \sum_{i:\bar{a}_i<1}^{r}\Big(-\lambda_i(1-\dot{u}_i) + \sum_{\substack{j=1 \\ j\neq i}}^{r}\mu_{ji}e_{ji} + Q_{ii}\sum_{j=1}^{r}\mu_{ji}f_{ji}\Big) \tag{EC.38}$$

$$+ \sum_{i=1}^{r}\sum_{\substack{j=1 \\ j\neq i}}^{r}\Big(-\mu_{ij}e_{ij} + Q_{ij}\sum_{k=1}^{r}\mu_{ki}f_{ki}\Big)\mathbf{1}\big(\hat{e}_{ij}>0\big) \tag{EC.39}$$

$$+ \sum_{i=1}^{r}\sum_{j=1}^{r}\Big(\lambda_i P_{ij}(1-\dot{u}_i) - \mu_{ij}f_{ij}\Big)\mathbf{1}\big(\hat{f}_{ij}>0\big), \tag{EC.40}$$

The first line, (EC.38), equals

$$\sum_{i:\bar{a}_i<1}^{r}\Big(-\lambda_i(1-\dot{u}_i) + \sum_{\substack{j=1 \\ j\neq i}}^{r}\mu_{ji}\bar{e}_{ji} + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\bar{f}_{ji}\Big) + \sum_{i:\bar{a}_i<1}^{r}\Big(\sum_{\substack{j=1 \\ j\neq i}}^{r}\mu_{ji}\hat{e}_{ji} + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}\Big)$$

$$= \sum_{i:\bar{a}_i<1}^{r}\Big(-\lambda_i(1-\dot{u}_i) + \lambda_i\bar{a}_i\Big) + \sum_{i:\bar{a}_i<1}^{r}\Big(\sum_{\substack{j=1 \\ j\neq i}}^{r}\mu_{ji}\hat{e}_{ji} + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}\Big)$$

$$= \sum_{i:\bar{a}_i<1}^{r}\lambda_i(\bar{a}_i-1)\mathbf{1}(\dot{u}_i=0) + \sum_{i:\bar{a}_i<1}^{r}\Big(-\lambda_i(1-\dot{u}_i) + \lambda_i\bar{a}_i\Big)\mathbf{1}(\dot{u}_i>0)$$

$$+ \sum_{i:\bar{a}_i<1}^{r}\Big(\sum_{\substack{j=1 \\ j\neq i}}^{r}\mu_{ji}\hat{e}_{ji} + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}\Big)$$

$$= \sum_{i:\bar{a}_i<1}^{r} \lambda_i(\bar{a}_i-1)1(\dot{u}_i=0)\Big(1-\sum_{j=1}^{r}P_{ij}1(\hat{f}_{ij}>0)\Big)$$

$$+ \sum_{i:\bar{a}_i<1}^{r} \lambda_i(\bar{a}_i-1)1(\dot{u}_i=0)\sum_{j=1}^{r}P_{ij}1(\hat{f}_{ij}>0)$$

$$+ \sum_{i:\bar{a}_i<1}^{r}\Big(-\lambda_i\big(1-\dot{u}_i\big)+\lambda_i\bar{a}_i\Big)1(\dot{u}_i>0)\Big(1-\sum_{j=1}^{r}P_{ij}1(\hat{f}_{ij}>0)\Big)$$

$$+ \sum_{i:\bar{a}_i<1}^{r}\Big(-\lambda_i\big(1-\dot{u}_i\big)+\lambda_i\bar{a}_i\Big)1(\dot{u}_i>0)\sum_{j=1}^{r}P_{ij}1(\hat{f}_{ij}>0)$$

$$+ \sum_{i:\bar{a}_i<1}^{r}\Big(\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}\Big), \tag{EC.41}$$

where the first equality comes from (30c). The second line, (EC.39), equals

$$\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}\Big(-\mu_{ij}\bar{e}_{ij}+Q_{ij}\sum_{k=1}^{r}\mu_{ki}\bar{f}_{ki}\Big)1\big(\hat{e}_{ij}>0\big)+\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}\Big(-\mu_{ij}\hat{e}_{ij}+Q_{ij}\sum_{k=1}^{r}\mu_{ki}\hat{f}_{ki}\Big)1\big(\hat{e}_{ij}>0\big)$$

$$= -\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}1\big(\hat{e}_{ji}>0\big)+\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}Q_{ij}1\big(\hat{e}_{ij}>0\big)\sum_{k=1}^{r}\mu_{ki}\hat{f}_{ki}, \tag{EC.42}$$

where in the equality we used (30b). Similarly, we use (30a) to see that the third line, (EC.40), equals

$$\sum_{i=1}^{r}\sum_{j=1}^{r}\Big(\lambda_iP_{ij}\big(1-\dot{u}_i\big)-\mu_{ij}\bar{f}_{ij}\Big)1\big(\hat{f}_{ij}>0\big)-\sum_{i=1}^{r}\sum_{j=1}^{r}\mu_{ij}\hat{f}_{ij}1\big(\hat{f}_{ij}>0\big)$$

$$= \sum_{i=1}^{r}1(\dot{u}_i=0)\lambda_i(1-\bar{a}_i)\sum_{j=1}^{r}P_{ij}1\big(\hat{f}_{ij}>0\big)$$

$$+ \sum_{i=1}^{r}1(\dot{u}_i>0)\sum_{j=1}^{r}P_{ij}\Big(\lambda_i\big(1-\dot{u}_i\big)-\lambda_i\bar{a}_i\Big)1\big(\hat{f}_{ij}>0\big)-\sum_{i=1}^{r}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1\big(\hat{f}_{ji}>0\big)$$

$$= \sum_{i=1}^{r}1(\dot{u}_i=0)\lambda_i(1-\bar{a}_i)\sum_{j=1}^{r}P_{ij}1\big(\hat{f}_{ij}>0\big)$$

$$+ \sum_{i:\bar{a}_i<1}^{r}\Big(\lambda_i\big(1-\dot{u}_i\big)-\lambda_i\bar{a}_i\Big)1(\dot{u}_i>0)\sum_{j=1}^{r}P_{ij}1\big(\hat{f}_{ij}>0\big)$$

$$+ \sum_{i:\bar{a}_i=1}^{r}\Big(\lambda_i\big(1-\dot{u}_i\big)-\lambda_i\bar{a}_i\Big)1(\dot{u}_i>0)\sum_{j=1}^{r}P_{ij}1\big(\hat{f}_{ij}>0\big)-\sum_{i=1}^{r}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1\big(\hat{f}_{ji}>0\big) \tag{EC.43}$$

In addition to (EC.41)–(EC.43), we will require the equation

$$\lambda_i(1-\dot{u}_i)-\lambda_ia_i=\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}, \quad \text{for } i \text{ such that } \dot{u}_i>0, \tag{EC.44}$$

which is argued in the same way as (EC.37). We now combine (EC.41)–(EC.44) to see that

$$\frac{1}{2}\dot{V}(e,f) = \sum_{i:\bar{a}_i<1}^{r} \lambda_i(\bar{a}_i-1)\Big(1-\sum_{j=1}^{r} P_{ij}1(\hat{f}_{ij}>0)\Big)1(\dot{u}_i=0) \tag{EC.45}$$

$$-\sum_{i:\bar{a}_i<1}^{r}\Big(1-\sum_{j=1}^{r} P_{ij}1(\hat{f}_{ij}>0)\Big)\Big(\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}\Big)1(\dot{u}_i>0) \tag{EC.46}$$

$$+\sum_{i:\bar{a}_i=1}^{r}\Big(\sum_{k=1}^{r} P_{ik}1(\hat{f}_{ik}>0)\Big)\Big(\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}\Big)1(\dot{u}_i>0) \tag{EC.47}$$

$$+\sum_{i:\bar{a}_i<1}^{r}\Big(\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}\Big) \tag{EC.48}$$

$$-\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}1\big(\hat{e}_{ji}>0\big)+\sum_{i=1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}Q_{ij}1\big(\hat{e}_{ij}>0\big)\sum_{k=1}^{r}\mu_{ki}\hat{f}_{ki} \tag{EC.49}$$

$$-\sum_{i=1}^{r}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1\big(\hat{f}_{ji}>0\big). \tag{EC.50}$$

Since the term above is very bulky, we manipulate one line at a time to help exposition. We leave (EC.45) as is, and decompose (EC.46) into

$$-\sum_{i:\bar{a}_i<1}^{r}\Big(1-\sum_{j=1}^{r}P_{ij}1(\hat{f}_{ij}>0)\Big)\Big(\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji}\geq 0)+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\geq 0)\Big)1(\dot{u}_i>0)$$
$$-\sum_{i:\bar{a}_i<1}^{r}\Big(1-\sum_{j=1}^{r}P_{ij}1(\hat{f}_{ij}>0)\Big)\Big(\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji}\leq 0)+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\leq 0)\Big)1(\dot{u}_i>0).$$

We leave (EC.47) as is. The term in (EC.48) equals

$$\sum_{i:\bar{a}_i<1}^{r}\Big(\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji}\geq 0)+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\geq 0)\Big)$$
$$+\sum_{i:\bar{a}_i<1}^{r}\Big(\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji}\leq 0)+Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\leq 0)\Big).$$

The term in (EC.49) equals

$$-\sum_{i:\bar{a}_i<1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}\mu_{ji}\hat{e}_{ji}1\big(\hat{e}_{ji}\geq 0\big)+\sum_{i:\bar{a}_i<1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}Q_{ij}1\big(\hat{e}_{ij}>0\big)\sum_{k=1}^{r}\mu_{ki}\hat{f}_{ki}1(\hat{f}_{ki}\leq 0)$$
$$+\sum_{i:\bar{a}_i<1}^{r}\sum_{\substack{j=1\\j\neq i}}^{r}Q_{ij}1\big(\hat{e}_{ij}>0\big)\sum_{k=1}^{r}\mu_{ki}\hat{f}_{ki}1(\hat{f}_{ki}\geq 0)$$

$$- \sum_{\substack{i:\bar{a}_i=1}} \sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1\big(\hat{e}_{ji} \geq 0\big) + \sum_{\substack{i:\bar{a}_i=1}} \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1\big(\hat{e}_{ij} > 0\big) \sum_{k=1}^{r} \mu_{ki} \hat{f}_{ki},$$

and the term in (EC.50) equals

$$- \sum_{i:\bar{a}_i<1} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1\big(\hat{f}_{ji} \geq 0\big) - \sum_{i:\bar{a}_i=1} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1\big(\hat{f}_{ji} \geq 0\big).$$

Putting all of these expansions back into (EC.45)–(EC.50), we see that

$$\frac{1}{2}\dot{V}(e,f) = \sum_{i:\bar{a}_i<1}^{r} \lambda_i(\bar{a}_i-1)\Big(1 - \sum_{j=1}^{r} P_{ij} 1(\hat{f}_{ij} > 0)\Big) 1(\dot{u}_i = 0) \tag{EC.51}$$

$$- \sum_{i:\bar{a}_i<1}^{r} \Big(1 - \sum_{j=1}^{r} P_{ij} 1(\hat{f}_{ij}>0)\Big)\Big(\sum_{\substack{j=1 \\ j\neq i}}^{r} \mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji}\geq 0) + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\geq 0)\Big)1(\dot{u}_i>0)$$
$$\tag{EC.52}$$

$$+ \sum_{i:\bar{a}_i<1}^{r} \Big(\sum_{j=1}^{r} P_{ij} 1(\hat{f}_{ij}>0)\Big)\Big(\sum_{\substack{j=1 \\ j\neq i}}^{r} \mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji}\leq 0) + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\leq 0)\Big)1(\dot{u}_i>0)$$
$$\tag{EC.53}$$

$$+ \sum_{i:\bar{a}_i<1}^{r} 1(\dot{u}_i=0)\Big(\sum_{\substack{j=1 \\ j\neq i}}^{r} \mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji}\leq 0) + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\leq 0)\Big) \tag{EC.54}$$

$$+ \sum_{i:\bar{a}_i<1}^{r} \sum_{\substack{j=1 \\ j\neq i}}^{r} Q_{ij} 1\big(\hat{e}_{ij}>0\big) \sum_{k=1}^{r} \mu_{ki}\hat{f}_{ki}1(\hat{f}_{ki}\leq 0) \tag{EC.55}$$

$$- \sum_{i:\bar{a}_i<1}^{r} \Big(1 - Q_{ii} - \sum_{\substack{j=1 \\ j\neq i}}^{r} Q_{ij} 1\big(\hat{e}_{ij}>0\big)\Big) \sum_{j=1}^{r} \mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\geq 0) \tag{EC.56}$$

$$+ \sum_{i:\bar{a}_i=1}^{r} \Big(\sum_{k=1}^{r} P_{ik} 1(\hat{f}_{ik}>0)\Big)\Big(\sum_{\substack{j=1 \\ j\neq i}}^{r} \mu_{ji}\hat{e}_{ji} + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}\Big)1(\dot{u}_i>0) \tag{EC.57}$$

$$- \sum_{i:\bar{a}_i=1}^{r} \sum_{\substack{j=1 \\ j\neq i}}^{r} \mu_{ji}\hat{e}_{ji}1\big(\hat{e}_{ji}\geq 0\big) + \sum_{i:\bar{a}_i=1}^{r} \sum_{\substack{j=1 \\ j\neq i}}^{r} Q_{ij}1\big(\hat{e}_{ij}>0\big)\sum_{k=1}^{r}\mu_{ki}\hat{f}_{ki} \tag{EC.58}$$

$$- \sum_{i:\bar{a}_i=1}^{r} \sum_{j=1}^{r} \mu_{ji}\hat{f}_{ji}1\big(\hat{f}_{ji}\geq 0\big). \tag{EC.59}$$

It remains to manipulate the terms in (EC.57)–(EC.59) to get them into the form we need. We begin with (EC.57), which equals

$$\sum_{i:\bar{a}_i=1}^{r} \Big(\sum_{k=1}^{r} P_{ik} 1(\hat{f}_{ik}>0)\Big)\Big(\sum_{\substack{j=1 \\ j\neq i}}^{r} \mu_{ji}\hat{e}_{ji}1(\hat{e}_{ji}\leq 0) + Q_{ii}\sum_{j=1}^{r}\mu_{ji}\hat{f}_{ji}1(\hat{f}_{ji}\leq 0)\Big)1(\dot{u}_i>0)$$

$$+ \sum_{i:\bar{a}_i=1}^{r} \Big(\sum_{k=1}^{r} P_{ik} 1(\hat{f}_{ik} > 0) \Big) \Big(\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \geq 0) + Q_{ii} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0) \Big) 1(\dot{u}_i > 0).$$

The term in (EC.58) equals

$$- \sum_{i:\bar{a}_i=1}^{r} \sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \geq 0) + \sum_{i:\bar{a}_i=1}^{r} \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1(\hat{e}_{ij} > 0) \sum_{k=1}^{r} \mu_{ki} \hat{f}_{ki} 1(\hat{f}_{ki} \leq 0)$$
$$+ \sum_{i:\bar{a}_i=1}^{r} \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1(\hat{e}_{ij} > 0) \sum_{k=1}^{r} \mu_{ki} \hat{f}_{ki} 1(\hat{f}_{ki} \geq 0),$$

and the term in (EC.59) equals

$$- \sum_{i:\bar{a}_i=1}^{r} \Big(1 - Q_{ii} 1(\dot{u}_i > 0) \Big) \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0) - \sum_{i:\bar{a}_i=1}^{r} Q_{ii} 1(\dot{u}_i > 0) \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0).$$

Inserting these expansions back into (EC.57)–(EC.59), we see that

$$\sum_{i:\bar{a}_i=1}^{r} \Big(\sum_{k=1}^{r} P_{ik} 1(\hat{f}_{ik} > 0) \Big) \Big(\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} + Q_{ii} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} \Big) 1(\dot{u}_i > 0)$$
$$- \sum_{i:\bar{a}_i=1}^{r} \sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \geq 0) + \sum_{i:\bar{a}_i=1}^{r} \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1(\hat{e}_{ij} > 0) \sum_{k=1}^{r} \mu_{ki} \hat{f}_{ki}$$
$$- \sum_{i:\bar{a}_i=1}^{r} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0)$$
$$= \sum_{i:\bar{a}_i=1}^{r} \Big(\sum_{k=1}^{r} P_{ik} 1(\hat{f}_{ik} > 0) \Big) \Big(\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \leq 0) + Q_{ii} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \leq 0) \Big) 1(\dot{u}_i > 0)$$
$$- \sum_{i:\bar{a}_i=1}^{r} \Big(1 - \sum_{k=1}^{r} P_{ik} 1(\hat{f}_{ik} > 0) \Big) \Big(\sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \geq 0) + Q_{ii} \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0) \Big) 1(\dot{u}_i > 0)$$
$$- \sum_{i:\bar{a}_i=1}^{r} 1(\dot{u}_i = 0) \sum_{\substack{j=1 \\ j \neq i}}^{r} \mu_{ji} \hat{e}_{ji} 1(\hat{e}_{ji} \geq 0) + \sum_{i:\bar{a}_i=1}^{r} \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1(\hat{e}_{ij} > 0) \sum_{k=1}^{r} \mu_{ki} \hat{f}_{ki} 1(\hat{f}_{ki} \leq 0)$$
$$- \sum_{i:\bar{a}_i=1}^{r} \Big(1 - Q_{ii} 1(\dot{u}_i > 0) - \sum_{\substack{j=1 \\ j \neq i}}^{r} Q_{ij} 1(\hat{e}_{ij} > 0) \Big) \sum_{j=1}^{r} \mu_{ji} \hat{f}_{ji} 1(\hat{f}_{ji} \geq 0).$$

The form of $\frac{1}{2}\dot{V}(e,f)$ we obtain by plugging the above back into (EC.51)–(EC.59) can be compared with (EC.35) to see that it matches.

Now suppose that $\{i : \bar{a}_i = 1\} = \emptyset$, or $\{i : \bar{a}_i = 1\} \neq \emptyset$ and $\sum_{i:\bar{a}_i=1}^{r} e_{ii} > \bar{m}$. We argue that we have already done all the hard work to justify (EC.36). Indeed, the same logic used to derive (EC.38)–(EC.40) implies that

$$
\begin{aligned}
\frac{1}{2}\dot{V}(e,f) = {} & \sum_{i=1}^{r} \left(-\lambda_i(1-\dot{u}_i) + \sum_{\substack{j=1 \\ j\neq i}}^{r} \mu_{ji}e_{ji} + Q_{ii}\sum_{j=1}^{r}\mu_{ji}f_{ji} \right) \\
& + \sum_{i=1}^{r}\sum_{\substack{j=1 \\ j\neq i}}^{r} \left(-\mu_{ij}e_{ij} + Q_{ij}\sum_{k=1}^{r}\mu_{ki}f_{ki} \right)1(\hat{e}_{ij} > 0) \\
& + \sum_{i=1}^{r}\sum_{j=1}^{r} \left(\lambda_i P_{ij}(1-\dot{u}_i) - \mu_{ij}f_{ij} \right)1(\hat{f}_{ij} > 0).
\end{aligned}
$$

The difference between the equation above and that of (EC.38)–(EC.40) is that the summation in the first line is over all $i = 1, \dots, r$, as opposed to only those i for which $\bar{a}_i < 1$. It can therefore be verified, by repeating the logic of this proof, that $\frac{1}{2}\dot{V}(e,f)$ equals (EC.51)–(EC.56) with summations over all $i = 1, \dots, r$, instead of only i such that $\bar{a}_i < 1$. This verifies (EC.36) and concludes the proof of this lemma. $\blacksquare$

EC.2.4. Proof of Lemma EC.3

Proof of Lemma EC.3.   Lemma EC.3 follows from the argument used in (Khalil 2002, Lemma 4.1) after the following observation. In Khalil (2002), the author states Lemma 4.1 for functions $x : \mathbb{R} \to \mathbb{R}^n$ that satisfy

$$
\dot{x}(t) = g(x(t)), \tag{EC.60}
$$

for some locally Lipschitz $g : \mathbb{R}^n \to \mathbb{R}^n$. Only three properties of $x(t)$ are really necessary for Lemma 4.1: that $x(t)$ is guaranteed to exist, be unique, and be continuous with respect to its initial condition. The latter property means that if $x(t)$ and $\tilde{x}(t)$ both satisfy (EC.60) with $x(0) \neq \tilde{x}(0)$, then for every $t \geq 0$ and every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$
|x(0) - \tilde{x}(0)| < \delta \quad \Rightarrow \quad |x(t) - \tilde{x}(t)| < \epsilon. \tag{EC.61}
$$

In our case, $\big(e(t), f(t)\big)$ are defined as the solution to (19)–(22) and cannot be written in the form of (EC.60). In fact, $\big(\dot{e}(t), \dot{f}(t)\big)$ is not even guaranteed to exist for all $t \geq 0$. Nevertheless, we know that $\big(e(t), f(t)\big)$ is unique by Lemma EC.1, and that it satisfies the analogue of (EC.61) by (EC.13) of Lemma EC.1. Hence, the proof of Lemma 4.1 can be carried through to prove Lemma EC.3 as well.

EC.3. Two Examples of Networks

In this section we describe both the 9-region, and 5-region networks we used for the numerical examples in Section 3.

EC.3.1. 9-Region Didi Network

The Di-Tech Challenge data set contains individual order information for trips taken between January 1, 2016 until January 21, 2016, in an unspecified city in China. The city is partitioned into a number of distinct geographical regions. The data is divided into 10-minute time slots, i.e. 144 times slots per day. Each data entry represents a single order. An order is a passenger request for a car, and may or may not be fulfilled due to lack of cars in proximity. A single data entry contains information about the origin and intended destination of the order, a time-stamp of when the order was made, whether the order was fulfilled, and in the case when the order was fulfilled, the ID of the car fulfilling the order, as well as the total trip price. Although the data set had more than nine regions, we focus on these because they are the 'major' ones. That is, they have a much higher volume of request rates compared to the rest of the regions.

We restrict our attention to data between 5-6pm each day, because we identified this to be the time of the evening rush hour; see Figure 4 in Section 3.2. Using the data set, we extract a nine-region network. We estimate λ, P, and μ as follows. To calculate P_{ij}, we consider all rides

happening in the 5-6pm window each day. We tally the total number of orders from i to j, and

divide by the total number of orders originating at i. The result is the matrix P, which equals

$$
\begin{pmatrix}
\begin{array}{c|ccccccccc}
\text{Region} & 10 & 11 & 18 & 13 & 19 & 27 & 45 & 47 & 50 \\
\hline
10 & 0.230 & 0.297 & 0.372 & 0.004 & 0.026 & 0.029 & 0.009 & 0.018 & 0.015 \\
11 & 0.044 & 0.655 & 0.146 & 0.005 & 0.079 & 0.038 & 0.018 & 0.005 & 0.011 \\
18 & 0.165 & 0.291 & 0.288 & 0.007 & 0.054 & 0.126 & 0.017 & 0.025 & 0.027 \\
13 & 0.0013 & 0.010 & 0.006 & 0.139 & 0.031 & 0.185 & 0.101 & 0.117 & 0.409 \\
19 & 0.005 & 0.096 & 0.026 & 0.037 & 0.25 & 0.333 & 0.218 & 0.012 & 0.027 \\
27 & 0.004 & 0.031 & 0.032 & 0.088 & 0.121 & 0.426 & 0.148 & 0.059 & 0.092 \\
45 & 0.002 & 0.023 & 0.011 & 0.066 & 0.142 & 0.269 & 0.399 & 0.020 & 0.069 \\
47 & 0.004 & 0.008 & 0.023 & 0.067 & 0.011 & 0.095 & 0.019 & 0.400 & 0.374 \\
50 & 0.001 & 0.004 & 0.005 & 0.095 & 0.010 & 0.059 & 0.030 & 0.185 & 0.610 \\
\end{array}
\end{pmatrix}. \tag{EC.62}
$$

To calculate μ, we used the average trip cost as a proxy for travel times, since travel times are not

provided in the data set. Trip costs are a reasonable proxy because the price of a trip is typically a

linear function of distance traveled, and time spent in car. To estimate μ_{ij}, we first calculated the

average trip cost between regions i and j, and then set the average travel time to equal the average

trip cost. Since our time unit is a time-slot, which is 10-minute interval, we set $1/\mu_{ij}$ to equal the

average trip cost divided by 10. For example, the average trip cost between region 47 and region

50 is 14.1 CNY, so we assumed the average travel time is 14 minutes, and set $\mu_{47,50} = 1.41$. The

resulting matrix $1/\mu$ is

$$
\begin{pmatrix}
\begin{array}{c|ccccccccc}
\text{Region} & 10 & 11 & 18 & 13 & 19 & 27 & 45 & 47 & 50 \\
\hline
10 & 0.83 & 1.87 & 1.07 & 3.89 & 3.25 & 2.79 & 4.25 & 2.94 & 4.37 \\
11 & 1.78 & 0.89 & 1.18 & 3.24 & 1.24 & 1.99 & 2.89 & 3.46 & 4.18 \\
18 & 1.02 & 1.31 & 0.78 & 2.82 & 1.45 & 1.36 & 3.26 & 2.17 & 3.04 \\
13 & 3.52 & 3.13 & 2.76 & 0.93 & 1.5 & 1.26 & 1.49 & 1.75 & 1.6 \\
19 & 2.86 & 1.42 & 1.64 & 1.55 & 0.84 & 1.04 & 1.45 & 2.88 & 2.89 \\
27 & 2.61 & 2.17 & 1.54 & 1.31 & 1.15 & 0.81 & 1.86 & 1.78 & 2.2 \\
45 & 4.38 & 3.02 & 2.79 & 1.36 & 1.35 & 1.65 & 0.94 & 3.1 & 3 \\
47 & 2.93 & 3.06 & 2.26 & 1.75 & 2.69 & 1.62 & 3.23 & 0.9 & 1.48 \\
50 & 3.58 & 4.18 & 2.8 & 1.49 & 2.46 & 2.02 & 2.72 & 1.43 & 1.01 \\
\end{array}
\end{pmatrix} . \tag{EC.63}
$$

To determine the arrival rate to region i, we counted the average number of orders to region i per time slot. This gave us an estimate $N\lambda_i$. Our data did not provide the exact number N of cars in the network. To determine a reasonable choice for N, we summed up the number of fulfilled orders across all 9 regions in Figure 2. As a result, we chose $N = 2000$. Although not exact, this number is of the correct order of magnitude. Furthermore, our numerical results in Section 3 remained (qualitatively) consistent even for different choices of N around 2000. Hence, we set our vector λ to

$$
\begin{pmatrix}
\begin{array}{c|ccccccccc}
\text{Region } i & 10 & 11 & 18 & 13 & 19 & 27 & 45 & 47 & 50 \\
\hline
\lambda_i & 0.0131 & 0.0624 & 0.0381 & 0.0652 & 0.0870 & 0.1178 & 0.0762 & 0.1438 & 0.2751 \\
\end{array}
\end{pmatrix} \tag{EC.64}
$$

EC.3.2. 5-Region Network

Let us consider a simplified model of a city illustrated in Figure EC.1 that consists of 5 regions: a downtown area D, a midtown area M, and three suburban areas S_1, S_2, S_3. The downtown area represents a central business district, where many people work but few people live. Midtown represents a region with restaurants and night-life, where people visit after work. The suburb regions are residential areas, and do not have as many entertainment options as midtown. For
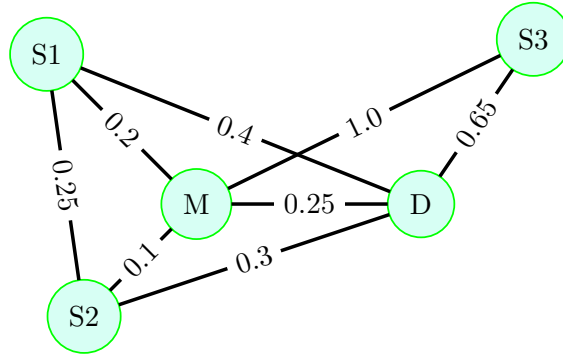
Figure EC.1   The 5-region city where edge weights represent mean travel times (in hours). Travel times longer
than 1 hour are omitted to help with vizualisation.

convenience, we enumerate S_1, S_2, S_3, M, D as $1, 2, 3, 4, 5$, respectively. The parameters from 5-11pm
are as follows:

1. From 5-7pm, the city experiences a rush hour as people go home from work. Most of the
   traffic originates in downtown, and flows into the suburbs as people go home after work. The
   parameters during this time-slot are

$$
\lambda = \begin{pmatrix} 0.108 \\ 0.108 \\ 0.108 \\ 0.108 \\ 1.08 \end{pmatrix}, \quad
P = \begin{pmatrix} 0.6 & 0.1 & 0 & 0.3 & 0 \\ 0.1 & 0.6 & 0 & 0.3 & 0 \\ 0 & 0 & 0.7 & 0.3 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.1 & 0 \end{pmatrix}, \quad
\left(1/\mu_{ij} \right) = \begin{pmatrix} 0.15 & 0.25 & 1.25 & 0.2 & 0.4 \\ 0.25 & 0.10 & 1.1 & 0.1 & 0.3 \\ 1.25 & 1.1 & 0.1 & 1 & 0.65 \\ 0.25 & 0.15 & 1 & 0.15 & 0.25 \\ 0.5 & 0.4 & 0.75 & 0.25 & 0.2 \end{pmatrix}.
$$

   Above, $\left(1/\mu_{ij} \right)$ is the matrix of mean travel times (in hours) from region i to j, and λ is a
   vector representing the number of passenger arrivals per hour, per car.

2. From 7-9pm, most of the traffic is headed into midtown as people go out in the evening to restaurants and for entertainment. The parameters during this time-slot are

$$\lambda = \begin{pmatrix} 0.72 \\ 0.48 \\ 0.48 \\ 0.48 \\ 0.12 \end{pmatrix}, \quad P = \begin{pmatrix} 0.1 & 0 & 0 & 0.9 & 0 \\ 0 & 0.1 & 0 & 0.9 & 0 \\ 0 & 0 & 0.1 & 0.9 & 0 \\ 0.05 & 0.05 & 0.05 & 0.8 & 0.05 \\ 0 & 0 & 0 & 0.9 & 0.1 \end{pmatrix}, \quad (1/\mu_{ij}) = \begin{pmatrix} 0.15 & 0.25 & 1.25 & 0.2 & 0.4 \\ 0.25 & 0.10 & 1.1 & 0.1 & 0.3 \\ 1.25 & 1.1 & 0.1 & 1 & 0.65 \\ 0.2 & 0.1 & 1 & 0.15 & 0.25 \\ 0.4 & 0.3 & 0.65 & 0.25 & 0.2 \end{pmatrix}$$

3. From 9-11pm, traffic flows mainly from midtown to the suburbs as people go home for the night. The parameters during this time-slot are

$$\lambda = \begin{pmatrix} 0.12 \\ 0.12 \\ 0.12 \\ 1.32 \\ 0.12 \end{pmatrix}, \quad P = \begin{pmatrix} 0.9 & 0.05 & 0 & 0.05 & 0 \\ 0.05 & 0.9 & 0 & 0.05 & 0 \\ 0 & 0 & 0.9 & 0.1 & 0 \\ 0.3 & 0.3 & 0.3 & 0.05 & 0.05 \\ 0 & 0 & 0 & 0.1 & 0.9 \end{pmatrix}, \quad (1/\mu_{ij}) = \begin{pmatrix} 0.15 & 0.25 & 1.25 & 0.2 & 0.4 \\ 0.25 & 0.10 & 1.1 & 0.1 & 0.3 \\ 1.25 & 1.1 & 0.1 & 1 & 0.65 \\ 0.2 & 0.1 & 1 & 0.15 & 0.25 \\ 0.4 & 0.3 & 0.65 & 0.25 & 0.2 \end{pmatrix}$$

EC.4. Achieving 100% Availability Globally

It may be the case, e.g. during rush hours, that customer demand is so high that there are simply not enough cars in the system to fulfill all passenger requests, no matter what kind of routing policy is used. A natural question to ask then is how many cars does one need to achieve 100% availability in every region? In this section we address this question using the fluid-based optimization problem.

Fix $N > 0$ and some static routing policy Q. Since the CTMC $\left(E^{(N)}, F^{(N)}\right)$ is positive recurrent and has a finite state space, it must be the case that $A_i^{(N)}(\infty) < 1$. In other words, with a finite number of cars, the availability at any region can never be 100%. However, it is possible to have $A_i^{(N)}(\infty) \to 1$ as $N \to \infty$. We say that our system (asymptotically) achieves perfect availability (under routing policy Q) if

$$\lim_{N \to \infty} A_i^{(N)}(\infty) = 1, \quad 1 \le i \le r.$$

We also say that perfect availability is feasible if there exists *some* routing matrix Q under which the system achieves perfect availability.

The question of whether perfect availability is feasible is a question of whether there is enough supply of cars to meet passenger demand. If perfect availability is not feasible, we want to know the amount by which to increase our fleet size in order to make it feasible. Conversely, if perfect availability is feasible, we want to know how much slack, or excess capacity, our system has. Since the passenger arrival rate to each region, $N\lambda_i$, depends on the number of cars N, we need to clarify what we mean by increasing or decreasing the fleet size.

Consider a system with N cars, passenger arrival rate $N\lambda_i$, travel choices P_{ij}, and travel time means $1/\mu_{ij}$, and recall that the associated fluid-based optimization problem is given by (4a)–(4g). We can consider a related system, where all parameters are the same except that the number of cars is now κN for some $\kappa > 0$. In this system, N is no longer the number of cars, but represents the size of the market. The passenger arrival rate to region i is $N\lambda_i = \kappa N(\lambda_i/\kappa)$, and therefore, the associated fluid-based optimization problem is (4a)–(4g), but with λ_i replaced by λ_i/κ there. That is, multiplying the number of cars in the system by a factor of κ is equivalent to dividing λ_i by a factor of κ. This makes sense because λ_i is the arrival rate of passengers *per car* to region i.

The following result states that determining the balance between supply and demand needed to achieve perfect availability can be done by solving a linear program. The proof can be found in Appendix EC.4.1.

THEOREM EC.2. *The feasibility region of the linear program*

$$\min_q \sum_{i=1}^{r}\sum_{j=1}^{r} \bar{f}_{ij} + \sum_{i=1}^{r}\sum_{j=1,j\neq i}^{r} \bar{e}_{ij} \tag{EC.65a}$$

$$\textit{subject to} \quad \lambda_i P_{ij} = \mu_{ij}\bar{f}_{ij}, \quad 1 \leq i,j \leq r, \tag{EC.65b}$$

$$\mu_{ij}\bar{e}_{ij} = q_{ij}\sum_{k=1}^{r}\mu_{ki}\bar{f}_{ki}, \quad 1 \leq i,j \leq r,\ j \neq i, \tag{EC.65c}$$

$$\lambda_i = \sum_{k=1,k\neq i}^{r}\mu_{ki}\bar{e}_{ki} + q_{ii}\sum_{k=1}^{r}\mu_{ki}\bar{f}_{ki}, \quad 1 \leq i \leq r, \tag{EC.65d}$$

$$q_{ij} \geq 0, \quad \sum_{j=1}^{r} q_{ij} = 1, \quad 1 \leq i,j \leq r. \tag{EC.65e}$$

is non-empty. Let $\kappa > 0$ be the optimal objective value, and assume that

$$\frac{1}{\mu_{ik}} \leq \frac{1}{\mu_{ij}} + \frac{1}{\mu_{jk}}, \quad 1 \leq i \neq j \neq k \leq r,$$

i.e. travel time means satisfy the triangle inequality.

1. *If $\kappa > 1$ then perfect availability is not feasible, but it becomes feasible if λ is reduced to λ/κ, i.e. an increase in fleet size by a factor of κ.*

2. *If $\kappa \leq 1$ then perfect availability is feasible, and it remains feasible even if λ is increased to λ/κ.*

The value κ from Theorem EC.2 is the minimal 'fluid mass' needed in the system to achieve perfect availability, and can be interpreted a measure of imbalance of supply and demand in the system. If $\kappa > 1$, then passenger demand exceeds vehicle supply, and if $\kappa < 1$ then the reverse is true.

EC.4.1. Proof of Theorem EC.2

Proof of Theorem EC.2   To see that the feasibility region is non-empty (and hence an optimal solution exists), observe that

$$q_{ij} = \frac{\mu_{ji} f_{ji}}{\sum_{k=1}^{r} \mu_{ki} f_{ki}}, \quad 1 \leq i, j \leq r \tag{EC.66}$$

is a feasible solution. The intuition behind deriving (EC.66) is that every car that completes a trip from i to j with a passenger must drive back empty from j to i.

Recall that $\kappa > 0$ is the optimal value of the linear program, and let q^{κ} be the solution that achieves this minimum. Also let $\bar{e}^{\kappa}$ and $\bar{f}^{\kappa}$ be the corresponding values of $\bar{e}$ and $\bar{f}$ under q^{κ}, i.e. the optimal value $\kappa = \sum_{i=1}^{r} \sum_{j=1}^{r} \bar{f}_{ij}^{\kappa} + \sum_{i=1}^{r} \sum_{j=1, j \neq i}^{r} \bar{e}_{ij}^{\kappa}$; observe that (EC.65b)–(EC.65e) place no constraints on $\bar{e}_{ii}^{\kappa}$, so we will choose our $\bar{e}^{\kappa}$ with $\bar{e}_{ii}^{\kappa} = 0$. Lastly, let $\bar{a}^{\kappa} \in \mathbb{R}^r$ be a vector whose elements all equal to one.

Suppose first that $\kappa > 1$. To show that perfect availability is not feasible, we argue that if $(q, \bar{e}, \bar{f}, \bar{a})$ is a point in the feasible region of the fluid-based optimization (4a)–(4g), then it must be that $\bar{a}_i < 1$ for some i. Assume this is not the case, i.e. $(q, \bar{e}, \bar{f}, \bar{a})$ satisfies the constraints of

the fluid-based optimization and $\bar{a}_i = 1$ for all $1 \le i \le r$. Then (4b)–(4d) and (4g) are identical to (EC.65b)–(EC.65d) and (EC.65e), respectively. Since the optimal value in (EC.65a) is greater than one, it means that constraint (4f) of the fluid-based optimization can never be satisfied under any q, which proves that perfect availability is not feasible.

Now consider a new system where λ_i is replaced instead by λ_i/κ. It can be checked that (EC.65b)–(EC.65e) in this new system is satisfied by q^κ, and that the resulting objective value is

$$
\frac{1}{\kappa}\Big(\sum_{i=1}^{r}\sum_{j=1}^{r}\bar{f}_{ij}^{\kappa}+\sum_{i=1}^{r}\sum_{j=1,j\neq i}^{r}\bar{e}_{ij}^{\kappa}\Big)=1.
$$

One can then check that $(q^\kappa, \bar{e}^\kappa/\kappa, \bar{f}^\kappa/\kappa, \bar{a}^\kappa)$ is a feasible solution to the fluid-based optimization problem of this new system. Assuming for now that $q_{ii}^\kappa > 0$ for all $1 \le i \le r$, we can invoke Theorem 1 to conclude that the new system achieves perfect availability under q^κ.

Now if $\kappa \le 1$, a similar argument can be used to see that the system achieves perfect availability under routing matrix q^κ, i.e. one confirms that $(q^\kappa, \bar{e}^\kappa, \bar{f}^\kappa, \bar{a}^\kappa)$ is a feasible solution to the fluid-based optimization. Furthermore, if we consider a new system where λ_i is replaced by λ_i/κ, then the new system still achieves perfect availability when q^κ is taken to be the routing matrix; this can again be verified just like in the $\kappa > 1$ case.

To conclude the proof, we need to show that we can always find an optimal solution to (EC.65a)–(EC.65e) such that $q_{ii}^\kappa > 0$ for all $1 \le i \le r$ (this is a minor technical condition needed to invoke Theorem 1). Suppose q^κ is an optimal solution and $q_{ii}^\kappa = 0$ for some i; we now construct another optimal solution $\hat{q}^\kappa$ such that $\hat{q}_{ii}^\kappa > 0$. We know that

$$
0 < \lambda_i = \sum_{k=1,k\neq i}^{r}\mu_{ki}\bar{e}_{ki}^{\kappa}+q_{ii}^{\kappa}\sum_{k=1}^{r}\mu_{ki}\bar{f}_{ki}^{\kappa}=\sum_{k=1,k\neq i}^{r}\mu_{ki}\bar{e}_{ki}^{\kappa},
$$

where the first equality is from (EC.65d). The above implies that there must exist some ℓ such that $\bar{e}_{\ell i}^\kappa > 0$, and consequently (by (EC.65c) and (EC.65b)) $q_{\ell i}^\kappa > 0$. Furthermore, since $q_{ii}^\kappa = 0$, there must exist some m such that $q_{im}^\kappa > 0$, and hence $\bar{e}_{im}^\kappa > 0$.

The idea of the following argument is to redirect some empty cars going from ℓ to i to instead go from ℓ to m, i.e. reduce $q_{\ell i}^\kappa$ to increase $q_{\ell m}^\kappa$. This will allow region i to keep some of the cars it

otherwise would have sent to m, i.e. reduce q_{im}^κ to increase q_{ii}^κ. It remains to specify precisely the changes to q^κ so that the objective value in (EC.65a) does not increase. To this end, fix $\varepsilon > 0$ and let

$$\hat{q}_{\ell i}^\kappa = q_{\ell i}^\kappa - \varepsilon$$

$$\hat{q}_{\ell m}^\kappa = q_{\ell m}^\kappa + \varepsilon$$

$$\hat{q}_{im}^\kappa = q_{im}^\kappa - \varepsilon \frac{\sum_{k=1}^r \lambda_k P_{k\ell}}{\sum_{k=1}^r \lambda_k P_{ki}}$$

$$\hat{q}_{ii}^\kappa = q_{ii}^\kappa + \varepsilon \frac{\sum_{k=1}^r \lambda_k P_{k\ell}}{\sum_{k=1}^r \lambda_k P_{ki}},$$

and let all other elements of $\hat{q}^\kappa$ be the same as those of q^κ. Provided ε is small enough, $\hat{q}^\kappa$ satisfies (EC.65e), i.e. its rows are probability distributions. It is also not hard to check that $\hat{q}^\kappa$ satisfies (EC.65b)–(EC.65d), and that the objective value under $\hat{q}^\kappa$ is

$$\sum_{i=1}^r \sum_{j=1}^r \bar{f}_{ij}^\kappa + \sum_{i=1}^r \sum_{j=1, j \neq i}^r \bar{e}_{ij}^\kappa + \varepsilon \sum_{k=1}^r \lambda_k P_{k\ell} \left(\frac{1}{\mu_{\ell m}} - \frac{1}{\mu_{\ell i}} - \frac{1}{\mu_{im}} \right),$$

which is not larger than the objective value under q^κ by our assumption that the mean travel times satisfy the triangle inequality. Therefore, $\hat{q}^\kappa$ is an optimal solution, and this concludes the proof.

EC.5. Miscellaneous Proofs

EC.5.1. Proof of Lemma 1

Proof of Lemma 1. Given $(\bar{e}, \bar{f}, \bar{a})$ and q that satisfy (4b)–(4g), it can be easily verified that $(\bar{e}, \bar{f}, \bar{a})$ satisfies conditions (10a)–(10g) based on the fact that $0 \leq q_{ij} \leq 1$.

Now given $(\bar{e}, \bar{f}, \bar{a})$ that satisfies (10a)–(10g), we define

$$q_{ij} = \frac{\mu_{ij} \bar{e}_{ij}}{\sum_{k=1}^r \mu_{ki} \bar{f}_{ki}}, \quad 1 \leq i \neq j \leq r, \quad q_{ii} = \frac{\lambda_i \bar{a}_i - \sum_{k=1, k \neq i}^r \mu_{ki} \bar{e}_{ki}}{\sum_{k=1}^r \mu_{ki} \bar{f}_{ki}} \quad 1 \leq i \leq r,$$

and let q be the $r \times r$ matrix whose entries are q_{ij}. Then conditions (4c) and (4d) hold according to the definition of q. Furthermore, $q_{ij} \geq 0$ because $\mu_{ij} \geq 0$, $\bar{e}_{ij} \geq 0$, and $\lambda_i \bar{a}_i \geq \sum_{k=1, k \neq i}^r \mu_{ki} \bar{e}_{ki}$ by (10c). Finally,

$$\sum_{j=1}^r q_{ij} = \sum_{j=1, j \neq i}^r \frac{\mu_{ij} \bar{e}_{ij}}{\sum_{k=1}^r \mu_{ki} \bar{f}_{ki}} + \frac{\lambda_i \bar{a}_i - \sum_{k=1, k \neq i}^r \mu_{ki} \bar{e}_{ki}}{\sum_{k=1}^r \mu_{ki} \bar{f}_{ki}}$$

$$= \frac{\lambda_i \bar{a}_i + \sum_{j=1, j\neq i}^r \mu_{ij}\bar{e}_{ij} - \sum_{k=1, k\neq i}^r \mu_{ki}\bar{e}_{ki}}{\sum_{k=1}^r \mu_{ki}\bar{f}_{ki}}$$

$$\overset{(a)}{=} \frac{\sum_{k=1}^r \mu_{ki}\bar{f}_{ki}}{\sum_{k=1}^r \mu_{ki}\bar{f}_{ki}}$$

$$= 1,$$

where equality (a) is obtained based on (10d). Therefore, (4g) holds and we can conclude that $(\bar{e}, \bar{f}, \bar{a})$ and our newly defined q satisfy (4b)–(4g).

EC.5.2. Proof of Lemma 2

Proof of Lemma 2.   Suppose $\bar{a}_i^* < 1$, for all $1 \le i \le r$. We argue by contradiction that this implies $\bar{e}_{ii}^* = 0$, for all i. Suppose there exists region i' such that $\bar{e}_{i'i'}^* > 0$. We now construct another solution $(\tilde{e}, \tilde{f}, \tilde{a})$ that is better than $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$. Assume for now that there exists an $r \times r$ matrix with non-negative entries π_{ij}, such that $\sum_{i=1}^r \sum_{j=1}^r \pi_{ij} = 1$, and

$$P_{ij} \sum_k \pi_{ki}\mu_{ki} = \mu_{ij}\pi_{ij}, \quad 1 \le i, j \le r. \tag{EC.67}$$

Fix $\epsilon > 0$ to be specified later, and let

$$\tilde{e}_{i'i'} = \bar{e}_{i'i'}^* - \epsilon$$

$$\tilde{e}_{ii} = \bar{e}_{ii}^*, \quad i \neq i',$$

$$\tilde{f}_{ij} = \bar{f}_{ij}^* + \epsilon\pi_{ij}, \quad 1 \le i, j \le r,$$

$$\tilde{e}_{ij} = \bar{e}_{ij}^*, \quad i \neq j,$$

$$\tilde{a}_i = \bar{a}_i^* + \epsilon\frac{\sum_{k=1}^r \pi_{ki}\mu_{ki}}{\lambda_i}, \quad 1 \le i \le r.$$

Since $\tilde{a}_i \ge \bar{a}_i^*$ for all i, it follows that

$$\sum_{i=1}^r \sum_{j=1}^r \bar{a}_i^* \lambda_i P_{ij} c_{ij} \le \sum_{i=1}^r \sum_{j=1}^r \tilde{a}_i \lambda_i P_{ij} c_{ij}. \tag{EC.68}$$

We now check that $(\tilde{e}, \tilde{f}, \tilde{a})$ satisfies (10a)–(10f), and is therefore a feasible solution. Since $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$ is a feasible solution and satisfies (10a), it follows that

$$\lambda_i P_{ij} \tilde{a}_i = \lambda_i P_{ij} \left(\bar{a}_i^* + \epsilon\frac{\sum_{k=1}^r \pi_{ki}\mu_{ki}}{\lambda_i} \right) = \mu_{ij}\bar{f}_{ij}^* + \epsilon P_{ij} \sum_{k=1}^r \pi_{ki}\mu_{ki}$$

$$= \mu_{ij}\bar{f}_{ij}^* + \epsilon\mu_{ij}\pi_{ij}$$

$$= \mu_{ij}\tilde{f}_{ij}, \quad 1 \le i, j \le r,$$

meaning $(\tilde{e}, \tilde{f}, \tilde{a})$ satisfies (10a). Next, we see that

$$\mu_{ij}\tilde{e}_{ij} = \mu_{ij}\bar{e}_{ij}^* \le \sum_{k=1}^{r} \mu_{ki}\bar{f}_{ki}^* \le \sum_{k=1}^{r} \mu_{ki}\tilde{f}_{ki}, \quad 1 \le i \ne j \le r,$$

where the first inequality follows because $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$ satisfies (10b). Thus, $(\tilde{e}, \tilde{f}, \tilde{a})$ satisfies (10b).

Next we check constraint (10c). Since $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$ satisfies (10c), we know that

$$\sum_{k=1, k \ne i}^{r} \mu_{ki}\tilde{e}_{ki} = \sum_{k=1, k \ne i}^{r} \mu_{ki}e_{ki}^* \le \lambda_i\bar{a}_i^* \le \lambda_i\tilde{a}_i, \quad 1 \le i \le r,$$

and

$$\begin{aligned}
\lambda_i\tilde{a}_i = \lambda_i\left(\bar{a}_i^* + \epsilon\frac{\sum_{k=1}^{r}\pi_{ki}\mu_{ki}}{\lambda_i}\right) &\le \sum_{k=1, k \ne i}^{r} \mu_{ki}\bar{e}_{ki}^* + \sum_{k=1}^{r} \mu_{ki}\bar{f}_{ki}^* + \epsilon\sum_{k=1}^{r}\pi_{ki}\mu_{ki} \\
&= \sum_{k=1, k \ne i}^{r} \mu_{ki}\bar{e}_{ki}^* + \sum_{k=1}^{r} \mu_{ki}(\bar{f}_{ki}^* + \epsilon\pi_{ki}) \\
&= \sum_{k=1, k \ne i}^{r} \mu_{ki}\tilde{e}_{ki} + \sum_{k=1}^{r} \mu_{ki}\tilde{f}_{ki}, \quad 1 \le i \le r.
\end{aligned}$$

Therefore $(\tilde{e}, \tilde{f}, \tilde{a})$ satisfies (10c). To verify (10d), observe that

$$\begin{aligned}
&\lambda_i\tilde{a}_i + \sum_{j=1, j \ne i}^{r} \mu_{ij}\tilde{e}_{ij} - \sum_{k=1, k \ne i}^{r} \mu_{ki}\tilde{e}_{ki} - \sum_{k=1}^{r} \mu_{ki}\tilde{f}_{ki} \\
=&\lambda_i\bar{a}_i^* + \sum_{j=1, j \ne i}^{r} \mu_{ij}\bar{e}_{ij}^* - \sum_{k=1, k \ne i}^{r} \mu_{ki}\bar{e}_{ki}^* - \sum_{k=1}^{r} \mu_{ki}\bar{f}_{ki}^* + \epsilon\sum_{k=1}^{r}\pi_{ki}\mu_{ki} - \epsilon\sum_{k=1}^{r}\mu_{ki}\pi_{ki} \\
=&0,
\end{aligned}$$

and therefore (10d) holds under $(\tilde{e}, \tilde{f}, \tilde{a})$. Lastly, ϵ can be chosen small enough so that both (10e) and (10f) hold (the former also uses the fact that $\sum_{ij}\pi_{ij} = 1$). We conclude that $(\tilde{e}, \tilde{f}, \tilde{a})$ is a feasible solution to (14)–(15) that is better than $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$, which contradicts the fact that $(\bar{e}^*, \bar{f}^*, \bar{a}^*)$ is an optimal solution.

It remains to verify that we can choose non-negative π_{ij}'s to satisfy $\sum_{ij}\pi_{ij} = 1$ and (EC.67). Consider a CTMC defined on the space $\{1, \cdots, r\}^2$. For all $1 \le i, j, k \le r$, the transition rate from

(k, i) to (i, k) is $\mu_{ki} P_{ik}$. No other transitions are possible. Since the CTMC is defined on a finite state space, it has a stationary distribution. Furthermore, any stationary distribution ν must satisfy the flow-balance equations

$$\sum_{k=1}^{r} \nu_{ki} \mu_{ki} P_{ij} = \sum_{\ell=1}^{r} \nu_{ij} \mu_{ij} P_{j\ell}, \quad 1 \le i, j \le r,$$

or

$$P_{ij} \sum_{k=1}^{r} \nu_{ki} \mu_{ki} = \nu_{ij} \mu_{ij}, \quad 1 \le i, j \le r,$$

which are precisely the same as (EC.67). Therefore we can take the π_{ij}'s to be any of the stationary distributions of such a CTMC.

In the case when there exists i' such that $\bar{a}_{i'}^* = 1$, the claim of the lemma is straightforward to verify.

EC.5.3. Proof of Theorem 1

The proof of (5)–(8) hinges on combining Theorems 3 and 4 with Theorem 5.1 of Anantharam and Benchekroun (1993). Below, we repeat the argument from Anantharam and Benchekroun (1993) for completeness. We know that the sequence $\{(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty))\}_N$ is tight, because the support of $(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty))$ is the compact set $\mathcal{T}$. It follows by Prohorov's Theorem Billingsley (1999) that the sequence is also relatively compact. We will now show that any subsequence of $\{(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty))\}_N$ has a further subsequence that converges weakly to a probability measure that assigns a mass of one to the equilibrium set $\mathcal{E}$, thereby proving (5)–(8).

Fix $N > 0$ and initialize $(\bar{E}^{(N)}(0), \bar{F}^{(N)}(0))$ according to $(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty))$. Prohorov's Theorem implies that for any subsequence

$$\{(\bar{E}^{(N')}(0), \bar{F}^{(N')}(0))\}_{N'} \subset \{(\bar{E}^{(N)}(0), \bar{F}^{(N)}(0))\}_N,$$

there exists a further subsequence

$$\{(\bar{E}^{(N'')}(0), \bar{F}^{(N'')}(0))\}_{N''} \subset \{(\bar{E}^{(N')}(0), \bar{F}^{(N')}(0))\}_{N'}$$

that converges weakly to some probability measure (e_0, f_0) with support in $\mathcal{T}$. Now for any $t \geq 0$,

$$(\bar{E}^{(N'')}(0), \bar{F}^{(N'')}(0)) \overset{d}{=} (\bar{E}^{(N'')}(t), \bar{F}^{(N'')}(t)) \Rightarrow (e(t), f(t)),$$

as $N'' \to \infty$, where $(e(t), f(t))$ is the fluid model with intial condition $(e(0), f(0)) = (e_0, f_0)$, and the weak convergence follows from Theorem 3. Since $(e(t), f(t))$ converges to the set $\mathcal{E}$ as $t \to \infty$, it must be the case that $(e_0, f_0) \in \mathcal{E}$ with probability one. This proves (5)–(8).

To prove (9), we need to use the generator of $(\bar{E}^{(N)}, \bar{F}^{(N)})$, which we call $G^{(N)}$. Since $(\bar{E}^{(N)}, \bar{F}^{(N)})$ takes values in a bounded set, Proposition 3 of Glynn and Zeevi (2008) tells us that any function $g : \mathcal{T} \to \mathbb{R}$ satisfies

$$\mathbb{E}\big[G^{(N)}g\big(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty)\big)\big] = 0. \tag{EC.69}$$

In particular, fix i, j between $1, \ldots, r$ and choose $g(e, f) = f_{ij}$. Then

$$G^{(N)}g(e, f) = N\lambda_i P_{ij} 1(e_{ii} > 0)\big((f_{ij} + 1/N) - f_{ij}\big) + \mu_{ij} N f_{ij}\big((f_{ij} - 1/N) - f_{ij}\big),$$

which implies

$$\mathbb{E}\big[G^{(N)}g\big(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty)\big)\big] = \lambda_i P_{ij} \mathbb{P}(\bar{E}_{ii}^{(N)}(\infty) > 0) - \mu_{ij} \mathbb{E}\bar{F}_{ij}^{(N)}(\infty) = 0. \tag{EC.70}$$

Hence,

$$\lim_{N \to \infty} \mathbb{P}(\bar{E}_{ii}^{(N)}(\infty) > 0) = \lim_{N \to \infty} \frac{\mu_{ij}}{\lambda_i P_{ij}} \mathbb{E}\bar{F}_{ij}^{(N)}(\infty) = \frac{\mu_{ij}}{\lambda_i P_{ij}} \bar{f}_{ij} = \bar{a}_i,$$

where in the second equality we used (5) and the fact that $\bar{F}_{ij}^{(N)}(\infty) \in [0, 1]$ to conclude that the sequence of expected values $\mathbb{E}[\bar{F}_{ij}^{(N)}(\infty)]$ converges to $\bar{f}_{ij}$, and in the last equality we used (30a).

EC.5.4. Proof of Theorem 2

Proof of Theorem 2. We will show that the performance measures $\big(\mathbb{E}[\bar{E}^{(N)}(\infty)], \mathbb{E}[\bar{F}^{(N)}(\infty)], A^{(N)}(\infty)\big)$ are a feasible solution to the optimization problem (14)–(15). Lemma 2 then implies part (a) of the theorem is satisfied but only with a non-strict inequality.

To show the inequality is strict, note that the lemma also tells us that the optimal solution to (14)–(15) can never be achieved by performance measures of a CTMC with finitely many cars under any routing policy. This is because the CTMC is positive recurrent and has finitely many states, and so its stationary distribution assigns positive mass to each state. It follows that $A_i^{(N)}(\infty) = \mathbb{P}(\bar{E}_{ii}^{(N)}(\infty) > 0) < 1$ and $\mathbb{E}[\bar{E}_{ii}^{(N)}(\infty)] > 0$ for all regions i. The first claim in Lemma 2 then prevents any performance measures from being the optimal solution, which proves the strict inequality.

Part (b) of the theorem is an immediate consequence of Theorem 1 by setting $Q = q^*$. Recall from (EC.69) that any function $g : \mathcal{T} \to \mathbb{R}$ satisfies

$$\mathbb{E}\big[G^{(N)} g\big(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty)\big)\big] = 0. \tag{EC.71}$$

We now show that $\big(\mathbb{E}[\bar{E}^{(N)}(\infty)], \mathbb{E}[\bar{F}^{(N)}(\infty)], A^{(N)}(\infty)\big)$ satisfies (10a)–(10f).

- Condition (10a) was already verified in (EC.70).

- To check condition (10b), we fix $i \neq j$, and use the test function $g(e, f) = e_{ij}$. Then

$$G^{(N)} g(e, f) = Q_{ij}(e, f) \sum_{k=1}^{r} \mu_{ki} N f_{ki}\big((e_{ij} + 1/N) - e_{ij}\big) + \mu_{ij} N e_{ij}\big((e_{ij} - 1/N) - e_{ij}\big),$$

where $Q_{ij}(e, f)$ is the probability that upon dropping a passenger off at region i, a car drives empty to region j given the current state of the system is (e, f). Using (EC.71) and the fact that $Q_{ij}(e, f) \in [0, 1]$, we see that

$$\begin{aligned}
0 &= \mathbb{E}\left[Q_{ij}\big(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty)\big) \sum_{k=1}^{r} \mu_{ki} \bar{F}_{ki}^{(N)}(\infty) - \mu_{ij} \bar{E}_{ij}^{(N)}(\infty)\right] \tag{EC.72} \\
&\leq \sum_{k=1}^{r} \mu_{ki} \mathbb{E}[\bar{F}_{ki}^{(N)}(\infty)] - \mu_{ij} \mathbb{E}[\bar{E}_{ij}^{(N)}(\infty)].
\end{aligned}$$

- For condition (10c), we fix i and use the test function $g(e, f) = e_{ii}$. Then

$$\begin{aligned}
G^{(N)} g(e, f) &= N \lambda_i 1(e_{ii} > 0)\big((e_{ii} - 1/N) - e_{ii}\big) \\
&\quad + \left(Q_{ii}(e, f) \sum_{k=1}^{r} \mu_{ki} N f_{ki} + \sum_{k=1, k \neq i}^{r} \mu_{ki} N e_{ki}\right)\big((e_{ii} + 1/N) - e_{ii}\big).
\end{aligned}$$

Taking the expected value and using (EC.71), we see that

$$\lambda_i \mathbb{P}(\bar{E}_{ii}^N(\infty) > 0)$$
$$= \mathbb{E}\left[Q_{ii}(\bar{E}^{(N)}(\infty), \bar{F}^{(N)}(\infty)) \sum_{k=1}^r \mu_{ki} \bar{F}_{ki}^{(N)}(\infty) + \sum_{k=1, k\neq i}^r \mu_{ki} \bar{E}_{ki}^{(N)}(\infty)\right] \tag{EC.73}$$

Using the fact that $Q_{ii}(e, f) \in [0, 1]$, we conclude that

$$0 \le \sum_{k=1}^r \mu_{ki} \mathbb{E}[\bar{F}_{ki}^{(N)}(\infty)] + \sum_{k=1, k\neq i}^r \mu_{ki} \mathbb{E}[\bar{E}_{ki}^{(N)}(\infty)] - \lambda_i \mathbb{P}(\bar{E}_{ii}^{(N)}(\infty) > 0),$$

and

$$0 \ge \sum_{k=1, k\neq i}^r \mu_{ki} \mathbb{E}[\bar{E}_{ki}^{(N)}(\infty)] - \lambda_i \mathbb{P}(\bar{E}_{ii}^{(N)}(\infty) > 0).$$

- Fix i. To check condition (10d), we could use the test function $g(e, f) = \sum_{j=1}^r e_{ij}$. Alternatively, it is easier to just add up (EC.72) for all $j \neq i$ together with (EC.73) to arrive at

$$\lambda_i \mathbb{P}(\bar{E}_{ii}^{(N)}(\infty) > 0) + \sum_{j=1, j\neq i}^r \mu_{ij} \mathbb{E}[\bar{E}_{ij}^{(N)}(\infty)]$$
$$= \sum_{k=1}^r \mu_{ki} \mathbb{E}[\bar{F}_{ki}^{(N)}(\infty)] + \sum_{k=1, k\neq i}^r \mu_{ki} \mathbb{E}[\bar{E}_{ki}^{(N)}(\infty)].$$

- Conditions (10e) and (10f) hold trivially.

EC.5.5. Proof of Lemma 3

Proof of Lemma 3. Substituting (30a) and (30b) into (30c), we obtain

$$\lambda_i \bar{a}_i = \sum_{\substack{\ell=1 \\ \ell \neq i}}^r Q_{\ell i} \sum_{k=1}^r \mu_{k\ell} \bar{f}_{k\ell} + Q_{ii} \sum_{k=1}^r \mu_{ki} \bar{f}_{ki} = \sum_{\ell=1}^r Q_{\ell i} \sum_{k=1}^r \lambda_k P_{k\ell} \bar{a}_k$$
$$= \sum_{k=1}^r \lambda_k \bar{a}_k \sum_{\ell=1}^r P_{k\ell} Q_{\ell i}, \quad 1 \le i \le r.$$

In matrix form, these equations can be written as

$$(I - B)\Lambda \bar{a} = 0, \tag{EC.74}$$

where I is the $r \times r$ identity matrix, and B and Λ are $r \times r$ matrices defined as

$$B_{ij} = \sum_{\ell=1}^r P_{j\ell} Q_{\ell i}, \quad \text{and} \quad \Lambda = \text{diag}(\lambda). \tag{EC.75}$$

Observe that B is a column stochastic matrix, i.e. columns sum to one. We now argue that B is irreducible because the CTMC $\left(E^{(N)}, F^{(N)}\right)$ is. For any $1 \leq i, j \leq r$, the entry B_{ij} is the probability that a car picks up a passenger at region i, drives him to some region k, and then drives empty to region j to wait for a new passenger there (or stay and wait at region j if $k = j$). Therefore, B can be interpreted as the transition probability matrix of a discrete-time Markov chain (DTMC) that describes the motion of a single car in a network, i.e. how it serves passengers and makes routing decisions, as if travel times were zero. Irreducibility of B then means that starting from any region the car in the DTMC can visit any other region, which is clearly satisfied when $P_{ij} > 0$ and $Q_{ii} > 0$ for all $i, j = 1, \ldots, r$.

Since B is column stochastic and irreducible, (EC.74) has a unique solution up to a multiplicative constant. That is, any solution to (EC.74) must be of the form $ca^* \geq 0$, where $c > 0$, and $a^* \geq 0$ is a unique vector in $\mathbb{R}^r_+$. We now argue that (30d) and (30e) uniquely define c. First, we use (30e) and (30a)–(30b) to write

$$
\begin{aligned}
1 &= \sum_{i=1}^{r}\sum_{j=1}^{r} \bar{f}_{ij} + \sum_{i=1}^{r}\sum_{\substack{j=1 \\ j \neq i}}^{r} \bar{e}_{ij} + \sum_{i=1}^{r} \bar{e}_{ii} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} \frac{\lambda_i P_{ij}}{\mu_{ij}} \bar{a}_i + \sum_{i=1}^{r}\sum_{\substack{j=1 \\ j \neq i}}^{r} \frac{Q_{ij}}{\mu_{ij}} \sum_{k=1}^{r} \mu_{ki} \bar{f}_{ki} + \sum_{i=1}^{r} \bar{e}_{ii} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} \frac{\lambda_i P_{ij}}{\mu_{ij}} \bar{a}_i + \sum_{i=1}^{r}\sum_{\substack{j=1 \\ j \neq i}}^{r} \frac{Q_{ij}}{\mu_{ij}} \sum_{k=1}^{r} \lambda_k P_{ki} \bar{a}_k + \sum_{i=1}^{r} \bar{e}_{ii}.
\end{aligned}
$$

The equation above can be written as

$$
1 = \sum_{i=1}^{r} \tilde{c}_i \bar{a}_i + \sum_{i=1}^{r} \bar{e}_{ii} = c \sum_{i=1}^{r} \tilde{c}_i a_i^* + \sum_{i=1}^{r} \bar{e}_{ii}
$$

where $\tilde{c}_1, \ldots, \tilde{c}_r > 0$, and in the second equality we used $\bar{a} = ca^*$. Now if

$$
\frac{1}{\sum_{i=1}^{r} \tilde{c}_i a_i^*} a_i^* \leq 1, \quad 1 \leq i \leq r, \tag{EC.76}
$$

then $c = 1/\sum_{i=1}^{r} \tilde{c}_i a_i^*$ and $\bar{e}_{ii} = 0$ for all i are the unique choices under which both (EC.76) and (30d) hold, and we are done.

Now suppose (EC.76) is violated. We cannot choose $c = 1/\sum_{i=1}^{r} \tilde{c}_i a_i^*$, because that would violate the requirement that $\bar{a} \in [0,1]^r$. Instead, choose

$$c = \sup\{s > 0 : sa^* \in [0,1]^r\}, \tag{EC.77}$$

and observe that such a choice necessarily satisfies $c < 1/\sum_{i=1}^{r} \tilde{c}_i a_i^*$. To satisfy (30d), we must set $\bar{e}_{ii} = 0$ for all i such that $\bar{a}_i < 1$. The only restriction on $\bar{e}_{ii}$ for i such that $\bar{a}_i = 1$ is that

$$\sum_{i:\bar{a}_i=1}^{r} \bar{e}_{ii} = 1 - c\sum_{i=1}^{r} \tilde{c}_i a_i^*.$$

It remains to verify that the only viable choice of c is given by (EC.77). Choosing $c < \sup\{s > 0 : sa^* \in [0,1]^r\}$ implies $\bar{a} \in (0,1)^r$, and consequently, (30d) implies that $\bar{e}_{ii} = 0$ for all i. In such a case,

$$c\sum_{i=1}^{r} \tilde{c}_i a_i^* + \sum_{i=1}^{r} \bar{e}_{ii} = c\sum_{i=1}^{r} \tilde{c}_i a_i^* < 1,$$

because $\sup\{s > 0 : sa^* \in [0,1]^r\} < 1/\sum_{i=1}^{r} \tilde{c}_i a_i^*$, and a contradiction is reached. Choosing $c > \sup\{s > 0 : sa^* \in [0,1]^r\}$ forces some element of $\bar{a}$ to be greater than one, which violated the requirement that $\bar{a} \in [0,1]^r$. This concludes the proof.

Acknowledgments

References

Adelman D (2007) Price-directed control of a closed logistics queueing network. *Operations Research* 55(6):1022–1038, URL http://dx.doi.org/10.1287/opre.1070.0408.

Anantharam V, Benchekroun M (1993) A technique for computing sojourn times in large networks of interacting queues. *Probability in the Engineering and Informational Sciences* 7:441–464, ISSN 1469-8951, URL http://dx.doi.org/10.1017/S0269964800003065.

Anselmi J, D'Auria B, Walton N (2013) Closed queueing networks under congestion: nonbottleneck independence and bottleneck convergence. *Mathematics of Operations Research* 38(3):469–491, URL http://dx.doi.org/10.1287/moor.1120.0583.

Asmussen S (2003) *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)* (New York: Springer-Verlag), second edition, ISBN 0-387-00211-1, stochastic Modelling and Applied Probability.

Banerjee S, Freund D, Lykouris T (2016) Multi-objective pricing for shared vehicle systems, URL http://arxiv.org/abs/1608.06819v1, preprint.

Banerjee S, Freund D, Lykouris T (2017) Pricing and optimization in shared vehicle systems: An approximation framework, URL http://arxiv.org/abs/1608.06819v3, preprint.

Baskett F, Chandy KM, Muntz RR, Palacios FG (1975) Open, closed and mixed networks of queues with different classes of customers. *Journal of the Association for Computing Machinery* 22:248–260, URL http://dl.acm.org.proxy.library.cornell.edu/citation.cfm?id=321887.

Billingsley P (1999) *Convergence of probability measures* (New York: Wiley), second edition.

Bimpikis K, Candogan O, Daniela S (2016) Spatial pricing in ride-sharing networks. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2868080, submitted for publication.

Chafkin M (2016) Ubers first self-driving fleet arrives in pittsburgh this month. URL https://www.bloomberg.com/news/features/2016-08-18/uber-s-first-self-driving-fleet-arrives-in-pittsburgh-this-month-is06r7on.

Chemla D, Meunier F, Calvo RW (2013) Bike sharing systems: Solving the static rebalancing problem. *Discrete Optimization* 10(2):120 – 146, ISSN 1572-5286, URL http://dx.doi.org/http://dx.doi.org/10.1016/j.disopt.2012.11.005.

Chen H, Mandelbaum A (1991) Discrete flow networks: bottlenecks analysis and fluid approximations. *Mathematics of Operations Research* 16:408–446.

Czyzyk J, Mesnier MP, Moré JJ (1998) The NEOS server. *IEEE Comput. Sci. Eng.* 5(3):68–75, ISSN 1070-9924, URL http://dx.doi.org/10.1109/99.714603.

Dai JG, He S, Tezcan T (2010) Many-server diffusion limits for $G/Ph/n+GI$ queues. *Annals of Applied Probability* 20(5):1854–1890.

DRI (2016) Didi Research Institute website. http://research.xiaojukeji.com/index_en.html, accessed: 2016-06-30.

George DK (2012) *Stochastic modeling and decentralized control policies for large-scale vehicle sharing systems via closed queueing networks.* Ph.D. thesis, Industrial and Systems Engineering, Ohio State University, URL https://etd.ohiolink.edu/pg_10.

George DK, Xia CH (2011) Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research* 211(1):198 – 207, ISSN 0377-2217, URL http://www.sciencedirect.com/science/article/pii/S0377221710008817.

Glynn PW, Zeevi A (2008) Bounding stationary expectations of Markov processes. *Markov processes and related topics: a Festschrift for Thomas G. Kurtz*, volume 4 of *Inst. Math. Stat. Collect.*, 195–214 (Inst. Math. Statist., Beachwood, OH), URL http://dx.doi.org/10.1214/074921708000000381.

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* 16(1):13–39.

Harrison JM, Reiman MI (1981) Reflected Brownian motion on an orthant. *Annals of Probability* 9:302–308.

Henderson SG, O'Mahony E, Shmoys DB (2016) (Citi)Bike sharing, submitted for publication.

Iglesias R, Rossi F, Zhang R, Pavone M (2016) A BCMP network approach to modeling and controlling autonomous mobility-on-demand systems, URL http://arxiv.org/abs/1607.04357, preprint.

Khalil H (2002) *Nonlinear Systems.* Pearson Education (Prentice Hall), 3rd edition, ISBN 9780130673893.

Krichagina EV (1992) Asymptotic analysis of queueing networks. *Stochastics and Stochastic Reports* 40(1-2):43–76, URL http://dx.doi.org/10.1080/17442509208833781.

Ma S, Zheng Y, Wolfson O (2013) T-share: A large-scale dynamic taxi ridesharing service. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 410–421, ISSN 1063-6382, URL http://dx.doi.org/10.1109/ICDE.2013.6544843.

Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30:149–201.

Ozkan E, Ward AR (2016) Dynamic matching for real-time ridesharing. URL https://ssrn.com/abstract=2844451.

Pavone M, Smith SL, Frazzoli E, Rus D (2012) Robotic load balancing for mobility-on-demand systems. *The International Journal of Robotics Research* 31(7):839–854, URL http://dx.doi.org/10.1177/0278364912444766.

Reiman MI (1984) Open queueing networks in heavy traffic. *Mathematics of Operations Research* 9:441–458, URL http://dx.doi.org/10.1287/moor.9.3.441.

Reiser M, Lavenberg SS (1980) Mean-value analysis of closed multichain queuing networks. *J. ACM* 27(2):313–322, ISSN 0004-5411, URL http://dx.doi.org/10.1145/322186.322195.

Santos DO, Xavier EC (2015) Taxi and ride sharing: A dynamic dial-a-ride problem with money as an incentive. *Expert Systems with Applications* 42(19):6728 – 6737, ISSN 0957-4174, URL http://dx.doi.org/http://dx.doi.org/10.1016/j.eswa.2015.04.060.

Suri R, Sahu S (2007) Approximate mean value analysis for closed queuing networks with multiple-server stations. *IIE Annual Conference. Proceedings*, 1618 (Institute of Industrial Engineers-Publisher).

Waserhole A, Jost V (2013) Vehicle Sharing System Pricing Regulation: A Fluid Approximation, URL https://hal.archives-ouvertes.fr/hal-00727041, working paper or preprint.

Waserhole A, Jost V (2016) Pricing in vehicle sharing systems: optimization in queuing networks with product forms. *EURO Journal on Transportation and Logistics* 5(3):293–320, ISSN 2192-4384, URL http://dx.doi.org/10.1007/s13676-014-0054-4.

Yang P, Iyer K, Frazier PI (2016) Mean field equilibria for competitive exploration in resource sharing settings. *Proceedings of the 25th International Conference on World Wide Web*, 177–187, WWW '16 (Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee), ISBN 978-1-4503-4143-1, URL http://dx.doi.org/10.1145/2872427.2883011.

Zhang R, Pavone M (2016) Control of robotic mobility-on-demand systems. *Int. J. Rob. Res.* 35(1-3):186–203, ISSN 0278-3649, URL http://dx.doi.org/10.1177/0278364915581863.