INVITED PAPER

The Price of Anarchy in Transportation Networks: Data-Driven Evaluation and Reduction Strategies

This paper studies transportation networks under two different routing policies, the selfish user-centric routing one and the socially optimal system-centric one, and proposes an index, the Price of Anarchy (PoA), to increase efficiency.

By Jing Zhang, *Student Member IEEE*, Sepideh Pourazarm, *Student Member IEEE*, Christos G. Cassandras, *Fellow IEEE*, and Ioannis Ch. Paschalidis, *Fellow IEEE*

ABSTRACT | Among the many functions a smart city must support, transportation dominates in terms of resource consumption, strain on the environment, and frustration of its citizens. We study transportation networks under two different routing policies, the commonly assumed selfish user-centric routing policy and a socially optimal system-centric one. We consider a performance metric of efficiency—the Price of Anarchy (PoA)—defined as the ratio of the total travel latency cost under selfish routing over the corresponding quantity under socially optimal routing. We develop a data-driven approach to estimate the PoA, which we subsequently use to conduct a case study using extensive actual traffic data from the Eastern Massachusetts road network. To estimate the PoA, our approach learns from data a complete model of the transportation network, including origin–destination demand and user preferences. We leverage this model to propose possible strategies to reduce the PoA and increase efficiency.

KEYWORDS | Optimization; Price-of-Anarchy (PoA); sensitivity analysis; traffic assignment problem; transportation networks; variational inequalities

I. INTRODUCTION

As of 2014, 54% of the earth's population resides in urban areas, a percentage expected to reach 66% by 2050. This increase would amount to 2.5 billion people added to urban populations [1]. At the same time, there are now 28 megacities (with population ≥10 million) worldwide, accounting for 22% of the world's urban dwellers, and projections indicate more than 41 megacities by 2030. It stands to reason that the management and sustainability of urban areas has become one of the most critical challenges our societies face today, leading to a quest for "smart" cities.

Among the many functions a city supports, transportation dominates in terms of resource consumption, strain on the environment, and frustration of its citizens. Commuter delays have risen by 260% over the past 25 years and 28% of U.S. primary energy is now used in transportation [2]. It is estimated that the cumulative cost of traffic congestion by 2030 will reach \$2.8 trillion [3]—equal roughly to the U.S. annual tax revenue. This estimate accounts for direct costs to drivers (time, fuel) and indirect costs resulting from businesses passing these same costs on to consumers, but it does not include the equally alarming environmental impact due to a large proportion of toxic air pollutants attributed to mobile sources. At the individual citizen level, traffic congestion led to \$1740 in average costs per driver during 2014. If unchecked, this number is expected to grow by more than 60%, to \$2900 annually, by 2030 [3].

A transportation network is a system with noncooperative agents (drivers) in which each agent seeks to minimize

Manuscript received June 22, 2017; revised January 2, 2018; accepted January 3, 2018. Date of publication February 5, 2018; date of current version March 26, 2018. This work was supported in part by the National Science Foundation (NSF) under Grants CNS-1645681, ECCS-1509084, CCF-1527292, IIS-1237022, and IIP-1430145; by the U.S. Department of Energy (DOE) under Grant DE-AR0000796; by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-15-1-0471; by the Army Research Office (ARO) under Grant W911NF-12-1-0390; by Bosch; and by The MathWorks. *(Corresponding author: Ioannis Ch. Paschalidis.)*

J. Zhang and **S. Pourazarm** are with the Division of Systems Engineering, Boston University, Boston, MA 02446 USA (e-mail: jzh@bu.edu; sepid@bu.edu).

C. G. Cassandras and **I. Ch. Paschalidis** are with the Division of Systems Engineering and the Department of Electrical and Computer Engineering, Boston University, Boston, MA 02215 USA (e-mail: cgc@bu.edu; yannisp@bu.edu).

Digital Object Identifier: 10.1109/JPROC.2018.2790405

0018-9219 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

her own individual cost by choosing the best route (resources) to reach her destination without taking into account the overall system performance. In these systems, the cost for each agent depends on the resources chosen as well as the number of agents choosing the same resources. This results in a Nash equilibrium, i.e., a point where no agent can benefit by altering its actions, assuming that the actions of all the other agents remain fixed [4]. However, it is known that the user optimal policy leading to a Nash equilibrium is generally inefficient and results in a suboptimal behavior compared to the socially optimal policy that could be attained through a centrally controlled system [4]. In order to quantify this inefficiency due to selfish driving, we define the Price of Anarchy (PoA) as the ratio of the total travel latency cost under the user optimal (user-centric) routing policy versus the socially optimal (system-centric) one. The PoA is, therefore, a measure of the efficiency achieved by any transportation network as it currently operates.

The first issue addressed in the paper is how to measure the PoA from data. The user flow equilibrium in a transportation network is known as a Wardrop equilibrium [5] (an instantiation of the generic Nash equilibrium). It is the solution of the traffic assignment problem (TAP)[6], which we call the user-centric forward problem. To solve this TAP, we need to know *a priori*: 1) the specific travel latency cost functions involved [7]; and 2) the traffic demand expressed through an origin-destination (OD) demand matrix [6]. Starting from the equilibrium link flows (assuming they can be inferred or directly observed), we first estimate an initial OD demand matrix. We note that the OD demand estimation problem has been widely studied; see, e.g., [8], [9], and the references therein. Then, based on inverse optimization techniques recently developed in [10], we propose a novel user-centric inverse problem formulation. Specifically, given observed link flow data (Wardrop equilibrium), we estimate the associated travel latency cost functions. In other words, we seek cost functions which, when applied to the TAP, would yield the link flows that are actually observed. Once this is accomplished, based on a bilevel optimization problem formulation considered in [11] and [12], we develop an algorithmic procedure for iteratively adjusting the values of the OD demands so that the observed link flows are as close as possible to the solution of the user-centric forward problem (i.e., TAP). The OD demand and the user travel latency cost functions completely parametrize a predictive model of the transportation network. We use this model to calculate the total travel latency cost under the user optimal routing policy, thus obtaining the numerator of the PoA ratio.

Next, using the same predictive model, we formulate a system-centric forward problem [6], [13], a nonlinear program (NLP), in which all agents (drivers) cooperate to optimize the overall system performance. Its solution enables us to calculate the total travel latency cost under the socially optimal routing policy, i.e., the denominator of the PoA ratio. Thus, the combination of the inverse and forward optimization problems results in measuring the PoA for a given transportation network whose equilibrium link flows are observed based on collected traffic data.

Having an accurate predictive model allows us to go beyond estimation (of the PoA) and consider specific control actions that could reduce the PoA. To that end, we analyze the sensitivity of the optimal objective function value of an optimization problem formulation for the TAP with respect to key parameters, such as road capacities and free-flow travel times. The results can help prioritize road segments for interventions that can mitigate congestion. We derive sensitivity analysis formulas and propose their finite difference approximations.

As an illustration of our data-driven approach outlined above, we use actual traffic data from the Eastern Massachusetts (EMA) transportation network, in the form of spatial average speeds and road segment flow capacities. These data were provided to us by the Boston Region Metropolitan Planning Organization (MPO) and include average speeds over 13 000 road segments at every minute of 2012. By using a traffic flow model, we first infer equilibrium flows on each road segment and then apply our approach to evaluate the PoA for two highway subnetworks of the EMA network. In addition, we derive sensitivity analysis results and conduct a meta-analysis comparing the user-centric and socially optimal routing policies.

As a final step, we propose strategies for reducing the PoA. First, by taking advantage of the rapid emergence of connected automated vehicles (CAVs) [14]–[17], it has become feasible to automate routing decisions, thus solving a system-centric forward problem in which all CAVs (bypassing driver decisions) cooperate to optimize the overall system performance. Second, we propose a modification to existing GPS navigation algorithms recommending to all drivers socially optimal routes. Finally, our sensitivity analysis results provide the means to prioritize road segments for specific interventions that can mitigate congestion.

The rest of the paper is organized as follows. We review the related literature in Section II. In Section III, we introduce models and methods we use. In Section IV, we describe the data sets and explain the data processing procedures for a case study of the EMA network. Numerical results for the case study are shown in Section V. In Section VI, we propose possible strategies to reduce the PoA. We provide concluding remarks and point out some directions for future research in Section VII.

Notation: All vectors are column vectors. For economy of space, we write $\mathbf{x} = (x_1, \ldots, x_{\dim(\mathbf{x})})$ to denote the column vector $\mathbf{x}$, where $\dim(\mathbf{x})$ is the dimension of $\mathbf{x}$. We use $\mathbf{0}$ and $\mathbf{1}$ for the vectors with all entries equal to zero and one, respectively. We denote by $\mathbb{R}_+$ the set of all nonnegative real numbers. $\mathbf{M} \geq \mathbf{0}$ (resp., $\mathbf{x} \geq \mathbf{0}$) indicates that all entries of a matrix $\mathbf{M}$ (resp., vector $\mathbf{x}$) are nonnegative. We use "prime" to denote the transpose of a matrix or vector. Unless otherwise specified, $\|\cdot\|$ denotes the ℓ_2-norm. We let $|\mathcal{D}|$ denote the cardinality of a set $\mathcal{D}$, and $[\![\mathcal{D}]\!]$ the set $\{1, \ldots, |\mathcal{D}|\}$.

II. RELATED WORK

The classical static TAP [6], i.e., the user-centric forward problem in our terminology, has been widely studied; see, e.g., [18] and [19] for the single-class (i.e., all vehicles are modeled as belonging to the same class) transportation networks and [20]–[22] for the multiclass (i.e., different types of vehicles, such as cars or trucks, are modeled as belonging to different classes) transportation networks. The static TAP has also been generalized to the case that has a dynamic network equilibrium modeling capability; see, e.g., [23] and [24], among others.

Based on road traffic counts within selected time intervals (i.e., road traffic flows), the problem of estimating the OD demand matrix of a given transportation network has been considered in [8], [9], [25], and references therein. In particular, Hazelton [26] proposed a generalized least squares (GLS) method to estimate the OD demand matrices of uncongested networks, and Spiess [11], Lundgren and Peterson [12], and Yang *et al.* [27] considered networks that could include congested roads.

Sensitivity analyses of traffic equilibria were conducted in [23], [28], and [29], among others, by evaluating the directions of change that occur in the link flows with respect to the change of travel costs as parameters in the cost and demand functions.

Preliminary PoA evaluation results of this paper have been presented in two conferences, [30] and [31], where results of a case study for a much smaller subnetwork of EMA were reported and no PoA reduction strategies were proposed. A similar topic was also discussed in [32] and the references therein; in particular, based on real traffic data from the transportation network of Singapore, Monnot *et al.* [32] used a different framework from ours to quantify the PoA.

III. MODELS AND METHODS

A. Model for a Single-Class Transportation Network

We begin by reviewing the model of [30]. Denote a road network by $(\mathcal{V}, \mathcal{A}, \mathcal{W})$, where $(\mathcal{V}, \mathcal{A})$ forms a directed graph with $\mathcal{V}$ being the set of nodes and $\mathcal{A}$ the set of links, and $\mathcal{W} = \{\mathbf{w}_i : \mathbf{w}_i = (w_{si}, w_{ti}), i \in [\![\mathcal{W}]\!]\}$ indicates the set of all OD pairs. Note that only nodes of the road network can be origin/destination of flows; we make this standard modeling assumption to accommodate our graph-based view of the transportation system. Assume the graph $(\mathcal{V}, \mathcal{A})$ is strongly connected and let $\mathbf{N} \in \{0, 1,\ -1\}^{|\mathcal{V}| \times |\mathcal{A}|}$ be its node-link incidence matrix. Denote by $\mathbf{e}_a$ the vector with an entry being 1 corresponding to link a and all the other entries being 0. For any OD pair $\mathbf{w} = (w_s, w_t)$, denote by $d^{\mathbf{w}} \geq 0$ the amount of the flow demand from w_s to w_t. Let $\mathbf{d}^{\mathbf{w}} \in \mathbb{R}^{|\mathcal{V}|}$ be the vector which is all zeros, except for two entries $-d^{\mathbf{w}}$ and $d^{\mathbf{w}}$ corresponding to nodes w_s and w_t, respectively.

Denote by $\mathcal{R}_i$ the set of simple routes (a route without cycles is called a "simple route") for OD pair i. For each $a \in \mathcal{A}$, $i \in [\![\mathcal{W}]\!]$, $r \in \mathcal{R}_i$, define the link-route incidence by

$$\delta^i_{ra} = \begin{cases} 1, & \text{if route } r \in \mathcal{R}_i \text{ uses link } a \\ 0, & \text{otherwise.} \end{cases}$$

Let x_a denote the flow on link $a \in \mathcal{A}$ and $\mathbf{x} = (x_a;\ a \in \mathcal{A})$ the flow vector. Denote by $t_a(\mathbf{x}) : \mathbb{R}_+^{|\mathcal{A}|} \to \mathbb{R}_+$ the travel latency cost (i.e., travel time) function for link $a \in \mathcal{A}$. If for all $a \in \mathcal{A}$, $t_a(\mathbf{x})$ only dwepends on x_a, we say the cost function $\mathbf{t}(\mathbf{x}) = (t_a(x_a);\ a \in \mathcal{A})$ is separable [6]. Throughout the paper, we assume that the travel latency cost functions are separable and take the following form [7], [10]:

$$t_a(x_a) = t_a^0 f\left(\frac{x_a}{m_a}\right) \tag{1}$$

where t_a^0 is the free-flow travel time of $a \in \mathcal{A}$, $f(0) = 1$, $f(\cdot)$ is strictly increasing and continuously differentiable on $\mathbb{R}_+$, and m_a is the flow capacity of $a \in \mathcal{A}$. Note that the flow capacity is not a "hard" constraint; x_a could exceed m_a for various a at the cost of increased travel time.

Define the set of feasible flow vectors $\mathcal{F}$ as [10]

$$\mathcal{F} \overset{\text{def}}{=} \Big\{ \mathbf{x} : \exists \mathbf{x}^{\mathbf{w}} \in \mathbb{R}_+^{|\mathcal{A}|} \text{ s.t. } \mathbf{x} = \sum_{\mathbf{w} \in \mathcal{W}} \mathbf{x}^{\mathbf{w}}, \\ \mathbf{N}\mathbf{x}^{\mathbf{w}} = \mathbf{d}^{\mathbf{w}}, \forall \mathbf{w} \in \mathcal{W} \Big\}$$

where $\mathbf{x}^{\mathbf{w}}$ indicates the flow vector attributed to OD pair $\mathbf{w}$. In order to formulate appropriate forward and inverse optimization problems arising in transportation networks, we next state the definition of Wardrop equilibrium.

Definition 1 [6]: A feasible flow $\mathbf{x}^* \in \mathcal{F}$ is a Wardrop equilibrium if for every OD pair $\mathbf{w} = (w_s, w_t) \in \mathcal{W}$, and any route connecting (w_s, w_t) with positive flow in $\mathbf{x}^*$, the cost of traveling along that route is no greater than the cost of traveling along any other route that connects (w_s, w_t). Here, the cost of traveling along a route is the sum of the costs of each of its constituent links.

B. The User-Centric Forward Problem

As in [30], here we refer to the classical static TAP as the user-centric forward problem, whose goal is to find the Wardrop equilibrium for a given single-class transportation network with a given travel latency cost function and a given OD demand matrix. It is a well-known fact that, for network $(\mathcal{V}, \mathcal{A}, \mathcal{W})$, the TAP can be formulated as the following optimization problem [6], [18]:

$$\text{(userOpt)} \qquad \min_{\mathbf{x} \in \mathcal{F}} \sum_{a \in \mathcal{A}} \int_0^{x_a} t_a(s)\, ds. \tag{2}$$

As an alternative, we also formulate the TAP as a variational inequality (VI) problem.

Definition 2 [10]: The VI problem, denoted as VI$(\mathbf{t}, \mathcal{F})$, is to find an $\mathbf{x}^* \in \mathcal{F}$ subject to

$$\mathbf{t}(\mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*) \geq 0 \qquad \forall \mathbf{x} \in \mathcal{F}. \tag{3}$$

To proceed, let us first recall the definition of the strong monotonicity for a cost function: $\mathbf{t}(\cdot)$ is strongly monotone [6] on $\mathcal{F}$ if there exists a constant $\eta > 0$ such that

$$[\mathbf{t}(\mathbf{x}) - \mathbf{t}(\mathbf{y})]'(\mathbf{x} - \mathbf{y}) \geq \eta \, \|\mathbf{x} - \mathbf{y}\|^2 \qquad \forall \mathbf{x}, \mathbf{y} \in \mathcal{F}. \tag{4}$$

It is known that if t$(\cdot)$ is continuously differentiable on $\mathcal{F}$, then (4) is equivalent to the positive definiteness of the Jacobian of t$(\cdot)$ [6, p. 180]. Note that a strictly increasing $f(\cdot)$ in (1) would not necessarily ensure the strong monotonicity of $\mathbf{t}(\cdot)$; e.g., $f(x) \overset{\text{def}}{=} x^3$ and $\mathbf{t}(\mathbf{x}) \overset{\text{def}}{=} (x_1^3, x_2^3)$ would lead to the Jacobian of $\mathbf{t}(\mathbf{x})$ as

$$\begin{bmatrix} 3x_1^2 & 0 \\ 0 & 3x_2^2 \end{bmatrix}$$

which is not positive definite over $\mathbb{R}^2$. We next introduce a key assumption.

Assumption A: $\mathbf{t}(\cdot)$ is strongly monotone on $\mathcal{F}$ and continuously differentiable on $\mathbb{R}_+^{|\mathcal{A}|}$. $\mathcal{F}$ is nonempty and contains an interior point (Slater's condition [33]).

For the existence and uniqueness of the TAP, the following result is available.

Theorem 1 [6]: Assumption A implies that there exists a Wardrop equilibrium of the network $(\mathcal{V}, \mathcal{A}, \mathcal{W})$, which is the unique solution to VI$(\mathbf{t}, \mathcal{F})$.

C. The User-Centric Inverse Problem

To solve the user-centric forward problem, we need to know the travel latency cost function and the OD demand matrix. Assuming that we know the OD demand matrix and have observed the Wardrop equilibrium link flows, we seek to formulate the user-centric inverse problem (the inverse VI problem, in particular), so as to estimate the travel latency cost function. To provide some insight, given $|\mathcal{K}|$ samples of the link flow vector $\mathbf{x}$, one can think of them as flow observations on $|\mathcal{K}|$ different networks/subnetworks which are nevertheless produced by the exact same cost function. The inverse formulation seeks to determine the cost function so that each flow observation is as close to an equilibrium as possible. Given that the inverse problem will rely on measured flows, we should expect measurement noise which will prevent the flows from being an exact solution of the forward VI problem VI$(\mathbf{t}, \mathcal{F})$. Therefore, we will first define the notion of an approximate solution.

For a given $\epsilon > 0$, we define an ϵ-approximate solution to VI$(\mathbf{t}, \mathcal{F})$ by changing the right-hand side of (3) to $-\epsilon$.

Definition 3 [10]: Given $\epsilon > 0$, $\hat{\mathbf{x}} \in \mathcal{F}$ is called an ϵ-approximate solution to VI$(\mathbf{t}, \mathcal{F})$ if

$$\mathbf{t}(\hat{\mathbf{x}})'(\mathbf{x} - \hat{\mathbf{x}}) \geq -\epsilon \qquad \forall \mathbf{x} \in \mathcal{F}. \tag{5}$$

Assume now we are given $|\mathcal{K}|$ networks $(\mathcal{V}^{(k)}, \mathcal{A}^{(k)}, \mathcal{W}^{(k)})$, $k \in [\![\mathcal{K}]\!]$ [as a special case, these could be $|\mathcal{K}|$ replicas of the same network $(\mathcal{V}, \mathcal{A}, \mathcal{W})$], and the observed link flow data $\{\mathbf{x}^{(k)} = (x_a^{(k)}; a \in \mathcal{A}^{(k)}); k \in [\![\mathcal{K}]\!]\}$, where k is the network index and $x_a^{(k)}$ is the flow for link $a \in \mathcal{A}^{(k)}$ correspondingly. The inverse VI problem amounts to finding a function $\mathbf{t}$ such that $\mathbf{x}^{(k)}$ is an ϵ_k-approximate solution to VI$(\mathbf{t}, \mathcal{F}^{(k)})$ for each k. Denoting $\boldsymbol{\epsilon} = (\epsilon_k; k \in [\![\mathcal{K}]\!])$, we can formulate the inverse VI problem as [10]

$$\min_{t,\epsilon} \; \|\epsilon\| \tag{6}$$

$$\text{s.t. } \mathbf{t}(\mathbf{x}^{(k)})'(\mathbf{x} - \mathbf{x}^{(k)}) \geq -\epsilon_k, \qquad \forall \mathbf{x} \in \mathcal{F}^{(k)}, \qquad k \in [\![\mathcal{K}]\!]$$

$$\epsilon_k > 0, \; \forall k \in [\![\mathcal{K}]\!]$$

where the optimization is over the selection of function $\mathbf{t}$ and the vector $\boldsymbol{\epsilon}$.

Aiming at recovering a cost function t that has both good data reconciling and generalization properties (i.e., t should fit "old" data well but should not be overfitting; it must also have great power to predict "new" data), to make (6) solvable, we apply an estimation approach which expresses the function $f(\cdot)$ [in (1)] in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ [10], [34]. In particular, by [10, Th. 2], we reformulate the inverse VI problem (6) as

$$\text{(invVI-1)} \quad \min_{f, \mathbf{y}, \epsilon} \; \|\epsilon\| + \gamma \|f\|_{\mathcal{H}}^2 \tag{7}$$

$$\text{s.t. } \mathbf{e}'_a \mathbf{N}'_k \mathbf{y}^{\mathbf{w}} \leq t_a^0 f\left(\frac{x_a}{m_a}\right) \tag{8}$$

$$\forall \mathbf{w} \in \mathcal{W}^{(k)}, \; a \in \mathcal{A}^{(k)}, \; k \in [\![\mathcal{K}]\!]$$

$$\sum_{a \in \mathcal{A}^{(k)}} t_a^0 x_a f\left(\frac{x_a}{m_a}\right) - \sum_{\mathbf{w} \in \mathcal{W}^{(k)}} (\mathbf{d}^{\mathbf{w}})' \mathbf{y}^{\mathbf{w}} \leq \epsilon_k, \tag{9}$$

$$\forall k \in [\![\mathcal{K}]\!]$$

$$f\left(\frac{x_a}{m_a}\right) \leq f\left(\frac{x_{\tilde{a}}}{m_{\tilde{a}}}\right) \tag{10}$$

$$\forall a, \; \tilde{a} \in \cup_{k=1}^{|\mathcal{K}|} \; \mathcal{A}^{(k)} \; \text{ s.t. } \frac{x_a}{m_a} \leq \frac{x_{\tilde{a}}}{m_{\tilde{a}}}$$

$$\epsilon \geq \mathbf{0}, \quad f \in \mathcal{H} \tag{11}$$

$$f(0) = 1$$

which is a counterpart of [10, eq. (22)]. Note that $\mathbf{y} = \left(\mathbf{y}^{\mathbf{w}}; \mathbf{w} \in \mathcal{W}^{(k)}, k \in [\![\mathcal{K}]\!]\right)$ and $\boldsymbol{\epsilon} = (\epsilon_k; k \in [\![\mathcal{K}]\!])$ are decision vectors ($\mathbf{y}^{\mathbf{w}}$ is a dual variable which can be interpreted as the "price" of $\mathbf{d}^{\mathbf{w}}$, in particular). Note also that $\gamma > 0$ is a regularization parameter—a smaller γ should result in recovering a "tighter" $f(\cdot)$ in terms of data reconciling; a larger γ, on the other hand, would lead to a "better" $f(\cdot)$ in terms of generalization properties. Moreover, $\|f\|_{\mathcal{H}}^2$ denotes the squared norm of $f(\cdot)$ in $\mathcal{H}$, (8) is for dual feasibility, (9) is the suboptimality (primal-dual gap) constraint, (10) enforces $f(\cdot)$ to be nondecreasing, and (11) is a normalization constraint.

It can be seen that the above formulation is still too abstract for us to solve, because it is an optimization over functions. To make it tractable, in the following, we will specify $\mathcal{H}$ by selecting its reproducing kernel [34] to be a polynomial $\phi(x,y) = (c + xy)^n$ for some choice of $c \geq 0$ and $n \in \mathbb{N}$ (for the specifications of c and n, see [35]). Then, writing

$$\phi(x,y) = (c + xy)^n = \sum_{i=0}^{n} \binom{n}{i} c^{n-i} x^i y^i$$

by [34, eq. (3.2), (3.3), and (3.6)], we instantiate invVI-1 as

$$
\begin{aligned}
&\text{(invVI-2)} \min_{\boldsymbol{\beta},\mathbf{y},\boldsymbol{\epsilon}} \quad \|\boldsymbol{\epsilon}\| + \gamma \sum_{i=0}^{n} \frac{\beta_i^2}{\binom{n}{i} c^{n-i}}\\
&\text{s.t. } \mathbf{e}_a' \mathbf{N}_k' \mathbf{y}^{\mathbf{w}} \leq t_a^0 \sum_{i=0}^{n} \beta_i \left(\frac{x_a}{m_a}\right)^i\\
&\qquad \forall \mathbf{w} \in \mathcal{W}^{(k)},\ a \in \mathcal{A}^{(k)},\ k \in [\![\mathcal{K}]\!]\\
&\sum_{a \in \mathcal{A}_k} t_a^0 x_a \sum_{i=0}^{n} \beta_i \left(\frac{x_a}{m_a}\right)^i - \sum_{\mathbf{w} \in \mathcal{W}_k} (\mathbf{d}^{\mathbf{w}})' \mathbf{y}^{\mathbf{w}} \leq \epsilon_k,\\
&\qquad \forall k \in [\![\mathcal{K}]\!]\\
&\sum_{i=0}^{n} \beta_i \left(\frac{x_a}{m_a}\right)^i \leq \sum_{i=0}^{n} \beta_i \left(\frac{x_{\tilde{a}}}{m_{\tilde{a}}}\right)^i\\
&\qquad \forall a,\ \tilde{a} \in \cup_{k=1}^{|\mathcal{K}|} \mathcal{A}^{(k)} \text{ s.t. } \frac{x_a}{m_a} \leq \frac{x_{\tilde{a}}}{m_{\tilde{a}}}\\
&\boldsymbol{\epsilon} \geq \mathbf{0}, \quad \beta_0 = 1
\end{aligned}
$$

where the function $f(\cdot)$ in invVI-1 is parameterized by $\boldsymbol{\beta} = (\beta_i;\ i = 0,1,\ldots,n)$. Assuming an optimal $\boldsymbol{\beta}^* = (\beta_i^*; i = 0,1,\ldots,n)$ is obtained by solving invVI-2, then our estimator for $f(\cdot)$ is

$$\hat{f}(x) = \sum_{i=0}^{n} \beta_i^* x^i = 1 + \sum_{i=1}^{n} \beta_i^* x^i. \tag{12}$$

D. OD Demand Estimation

Given a network $(\mathcal{V}, \mathcal{A}, \mathcal{W})$, to estimate an initial OD demand matrix, we borrow the GLS method proposed in [26], which assumes that the transportation network $(\mathcal{V}, \mathcal{A}, \mathcal{W})$ is uncongested (in other words, for each OD pair the route choice probabilities are independent of traffic flow), and that the OD trips (traffic counts) are Poisson distributed. Note that such assumptions may be strong and we will relax them when finalizing our OD demand estimator by performing an adjustment procedure.

Denote by $\{\mathbf{x}^{(k)}; k \in [\![\mathcal{K}]\!]\}$ $|\mathcal{K}|$ observations of the flow vector. Let $\bar{\mathbf{x}} = (1/|\mathcal{K}|) \sum_{k=1}^{|\mathcal{K}|} \mathbf{x}^{(k)}$ be the sample mean vector and $\mathbf{S} = (1/(|\mathcal{K}| - 1)) \sum_{k=1}^{|\mathcal{K}|} (\mathbf{x}^{(k)} - \bar{\mathbf{x}})(\mathbf{x}^{(k)} - \bar{\mathbf{x}})'$ the sample covariance matrix. Let $\mathbf{P} = [p_{ir}]$ denote the route choice probability matrix, where p_{ir} is the probability that a traveler associated with OD pair i uses route r. Vectorize the OD demand matrix as $\mathbf{g} = (g_i; i \in [\![\mathcal{W}]\!])$. After finding feasible routes for each OD pair, thus obtaining the link-route incidence matrix $\mathbf{A}$, the GLS method amounts to solving the following optimization problem:

$$
\begin{aligned}
&\text{(P0)} \min_{\mathbf{P}\geq\mathbf{0},\, \mathbf{g}\geq\mathbf{0}} \sum_{k=1}^{|\mathcal{K}|} \left(\mathbf{x}^{(k)} - \mathbf{A}\mathbf{P}'\mathbf{g}\right)' \mathbf{S}^{-1} \left(\mathbf{x}^{(k)} - \mathbf{A}\mathbf{P}'\mathbf{g}\right)\\
&\qquad \text{s.t. } p_{ir} = 0 \quad \forall (i,r) \in \{(i,r) : r \notin \mathcal{R}_i\}\\
&\qquad\qquad \mathbf{P}\mathbf{1} = \mathbf{1}
\end{aligned}
$$

which minimizes a weighted sum of the squared errors in the flow observations. Directly solving (P0) is cumbersome due to the complicated form of the objective function, and we in turn decouple (P0) into two subproblems. To that end, we perform a variable substitution by setting $\boldsymbol{\xi} = \mathbf{P}'\mathbf{g}$ and we let $h(\mathbf{P},\mathbf{g})$ be an arbitrarily selected smooth scalar-valued function. Then, we solve sequentially the following two problems [30]:

$$\text{(P1)} \quad \min_{\boldsymbol{\xi}\geq\mathbf{0}} \quad \frac{|\mathcal{K}|}{2} \boldsymbol{\xi}'\mathbf{Q}\boldsymbol{\xi} - \mathbf{b}'\boldsymbol{\xi} \tag{13}$$

where $\mathbf{Q} = \mathbf{A}'\mathbf{S}^{-1}\mathbf{A}$ and $\mathbf{b} = \sum_{k=1}^{|\mathcal{K}|} \mathbf{A}'\mathbf{S}^{-1}\mathbf{x}^{(k)}$, and

$$
\begin{aligned}
&\text{(P2)} \qquad \min_{\mathbf{P}\geq\mathbf{0},\, \mathbf{g}\geq\mathbf{0}} \quad h(\mathbf{P},\mathbf{g}) &\qquad (14)\\
&\text{s.t.} \quad p_{ir} = 0,\ \forall (i,r) \in \{(i,r) : r \notin \mathcal{R}_i\}\\
&\qquad\quad \mathbf{P}'\mathbf{g} = \boldsymbol{\xi}^0\\
&\qquad\quad \mathbf{P}\mathbf{1} = \mathbf{1},
\end{aligned}
$$

where $\boldsymbol{\xi}^0$ is the optimal solution to (P1). Essentially, (P1) uses the variable substitution to eliminate the constrains on $\mathbf{P}$ and (P2) seeks to find a feasible $\mathbf{P}$ consistent with the optimal solution of (P1) and the relationship $\boldsymbol{\xi} = \mathbf{P}'\mathbf{g}$. We write the feasibility problem (P2) as an optimization problem with some "dummy" cost function because this allows us to use an optimization solver; in fact, we can simply set $h(\mathbf{P},\mathbf{g}) \equiv 0$. Specifically, (P1) [resp., (P2)] is a typical quadratic program (QP) [resp., quadratically constrained program (QCP)]. Letting $(\mathbf{P}^0, \mathbf{g}^0)$ be an optimal solution to (P2), then $\mathbf{g}^0$ is our initial estimate of the demand vector.

Remark 1: It is seen that each entry of $\mathbf{g}^0$ can always be expressed as a sum of certain entries in $\boldsymbol{\xi}^0$; in other words, given $\boldsymbol{\xi}^0 \geq \mathbf{0}$, (P2) always has a feasible solution. Thus, (P0) is actually equivalent to (P1) and (P2), in the sense that if $(\mathbf{P}^0, \mathbf{g}^0)$ is an optimal solution to (P0) [resp., (P2)], then it is also an optimal solution to (P2) [resp., (P0)]. In addition, we note that the GLS method above would encounter numerical difficulties when the network size is large, because there would be too many decision variables. Note also that this method is valid under a "no-congestion" assumption and, to take the congestion on the link flows into account, we in turn consider a bilevel optimization problem in the following.

Assume now the function $f(\cdot)$ in (1) is available. For any given feasible $\mathbf{g}$ $(\geq \mathbf{0})$, let $\mathbf{x}(\mathbf{g})$ be the optimal solution to the TAP (2). In the following, denote by $\tilde{\mathbf{x}} = (\tilde{x}_a;\ a \in \mathcal{A})$ the observed flow vector. Assuming an initial demand vector $\mathbf{g}^0$

is given [the $\mathbf{g}^0$ obtained by solving (P1) and (P2) is a good candidate; we will take it as $\mathbf{g}^0$ hereafter], we consider the following bilevel optimization problem [11], [12]:

$$\text{(BiLev)} \quad \min_{\mathbf{g}\geq\mathbf{0}} F(\mathbf{g}) \overset{\text{def}}{=} \gamma_1 \sum_{i\in[\![\mathcal{W}]\!]} \left(g_i - g_i^0\right)^2 + \gamma_2 \sum_{a\in\mathcal{A}} \left(x_a(\mathbf{g}) - \tilde{x}_a\right)^2 \qquad (15)$$

where $\gamma_1, \gamma_2 \geq 0$ are two weight parameters. The first term penalizes moving too far away from the initial demand, and the second term ensures that the optimal solution to the TAP is close to the flow observation. Note that the BiLev formulation (15) is more general than the one considered in [31], which includes the second term only. It is worth pointing out that $F(\mathbf{g})$ has a lower bound 0 which guarantees the convergence of the algorithm (see Algorithm 1) that we will apply.

Remark 2: From now on, let us fix $\gamma_2 = 1$ in (15). Intuitively, the closer the initial is $\mathbf{g}^0$ to the ground truth $\mathbf{g}^*$, the larger the γ_1 we should set; otherwise the contribution of the first term to the objective function will be small. In practice, however, we typically do not have exact information about how far $\mathbf{g}^0$ is from $\mathbf{g}^*$; we therefore have to appropriately tune γ_1. One possible criterion is that fixing the parameters involved in Algorithm 1, a "good" γ_1 should lead to a reduction of the objective function value of the BiLev as much as possible.

To solve the BiLev numerically, thus adjusting the demand vector iteratively, we leverage a gradient-based algorithm (Algorithm 1). In particular, suppose that the route probabilities are locally constant. For OD pair $i \in [\![\mathcal{W}]\!]$, we consider only the fastest route $r_i(\mathbf{g})$, where in each iteration, based on the updated link flows after the previous iteration, we update link travel times and assign them as current link weights in the graph model introduced in Section III-A. Then, we have [11]

$$\frac{\partial x_a(\mathbf{g})}{\partial g_i} \approx \delta_{r_i(\mathbf{g})a} = \begin{cases} 1, & \text{if } a \in r_i(\mathbf{g}) \\ 0, & \text{otherwise.} \end{cases} \qquad (16)$$

(Note that we have assumed the partial derivatives do exist; a comprehensive discussion on the existence and calculation of $\partial x_a(\mathbf{g}) / \partial g_i$ can be found in [29].) Thus, by (16) we obtain an approximation to the Jacobian matrix

$$\left[\frac{\partial x_a(\mathbf{g})}{\partial g_i};\ a \in \mathcal{A},\ i \in [\![\mathcal{W}]\!]\right]. \qquad (17)$$

Let us now compute the gradient of $F(\mathbf{g})$. We have

$$\begin{aligned} \nabla F(\mathbf{g}) &= \left(\frac{\partial F(\mathbf{g})}{\partial g_i};\ i \in [\![\mathcal{W}]\!]\right) \\ &= \Bigg(2\gamma_1\left(g_i - g_i^0\right) + 2\gamma_2 \sum_{a\in\mathcal{A}} \left(x_a(\mathbf{g}) - \tilde{x}_a\right) \frac{\partial x_a(\mathbf{g})}{\partial g_i}; \\ &\qquad i \in [\![\mathcal{W}]\!]\Bigg). \end{aligned} \qquad (18)$$

Remark 3: There are three reasons why we consider only the fastest routes for the purpose of calculating the Jacobian. 1) GPS navigation is widely used by vehicle drivers so that they tend to always select the fastest routes between their OD pairs. 2) There are very efficient algorithms for finding the fastest route for each OD pair. 3) If considering more than one route for an OD pair, then the route flows cannot be uniquely determined by solving the TAP (2), thus leading to unstable route-choice probabilities, which would undermine the accuracy of the approximation to the Jacobian matrix in (17).

We summarize the procedures for adjusting the OD demand matrices as Algorithm 1, whose convergence will be proven in the following proposition.

Algorithm 1. Adjusting OD Demand Matrices

```
Input: the road network $(\mathcal{V}, \mathcal{A}, \mathcal{W})$; the function $f(\cdot)$ in (1);
    the observed flow vector from given data $\tilde{\mathbf{x}} = (\tilde{x}_a;\ a \in \mathcal{A})$; the initial demand vector $\mathbf{g}^0 = (g_i^0;\ i \in [\![\mathcal{W}]\!])$; two
    positive integer parameters $\rho$, $T$; two real parameters $\varepsilon_1 \geq 0$, $\varepsilon_2 > 0$.
Step 1: Initialization. Take the demand vector $\mathbf{g}^0$ as the input, solve the TAP (2) (using the Method of Successive
    Averages (MSA) [36]) to obtain $\mathbf{x}^0$. Set $l = 0$. If $F\left(\mathbf{g}^0\right) = 0$, stop; otherwise, go onto Step 2.
Step 2: Computation of a descent direction. Calculate $\mathbf{h}^l = -\nabla F\left(\mathbf{g}^l\right)$ by (18).
Step 3: Calculation of a search direction. For $i \in [\![\mathcal{W}]\!]$ set
    $\bar{h}_i^l = \begin{cases} h_i^l, & \text{if } \left(g_i^l > \varepsilon_1\right) \text{ or } \left(g_i^l \leq \varepsilon_1 \text{ and } h_i^l > 0\right), \\ 0, & \text{otherwise.} \end{cases}$
Step 4: Armijo-type line search.
    4.1: Calculate the maximum possible step-size $\theta_{\max}^l = \min\left\{-g_i^l/\bar{h}_i^l;\ \bar{h}_i^l < 0, i \in [\![\mathcal{W}]\!]\right\}$.
    4.2: Determine $\theta^l = \arg\min_{\theta\in\mathcal{S}} F\left(\mathbf{g}^l + \theta\bar{\mathbf{h}}^l\right)$, where
        $\mathcal{S} \overset{\text{def}}{=} \left\{\theta_{\max}^l, \theta_{\max}^l/\rho, \theta_{\max}^l/\rho^2, \ldots, \theta_{\max}^l/\rho^T, 0\right\}$.
Step 5: Update and termination.
    5.1: Set $\mathbf{g}^{l+1} = \mathbf{g}^l + \theta^l\bar{\mathbf{h}}^l$. Using $\mathbf{g}^{l+1}$ as the input, solve the TAP (2) to obtain $\mathbf{x}^{l+1}$.
    5.2: If $\frac{F(\mathbf{g}^l) - F(\mathbf{g}^{l+1})}{F(\mathbf{g}^0)} < \varepsilon_2$, stop the iteration; otherwise, go onto Step 5.3.
    5.3: Set $l = l + 1$ and return to Step 2.
```

Proposition 2: Algorithm 1 converges.

Proof: If the initial demand vector $\mathbf{g}^0$ saisfies $F\left(\mathbf{g}^0\right) = 0$, then, by Step 1, the algorithm stops (trivial case). Otherwise, we have $F\left(\mathbf{g}^0\right) > 0$, and it is seen from (15) that the objective function $F(\mathbf{g})$ has a lower bound 0. In addition, by the line search and the update steps (Steps 4.2 and 5.1, in particular), we obtain

$$\begin{aligned} F\left(\mathbf{g}^{l+1}\right) &= F\left(\mathbf{g}^l + \theta^l\bar{\mathbf{h}}^l\right) \\ &= \min_{\theta\in\mathcal{S}} F\left(\mathbf{g}^l + \theta\bar{\mathbf{h}}^l\right) \leq F\left(\mathbf{g}^l\right) \qquad \forall l \end{aligned}$$

where the last inequality holds due to $0 \in \mathcal{S}$, indicating that the nonnegative objective function in (15) is nonincreasing as the number of iterations increases. Thus, by the well-known monotone convergence theorem, the convergence of the algorithm can be guaranteed.

Remark 4: Algorithm 1 is a variant of the algorithms proposed in [11] and [12]. We use a different method to calculate the step-sizes (resp., Jacobian matrix) than that in [11] (resp., [12]). The optimization problem BiLev is not convex because of the potential nonlinearity in $\mathbf{x}(\mathbf{g})$. Thus, one would not necessarily expect Algorithm 1's convergence to a global minimum. In addition, due to inaccuracies in the gradient calculation, one would not expect Algorithm 1's convergence to a local minimum either. A discussion on the performance of similar heuristics can be found in [12]. It is worth noting that, in [12], the proposed "descent" algorithm could possibly not "descend" in some iterations due to computational inaccuracy of the gradient. We will demonstrate our findings for the performance of Algorithm 1 by numerical experiments in Section V-B. We also note that, in terms of decreasing the objective function value of the BiLev, the performance of Algorithm 1 definitely depends heavily on the initial demand vector $\mathbf{g}^0$.

E. Price of Anarchy

As discussed in Section I, one of our goals is to measure inefficiency in the network due to the noncooperative behavior of drivers. Thus, we compare the network performance under a user-centric routing policy versus a system-centric one. As a metric for this comparison, we conceptually define the PoA as the ratio between the total travel latency cost, i.e., the total travel time over all drivers, obtained under Wardrop flows (user-centric routing policy) and that obtained under socially optimal flows (system-centric routing policy).

Given road network $(\mathcal{V}, \mathcal{A}, \mathcal{W})$, as in [30], we calculate its total travel latency cost as

$$L(\mathbf{x}) = \sum_{a\in\mathcal{A}} x_a t_a(x_a). \tag{19}$$

The socially optimal flow vector, denoted by $\mathbf{x}^{\text{social}} = (x_a^{\text{social}}; a \in \mathcal{A})$, is the solution to the following system-centric forward problem, which is an NLP [6], [13]:

$$(\text{socialOpt}) \quad \min_{\mathbf{x}\in\mathcal{F}} \sum_{a\in\mathcal{A}} x_a t_a(x_a). \tag{20}$$

We therefore explicitly define the PoA as

$$\text{PoA} \overset{\text{def}}{=} \frac{L(\mathbf{x}^{\text{user}})}{L(\mathbf{x}^{\text{social}})} = \frac{\sum_{a\in\mathcal{A}} x_a^{\text{user}} t_a(x_a^{\text{user}})}{\sum_{a\in\mathcal{A}} x_a^{\text{social}} t_a(x_a^{\text{social}})} \geq 1 \tag{21}$$

where $\mathbf{x}^{\text{user}} = (x_a^{\text{user}}; a \in \mathcal{A})$ is the Wardrop equilibrium flow vector assumed to be directly observable or indirectly inferrable. By the definition of $\mathbf{x}^{\text{social}}$, we always have PoA ≥ 1; the larger the PoA is, the larger the inefficiency induced by selfish drivers. Thus, PoA quantifies the inefficiency that a societal group has to deal with due to noncooperative behavior of its members.

We note that the objective function in (20) is different from its counterpart in (2); for a detailed explanation, see [18]. However, the two forward problem formulations have a very tight connection. Let us take a close look at the following equalities [6]:

$$\bar{t}_a(x_a) \overset{\text{def}}{=} \frac{d}{dx_a}(x_a t_a(x_a)) = t_a(x_a) + x_a \dot{t}_a(x_a) \qquad \forall a \in \mathcal{A}. \tag{22}$$

By (22) we see that the socialOpt in (20) is equivalent to

$$(\text{userOpt}) \quad \min_{\mathbf{x}\in\mathcal{F}} \sum_{a\in\mathcal{A}} \int_0^{x_a} \bar{t}_a(s)\, ds.$$

The remarkable implication of the above is that in order to find the socially optimal flows x_a^{social}, $a \in \mathcal{A}$, instead of directly solving (20), it suffices to solve (2) with $t_a(\cdot)$ replaced by $\bar{t}_a(\cdot)$. As noted in [6], the difference between the social cost and the user cost is $x_a \dot{t}_a(x_a)$, which can be interpreted as the cost a user (driver) imposes on the other users.

Let $\bar{\mathbf{t}}(\mathbf{x}) \overset{\text{def}}{=} (\bar{t}_a(x_a); a \in \mathcal{A})$. To ensure the existence and uniqueness of the solution to (20), we need the following assumption.

Assumption B: $\bar{\mathbf{t}}(\cdot)$ is strongly monotone on $\mathcal{F}$ and continuously differentiable on $\mathbb{R}_+^{|\mathcal{A}|}$. $\mathcal{F}$ satisfies Slater's condition [33].

We note that if Assumption A holds and, for all $a \in \mathcal{A}$, $t_a(x_a)$ is convex and twice continuously differentiable on $\mathbb{R}_+$ [e.g., $t_a(x_a) = 2x_a^2 + x_a + 1$], then Assumption B holds as well.

F. Sensitivity Analysis

To prioritize road segments for potential congestion reducing interventions by the local transportation authorities, we investigate the sensitivities of the optimal objective function value of (2) with respect to key parameters, specifically, free-flow travel time and flow capacity. In particular, we first derive two rigorous formulas, and then propose their finite difference approximations as an alternative.

Write $\mathbf{t}^0 \overset{\text{def}}{=} \left(t_a^0; a \in \mathcal{A}\right)$, $\mathbf{m} \overset{\text{def}}{=} (m_a; a \in \mathcal{A})$, and

$$V\left(\mathbf{t}^0, \mathbf{m}\right) \overset{\text{def}}{=} \min_{\mathbf{x}\in\mathcal{F}} \sum_{a\in\mathcal{A}} \int_0^{x_a} t_a^0 f\left(\frac{s}{m_a}\right) ds. \tag{23}$$

Differentiating (23), for each $a \in \mathcal{A}$, we obtain

$$\frac{\partial V\left(\mathbf{t}^0, \mathbf{m}\right)}{\partial t_a^0} = \int_0^{x_a^{\text{user}}} f\left(\frac{s}{m_a}\right) ds \tag{24}$$

$$\frac{\partial V\left(\mathbf{t}^0, \mathbf{m}\right)}{\partial m_a} = \int_0^{x_a^{\text{user}}} t_a^0 \dot{f}\left(\frac{s}{m_a}\right)\left(-\frac{s}{m_a^2}\right) ds \tag{25}$$

where $\dot{f}(\cdot)$ denotes the derivative of $f(\cdot)$.

Note that typically we have

$$\frac{\partial V(\mathbf{t}^0,\mathbf{m})}{\partial t_a^0} > 0$$

and

$$\frac{\partial V(\mathbf{t}^0,\mathbf{m})}{\partial m_a} < 0$$

meaning a slight decrease (resp., increase) of t_a^0 (resp., m_a) would reduce the objective function value of (2). Based on this observation, for $a' \in \mathcal{A}$, we define the following quantities:

$$\Delta V\left(\mathbf{t}^0,\mathbf{m};\Delta t_{a'}^0\right) \overset{\text{def}}{=} \min_{\mathbf{x}\in\mathcal{F}} \sum_{a\in\mathcal{A}} \int_0^{x_a} t_a^0 f\left(\frac{s}{m_a}\right) ds - \min_{\mathbf{x}\in\mathcal{F}} \left[\sum_{a\in\mathcal{A},a\neq a'} \int_0^{x_a} t_a^0 f\left(\frac{s}{m_a}\right) ds + \int_0^{x_{a'}} \left(t_{a'}^0 + \Delta t_{a'}^0\right) f\left(\frac{s}{m_{a'}}\right) ds\right] \quad (26)$$

and

$$\Delta V\left(\mathbf{t}^0,\mathbf{m};\Delta m_{a'}\right) \overset{\text{def}}{=} \min_{\mathbf{x}\in\mathcal{F}} \sum_{a\in\mathcal{A}} \int_0^{x_a} t_a^0 f\left(\frac{s}{m_a}\right) ds - \min_{\mathbf{x}\in\mathcal{F}} \left[\sum_{a\in\mathcal{A},a\neq a'} \int_0^{x_a} t_a^0 f\left(\frac{s}{m_a}\right) ds + \int_0^{x_{a'}} t_{a'}^0 f\left(\frac{s}{m_{a'} + \Delta m_{a'}}\right) ds\right] \quad (27)$$

where $\Delta t_{a'}^0 \overset{\text{def}}{=} -0.2 \times \min\{t_a^0; a \in \mathcal{A}\}$ and $\Delta m_{a'} \overset{\text{def}}{=} 0.2 \times \min\{m_a; a \in \mathcal{A}\}$. Note that, by construction, for each and every $a \in \mathcal{A}$, we approximately have $0 < \Delta V\left(\mathbf{t}^0,\mathbf{m};\Delta t_a^0\right) \propto \partial V(\mathbf{t}^0,\mathbf{m})/\partial t_a^0$ and $0 < \Delta V\left(\mathbf{t}^0,\mathbf{m};\Delta m_a\right) \propto \left|\partial V(\mathbf{t}^0,\mathbf{m})/\partial m_a\right|$.

IV. DATA SET DESCRIPTION AND PROCESSING

In this section, based on our data-driven approach outlined in Section III, we conduct a case study using actual traffic data from the EMA road network [35], [37].

A. Description of the EMA Data Set

We deal with two data sets concerning the EMA road network. 1) The speed data set, made available to us by the Boston Region Metropolitan Planning Organization (MPO), includes the spatial average speeds for more than 13 000 road segments (with an average length of 0.7 miles; see Fig. 1) of EMA, providing the average speed for every minute of 2012. For each road segment, identified with a unique tmc (traffic message channel) code, the data set provides information such as speed data (instantaneous, average and free-flow speed) in miles per hour (mph), date and time, and traveling time (in minutes) through that segment. 2) The flow capacity (in vehicles per hour) data set, also provided by the MPO, includes capacity data for more than 100 000 road segments (with an average length of 0.13 miles) in EMA. For more detailed information of these two data sets, see [30].

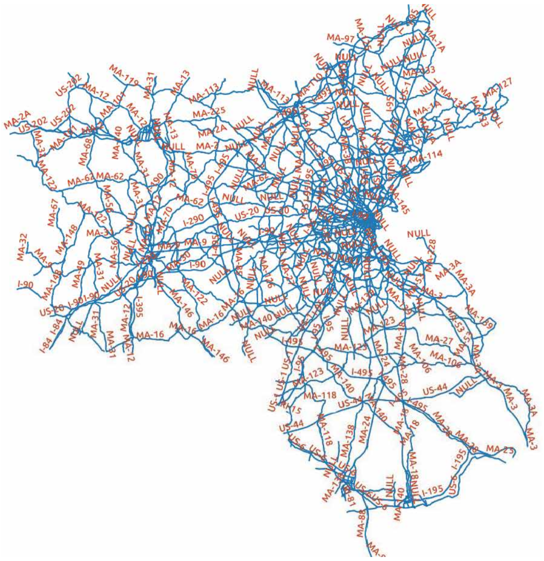


Fig. 1. ***All available road segments in EMA (from [30]).***

B. Preprocessing

In [30] and [31], we investigate two relatively small subnetworks [denoted by $\mathcal{I}_1$ and $\mathcal{I}_2$ and shown in Fig. 2(a) and (b), respectively] of the EMA road network. Here, we further consider a much larger subnetwork (denoted by $\mathcal{I}_3$ and shown in Fig. 3). Performing similar preprocessing procedures as those in [30] and [31], we end up with traffic flow data (Wardrop equilibria) and road (link) parameters (flow capacity and free-flow travel time) for the three subnetworks $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$, where $\mathcal{I}_1$ contains only interstate highways, $\mathcal{I}_2$ also contains state highways, and $\mathcal{I}_3$ covers a much wider area of EMA. Note that $\mathcal{I}_1$ (resp., $\mathcal{I}_2$, $\mathcal{I}_3$) consists of 8

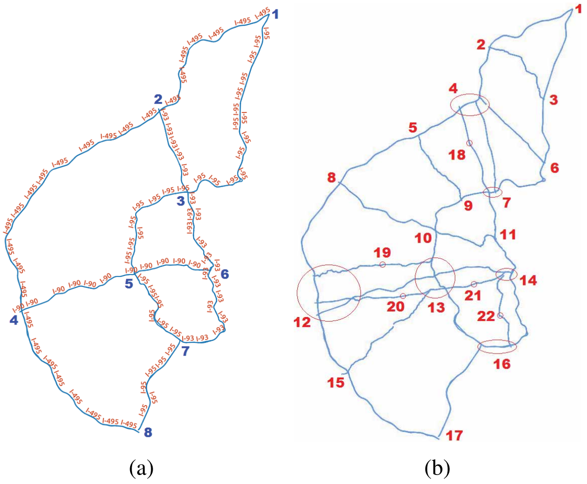


Fig. 2. ***(a) An interstate highway subnetwork of EMA ($\mathcal{I}_1$) (the blue numbers indicate node indices). (b) An extended highway subnetwork of EMA ($\mathcal{I}_2$) (the red numbers indicate node indices). (See [35] for the correspondences between nodes and link indices.)***

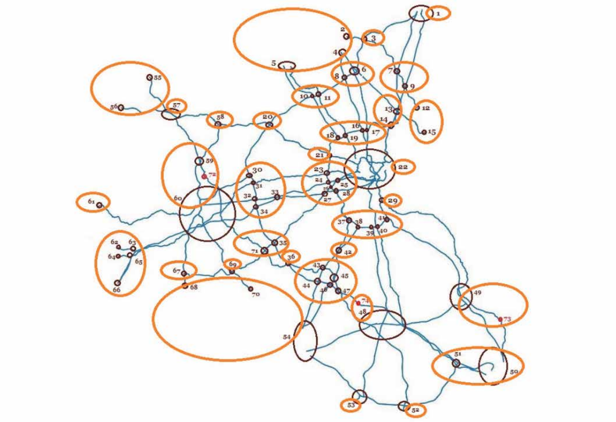


Fig. 3. ***A wider EMA highway subnetwork ($\mathcal{I}_3$); details on the correspondences between nodes and link indices are in [35]. ("nodes:zone" pairs– {1}: Seabrook (NH); {2, 4, 5}: NH; {3}: Haverhill; {6, 8}: Lawrence; {7, 9}: Georgetown; {10, 11}: Lowell; {12, 15}: Salem; {13, 14}: Peabody; {16, 17, 18, 19}: Burlington; {20}: Littleton; {21}: Lexington; {22}: Boston; {23, 24, 25, 26, 27, 28}: Waltham; {29}: Quincy; {30, 31, 32, 33, 34}: Marlborough/Framingham; {35, 71}: Milford; {36}: Franklin; {37, 38, 39, 40, 41}: Westwood/Quincy; {42}: Dedham; {43, 44, 45, 46, 47}: Foxborough; {48, 74}: Taunton; {49, 73}: Plymouth; {50, 51}: Cape Cod; {52}: Dartmouth; {53}: Fall River; {54, 68, 70}: RI; {55, 56}: VT; {57}: Westminster; {58}: Leominster; {59, 60, 72}: Worcester; {61}: Amherst; {62, 63, 64, 65, 66}: CT; {67}: Webster; {69}: Uxbridge.)***

(resp., 22, 74) nodes and 24 (resp., 74, 258) links. Assuming that each node could be an origin and a destination, then there are $8 \times (8-1) = 56$ (resp., $22 \times (22-1) = 462$) OD pairs in $\mathcal{I}_1$ (resp., $\mathcal{I}_2$). For $\mathcal{I}_3$, we simplify the analysis by grouping nodes within the same area, assigning them the same zone label, thus obtaining 34 zones (as opposed to 74 nodes). Assuming that each zone could be an origin and a destination, then there are $34 \times (34-1) = 1122$ OD pairs in $\mathcal{I}_3$. It is worth pointing out that nodes 18, 19, 20, 21, and 22 (resp., 72, 73, and 74) in $\mathcal{I}_2$ (resp., $\mathcal{I}_3$) are introduced for ensuring the identifiability of the OD demand matrices. More specifically, to "recover" uniquely an OD demand matrix from observed link flow data, the link-route incidence matrix **A** is required to satisfy certain structural properties; see [26, Lemma 2].

C. Estimating Initial OD Demand Matrices

Operating on $\mathcal{I}_1$, we solve the QP (P1) [cf., (13)] and the QCP (P2) [cf., (14)] using data corresponding to five different time periods [AM, MD (middle day), PM, NT (night), and WD (weekend)] of four months (January, April, July, and October) in 2012, thus obtaining 20 different OD demand matrices for these scenarios. Expanding each and every of the 20 OD demand matrices of $\mathcal{I}_1$ by setting the demand for any OD pair that belongs to $\mathcal{I}_2$ but does not belong to $\mathcal{I}_1$ to zero, we obtain "rough" initial demand matrices for $\mathcal{I}_2$.

On the other hand, for the much larger subnetwork $\mathcal{I}_3$, to obtain initial OD demand matrices corresponding to the same 20 scenarios, we perform a different simplification procedure. In particular, we only consider the shortest route for each OD pair of $\mathcal{I}_3$, thus leading to a deterministic route choice matrix **P** and significantly reducing the number of decision variables in the QCP (P2).

As noted in [30], the GLS method assumes the traffic network to be uncongested. It follows that the estimated OD demand matrices for nonpeak periods (MD/NT/WD) are relatively more accurate than those for peak periods (AM/PM). After obtaining estimates for travel latency cost functions in Section IV-D, based on the observed Wardrop flows and the initial estimates for the OD demand matrices, we will conduct demand adjustment procedures for $\mathcal{I}_2$ and $\mathcal{I}_3$ in Section IV-E.

D. Cost Function Estimation and Sensitivity Analysis

First, to validate the effectiveness and efficiency of the cost function estimator (12), we conduct numerical experiments over the Anaheim benchmark network [37], whose ground truth cost functions, OD demand matrices, and all necessary road parameters are available. Next, operating on $\mathcal{I}_1$ using the flow data and the OD demand matrices obtained in Sections IV-B and IV-C, respectively, we estimate the travel latency cost functions, $f(\cdot)$ in particular, for 20 different scenarios, via the estimator (12), by solving the QP invVI-2 accordingly. As in [30], to make the estimates reliable, for each scenario, we perform a threefold cross validation. Note that [30] applied a different estimator, which is numerically not as stable.

We assume that such estimates for $f(\cdot)$, as obtained from $\mathcal{I}_1$, can be shared by all the three subnetworks $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$; this makes sense, because the function $f(\cdot)$ is common for all links and, when estimating it through $\mathcal{I}_1$, we have already made use of a large amount of data (note that there are 24 links in $\mathcal{I}_1$ and the flow data and the corresponding OD demand matrices that we use have covered 120 different time instances for each of the 20 scenarios; for details, see [35]).

To illustrate our method of analyzing sensitivities for the TAP formulation (2), we again conduct numerical experiments on $\mathcal{I}_1$. In particular, we investigate a scenario corresponding to the AM peak period of April 2012.

E. OD Demand Adjustments

First, we demonstrate the effectiveness of Algorithm 1 using the Anaheim benchmark network. Then, assuming the per-road travel latency cost functions are available (we take the travel latency cost functions derived from $\mathcal{I}_1$ as in Section IV-D), we apply Algorithm 1 to $\mathcal{I}_2$, which contains $\mathcal{I}_1$ as one of its representative subnetworks. Note that the main difference between $\mathcal{I}_1$ and $\mathcal{I}_2$ is the modeling emphasis; specifically, $\mathcal{I}_1$ only takes account of interstate highways, while $\mathcal{I}_2$ also encompasses state highways, thus containing more details of the real road network of EMA. We can think of

$\mathcal{I}_1$ as a "landmark" subnetwork of $\mathcal{I}_2$. Based on the initially estimated demand matrices for $\mathcal{I}_1$, we will implement the following generic demand-adjusting scheme so as to derive the OD demand matrices for $\mathcal{I}_2$.

Given a network ($\mathcal{I}_2$ in our case) of any size we can select its "landmark" subnetworks ($\mathcal{I}_1$ in our case) (based on the information of road types, preidentified centroids, etc.) with acceptably smaller sizes; say we end up with N ($N = 1$ in our case) such subnetworks. Then, for each subnetwork, we estimate its demand matrix by solving sequentially the QP (P1) and the QCP (P2) (cf., Section IV-C). Setting the demand for any OD pair not belonging to this subnetwork to zero, we obtain a "rough" initial demand matrix for the entire network ($\mathcal{I}_2$ in our case). Next, we take the average of these initial demand matrices. Finally, we adjust the average demand matrix based on the flow observations of the entire network.

Next, taking again the travel latency cost functions derived from $\mathcal{I}_1$, we apply Algorithm 1 to $\mathcal{I}_3$, based on the initial OD demand matrices estimated from $\mathcal{I}_3$ (see Section IV-C) and the Wardrop flows inferred from $\mathcal{I}_3$ (see Section IV-B).

As noted in Remark 1, the reason for not directly solving (P1) and (P2) for the larger networks ($\mathcal{I}_2$ and $\mathcal{I}_3$ in our case) is that there are too many decision variables in (P2) and this would lead to numerical difficulties.

F. PoA Evaluation and Meta-Analysis

We calculate the PoA values for $\mathcal{I}_2$ and $\mathcal{I}_3$ for the PM period of April 2012. For each day, in (21) we take the average observed link flows over the PM period as the "user flows," and obtain the "social flows" by solving the NLP (20) using the estimated cost functions and demand matrix exclusively for the PM period. To solve (20), we use the IPOPT solver [38] which implements a primal–dual interior point method [39].

To better understand the performance of the road network under the user-centric versus the system-centric routing policy, we conduct a meta-analysis on $\mathcal{I}_3$. In particular, under the two policies, we compare congestion for various zones of the network, the maximum/minimum link flows, and link-specific congestion.

V. Numerical Results

For economy of space, we will not show the detailed results for the initial estimation of OD demand matrices. However, we report in Table 1 the entries of the route choice probability matrix **P** derived for $\mathcal{I}_1$ for some specific OD pairs (the complete results can be found in [35]). It is seen from Table 1 that we cannot always expect a higher probability for a shorter/faster route; randomness exists. However, this is not necessarily counterintuitive, because the selected routes for the same OD pair have close lengths/travel times. We note here that when identifying (and refining) the feasible routes for each OD pair of $\mathcal{I}_1$, we consider at most three shortest routes (ranked #1–#3) and discard all the routes with a length larger than that of the shortest route (ranked #1) by more than 50%. Note also that since this initial OD estimation procedure does not involve real-time updates of traffic conditions, we may use either travel times or lengths as weights for links in the graph model.

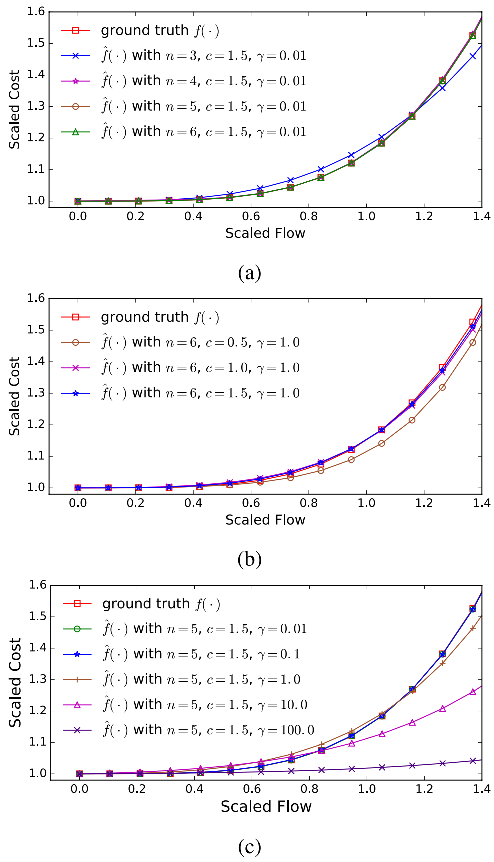
Table 1 Selected Route Choice Analysis Results for $\mathcal{I}_1$ (Corresponding to the PM Peak Period of April 2012)

OD pair	refined feasible route	route length (in miles)	free-flow travel time (in hours)	route choice probability
(1, 8)	$1 \to 2 \to 3 \to 5 \to 7 \to 8$	74.2072	1.0235	0.3265
	$1 \to 3 \to 5 \to 7 \to 8$	74.6696	1.0297	0.3394
	$1 \to 2 \to 3 \to 6 \to 7 \to 8$	74.8692	1.0522	0.3341
(2, 4)	$2 \to 4$	37.6346	0.5123	0.8274
	$2 \to 3 \to 5 \to 4$	43.4554	0.6010	0.1004
	$2 \to 3 \to 6 \to 5 \to 4$	50.7995	0.7274	0.0722
(3, 5)	$3 \to 5$	16.2154	0.2262	0.8375
	$3 \to 6 \to 5$	23.5596	0.3526	0.1625
(8, 3)	$8 \to 7 \to 5 \to 3$	43.3260	0.6065	0.4364
	$8 \to 7 \to 6 \to 3$	43.4313	0.6310	0.3022
	$8 \to 7 \to 5 \to 6 \to 3$	50.2382	0.7308	0.2614

In the following, we will focus on presenting the results for the estimates of the travel latency cost functions (derived for the Anaheim benchmark network and $\mathcal{I}_1$), the demand adjustment procedure (derived for the Anaheim benchmark network; note that we will not show the detailed demand adjustment results for $\mathcal{I}_2$ and $\mathcal{I}_3$, because we do not have the ground truth for a comparison), the PoA evaluations (derived for $\mathcal{I}_2$ and $\mathcal{I}_3$), the sensitivity analysis (derived for $\mathcal{I}_1$), and the meta-analysis (derived for $\mathcal{I}_3$).

A. Results From Estimating the Cost Functions

1) Results for the Anaheim Benchmark Network: The Anaheim network contains 38 zones [hence $38 \times (38 - 1) = 1406$ OD pairs], 416 nodes, and 914 links. The ground truth $f(\cdot)$ is taken as $f(z) = 1 + 0.15z^4, z \geq 0$. Fig. 4 shows the estimation results for $f(z)$ by solving invVI-2 corresponding to different parameter settings. In particular, Fig. 4(a) shows the curves of the ground truth $f(z)$ and the estimator $\hat{f}(z)$ corresponding to n taking values from $\{3, 4, 5, 6\}$ while keeping c and γ fixed to 1.5 and 0.01 respectively; it is seen that except for the case $n = 3$, all estimation curves are very close to the ground truth. Note that the ground truth $f(z)$ is a polynomial function with degree 4, which is greater than 3. This suggests the use of a value $n \geq 4$ in recovering the cost function $f(\cdot)$. The intuition here is that we can use a higher order polynomial with appropriate coefficients to approximate a lower order polynomial, but not *vice versa*. Fig. 4(b) shows the curves of the ground truth $f(z)$ and the estimator $\hat{f}(z)$ corresponding to c taking values from $\{0.5, 1.0, 1.5\}$ while keeping n and γ fixed to 6 and 1.0 respectively; it is seen that except for the case $c = 0.5$, the estimation curves are very close to the ground truth. This suggests that setting c reasonably larger should give better estimation results. Fig. 4(c) plots the curves of the ground truth $f(z)$ and the



(a)

(b)

(c)

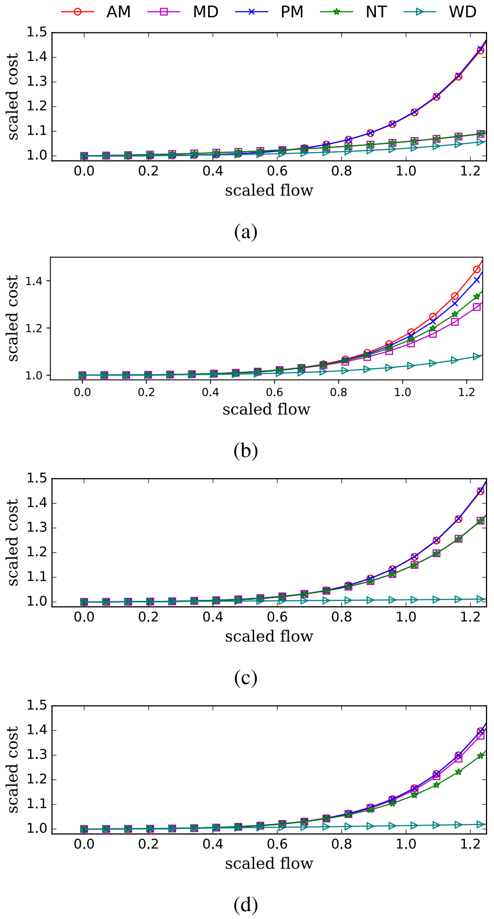
Fig. 4. ***Estimations for cost function $f(\cdot)$ by solving invVI-2 corresponding to different parameter settings (Anaheim). (a) Vary n (c and γ fixed). (b) Vary c (n and γ fixed). (c) Vary γ (n and c fixed).***

estimator $\hat{f}(z)$ corresponding to γ taking values from $\{0.01, 0.1, 1.0, 10.0, 100.0\}$ while keeping n and c fixed to 5 and 1.5 respectively; it is seen that as γ is set smaller and smaller, the estimation curve gets closer and closer to the ground truth. This suggests that choosing a smaller regularization parameter γ should give tighter estimation results in terms of data reconciling.

2) *Results for $\mathcal{I}_1$:* We show the comparison results of the cost functions in Fig. 5, where in each subfigure, we plot the curves of the estimated $f(\cdot)$ corresponding to five different time periods. For economy of space, we will not list the parameter setting details of n, c, and γ, which were selected by conducting a threefold cross validation.

We observe from Fig. 5(a)–(d) that the costs for peak periods (AM/PM) are more sensitive to traffic flows than for nonpeak periods (MD/NT/WD). This can be explained as follows: during rush hour, it is very common for vehicles to pass through a congested road network while during nonrush hour, drivers mostly enjoy an uncongested road network.

In addition, it is seen that, for different months, the cost curves for nonpeak periods differ more significantly than for peak periods. Aside from the observation and modeling errors, this can also be explained by seasonal traveling patterns.



(a)

(b)

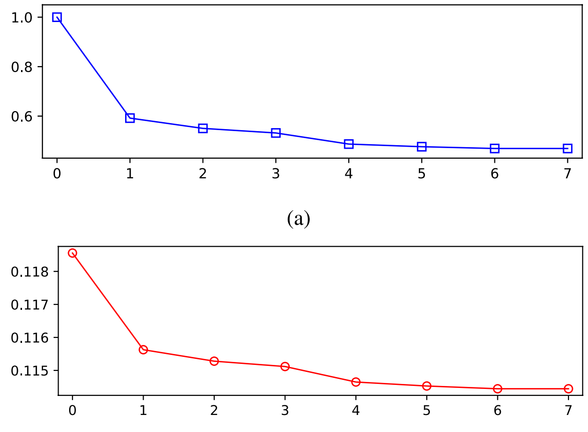
(c)

(d)

Fig. 5. ***Estimates for $f(\cdot)$ corresponding to different time periods [AM, MD (middle day), PM, NT (night), WD (weekend)], derived from data over $\mathcal{I}_1$ for 2012. (a) Jan. (b) Apr. (c) Jul. (d) Oct.***

B. Results From OD Demand Adjustment

We now present the OD demand adjustment results from the Anaheim network. For each OD pair, the initial demand is taken by scaling the ground truth demand using a random factor with uniform distribution over $[0.8, 1.2]$. The ground truth $f(\cdot)$ is taken as $f(z) = 1 + 0.15z^4$, $\forall z \geq 0$, and is assumed directly available. When implementing Algorithm 1, we set $\gamma_1 = 0$, $\gamma_2 = 1$, $\rho = 2$, $T = 10$, $\varepsilon_1 = 0$, and $\varepsilon_2 = 10^{-20}$. Fig. 6(a) shows that, after seven iterations, the objective function value of the BiLev (15) has been reduced by more than 50%. Fig. 6(b) shows that the distance between the adjusted demand and the ground truth demand keeps decreasing with the number of iterations, and the distance changes very slightly, meaning the adjustment procedure does not alter the initial demand much. Note that in Fig. 6(a), the vertical axis corresponds to the normalized objective function value of the BiLev, i.e., $F(\mathbf{g}^l)/F(\mathbf{g}^0)$ and, in Fig. 6(b), the vertical axis denotes the



(a)

(b)

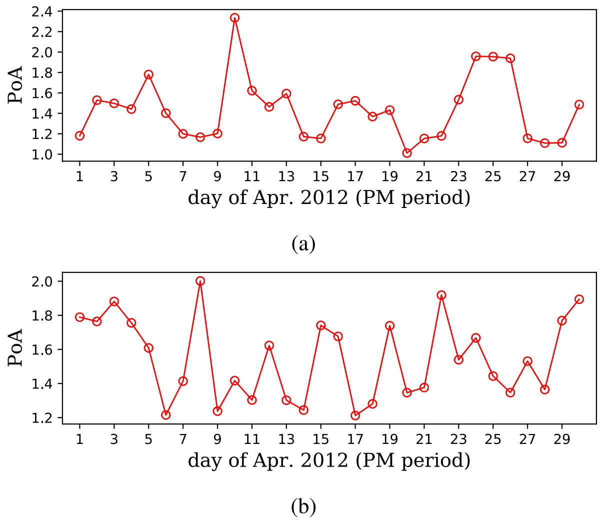
Fig. 6. ***Key quantities versus number of iterations (Anaheim). (a) $F(\mathbf{g}^l)/F(\mathbf{g}^0)$ vs. # of iterations. (b) $\|\mathbf{g}^l - \mathbf{g}^*\|/\|\mathbf{g}^*\|$ vs. # of iterations.***

normalized distance between the adjusted demand vector and the ground truth, i.e., $\|\mathbf{g}^l - \mathbf{g}^*\|/\|\mathbf{g}^*\|$, where $\mathbf{g}^*$ is the ground-truth demand vector.

C. Results for PoA Evaluation

After implementing the demand adjusting scheme, we obtain the demand matrices for $\mathcal{I}_2$ and $\mathcal{I}_3$ on a daily basis, as opposed to those for $\mathcal{I}_1$ on a monthly basis. Note that, even for the same period of a day and within the same month, slight demand variations among different days are possible; thus, our PoA results for $\mathcal{I}_2$ and $\mathcal{I}_3$ would be more accurate than those for $\mathcal{I}_1$ (shown in [30]).

The PoA values for $\mathcal{I}_2$ shown in Fig. 7(a) have larger variations than those for $\mathcal{I}_1$ in [30] and for $\mathcal{I}_3$ shown in Fig. 7(b); some are closer to 1 but some go beyond 2.2, meaning we have larger potential to improve the road network. It is also seen that, although $\mathcal{I}_2$ is extracted from $\mathcal{I}_3$, there is no obvious correlation between the PoA values estimated for $\mathcal{I}_2$ and $\mathcal{I}_3$. To explain this, one should note the fact that $\mathcal{I}_2$ is only a small subnetwork of $\mathcal{I}_3$, where the latter contains many more nodes/links/OD pairs [see Figs. 2(b) and 3]. Specifically, in Fig. 3, many more links have been added which significantly alter the feasible routing patterns relative to Fig. 2(b). Thus, even though there may be correlations at the individual link flow level, once we add links and then aggregate over all links, any correlation is likely weakened or lost. Moreover, the social optimization problems solved to obtain the denominator of the PoA ratio in (21) are very different since the subnetwork topologies are different. However, when taking the average of the PoA values for all 30 days of April 2012, all $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$ result in an average PoA approximately equal to 1.5, meaning we can gain an efficiency improvement of about 50%; thus, the results are consistent.



(a)

(b)

Fig. 7. ***Daily PoAs for $\mathcal{I}_2$ and $\mathcal{I}_3$ (PM period for April 2012): (a) for $\mathcal{I}_2$; (b) for $\mathcal{I}_3$.***

D. Results From Sensitivity Analysis

Investigating the AM peak period of April 2012 for $\mathcal{I}_1$, instead of directly applying (24) and (25), we calculate the two quantities defined in (26) and (27), and plot the results in Fig. 8, where the blue (resp., red) curve indicates the quantity $\Delta V(\mathbf{t}^0, \mathbf{m}; \Delta t_a^0)$ (resp., $\Delta V(\mathbf{t}^0, \mathbf{m}; \Delta m_a)$) for each and every link of $\mathcal{I}_1$. It is seen from Fig. 8 that the largest four values of $\Delta V(\mathbf{t}^0, \mathbf{m}; \Delta t_a^0)$ (resp., $\Delta V(\mathbf{t}^0, \mathbf{m}; \Delta m_a)$) correspond to links 10, 19, 9, and 5 (resp., 10, 19, 9, and 1). This suggests that, during the AM peak period of April 2012, the transportation management department could have most efficiently reduced the objective function value of the TAP (2), thus mitigating congestion, by taking actions with priorities on these links (e.g., improving road conditions to reduce the free-flow travel time for links 10, 19, 9, and 5, and increasing the number of lanes to enlarge the flow capacity for links 10, 19, 9, and 1).

E. Results From Meta-Analysis

We conduct meta-analysis for $\mathcal{I}_3$, under the user-centric routing policy versus the system-centric one. Our analysis includes the zone costs, the maximum/minimum link flows, and the link-specific congestion.

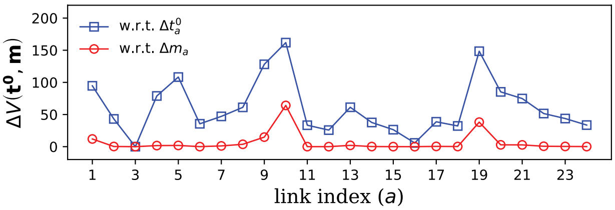
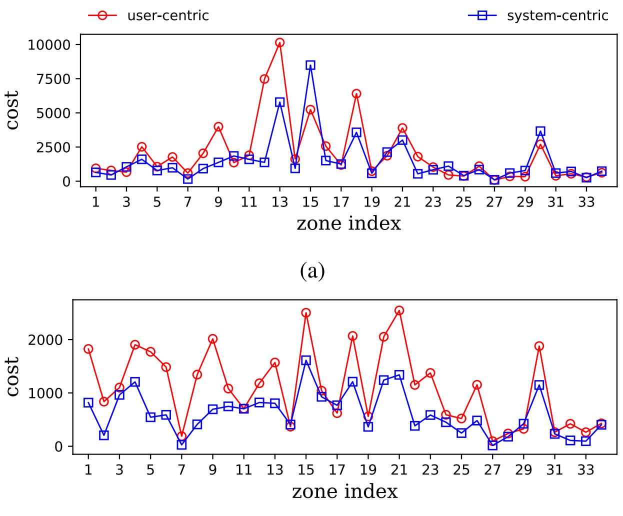


Fig. 8. ***Sensitivity analysis (finite difference approximation) results for $\mathcal{I}_1$; AM period of April 2012.***



(a)

(b)

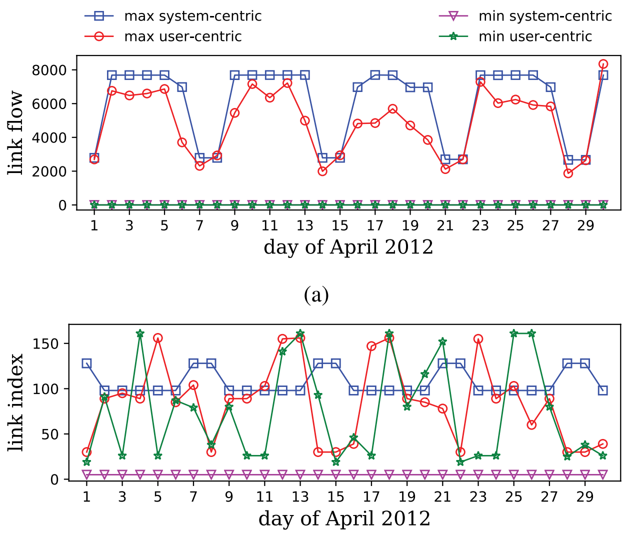
Fig. 9. ***Zone costs under user-centric versus system-centric routing policy (PM period of April 2012). (a) weekday (4/18/2012). (b) Weekend (4/15/2012).***

1) Meta-Analysis for Zone Costs: Let $\mathcal{A}_3^i$ denote the set of links related to zone i of $\mathcal{I}_3$ (each link in $\mathcal{A}_3^i$ has at least one node contained in zone i). Then, the total users' travel latency cost for zone i is defined as

$$C_i = \sum_{a \in \mathcal{A}_3^i} x_a t_a(x_a).$$

We consider two scenarios, one corresponding to the PM peak period of a typical weekday (Wednesday, April 18, 2012) and the other the PM period of a typical weekend (Sunday, April 15, 2012). The zone costs under the user-centric (resp., system-centric) routing policy are visualized in Fig. 9(a) [resp., Fig. 9(b)]. Three observations can be made. 1) Overall, most zone costs would be reduced when switching from the user-centric routing policy to the system-centric one. 2) In general, the zone costs for weekends are less than their counterparts for weekdays; this is consistent with intuition. 3) The decrease seems more consistent for all zones during weekends than during weekdays, suggesting it is easier to optimize the network during weekends; this is again consistent with intuition.

2) Meta-Analysis for Maximum/Minimum Link Flows The maximum/minimum link flows for the PM peak period of each and every day of April 2012 are plotted in Fig. 10(a), and the corresponding link indices are shown in Fig. 10(b). A major observation, based on Fig. 10(a), is that the maximum link flow values would increase for most of the days when switching the routing policy from the user-centric one to the system-centric one, which is desirable. In addition, it is seen that, among the entire month (April 2012), both the maximum link flows under the two routing policies have a weekly periodic distribution; this is consistent with intuition.



(a)

(b)

Fig. 10. ***Maximum/minimum link flows and the corresponding link indices under user-centric versus system-centric routing policy (PM period of April 2012).***

3) Meta-Analysis for Link Congestion: For any given link a, we define its congestion metric (CM) [40] as the ratio of the travel time to free-flow travel time

$$\mathrm{CM}_a \overset{\text{def}}{=} \frac{t_a(x_a)}{t_a^0} = f\left(\frac{x_a}{m_a}\right) \tag{28}$$

where $f(\cdot)$ is the cost function that we have estimated. By this definition, we always have $\mathrm{CM}_a \geq 1$.

We first consider a PM peak period scenario for a typical workday (Wednesday, April 18, 2012). The CM values of all the 258 links are plotted in Fig. 11 in a logarithmic scale (base 2). It is seen that, for some links (indexed with 79, 92, and 86) the CM value is significantly higher (gap > 1) under the user-centric routing policy than under the system-centric one. There are some links for which we have

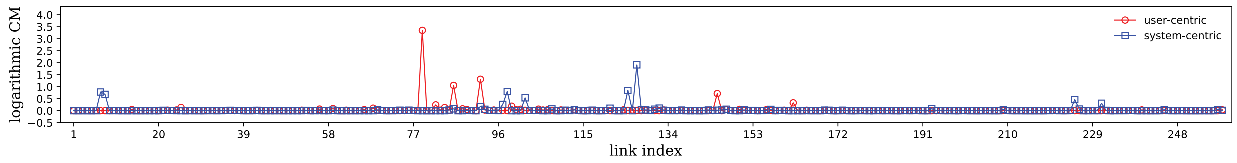


Fig. 11. ***Link congestion under user-centric versus system-centric routing policy (PM period of April 18, 2012).***

the opposite, but, overall, the CM peak is reduced under the system-centric policy. We then investigate a PM period scenario for a typical weekend (Sunday, April 15, 2012), and find that all the CM values for this scenario are very close to 1, meaning there was almost no congestion for all links; we have omitted the weekend CM plot for economy of space.

VI. STRATEGIES FOR POA REDUCTION

After quantifying the PoA, a natural question we must answer is the following: How can we reduce the PoA for a given transportation network? We propose three practical strategies for reducing the PoA, especially when PoA $\gg 1$.

First, by taking advantage of the rapid emergence of CAVs, it has become feasible to automate routing decisions, thus solving a system-centric forward problem [cf., (20)] in which all CAVs (bypassing driver decisions) cooperate to optimize the overall system performance.

Second, we propose a modification to existing GPS navigation algorithms recommending to all drivers socially optimal routes, which could be implemented by making use of (22). In particular, we can solve the user-centric forward problem (2), embedded in a typical GPS navigation application, with $t_a(\cdot)$ replaced by $\bar{t}_a(\cdot)$, whose common cornerstone part $f(\cdot)$ is estimated using (12). It is worth pointing out that some existing work simply took $f(\cdot)$ to be the Bureau of Public Roads (BPR)'s [7] empirical polynomial function $f(z) = 1 + 0.15z^4, \forall z \geq 0$, which would not be as accurate.

Finally, our sensitivity analysis results provide the means to prioritize road segments for specific interventions that can mitigate congestion.

VII. CONCLUSION AND FUTURE WORK

In this paper, we assess the efficiency of transportation networks under a selfish user-centric routing policy as opposed to a socially optimal system-centric routing policy. To that end, we define and quantify the PoA and propose possible strategies to reduce it. All the procedures involved are data driven, thus having the capability of dynamically optimizing any given transportation network (by using the data collected in real-time manner), in terms of reducing the PoA (especially when PoA $\gg 1$) such that it gets as close to 1 as possible.

We must keep in mind that, due to unavoidable inaccuracies in data and modeling, all the numerical results shown in Section V are only estimates. In particular, the speed-to-flow conversion model that we use (Greenshield's model) is a macroscopic model with naturally limited accuracy, the GLS method that we leverage also is based on an approximation, and the MSA subroutine in Algorithm 1 is an approximate scheme.

In terms of the computational challenges of our proposed approaches, we encountered numerical difficulties when solving (13) and (14) to obtain OD demands for large-sized (say a network like $\mathcal{I}_3$) networks. However, we subsequently developed a simplification procedure by considering only the fastest route for each OD pair, thus successfully resolving this issue. We conducted case studies on a workstation with 24-GB memory and a 12-core Intel Core i5 CPU, and for the largest network ($\mathcal{I}_3$) that we investigated, the total CPU time (including estimating OD demands, recovering link latency cost functions, adjusting OD demands, solving for socially optimal flows, and finally calculating PoA values) is about 10 h. The total CPU times for $\mathcal{I}_1$ and $\mathcal{I}_2$ are about 30 min and 2 h, respectively. We note that the most time-consuming task is adjusting OD demands using Algorithm 1. However, it is seen that Steps 2–4 of Algorithm 1 and the MSA subroutine can easily benefit from parallel computing. Thus, scalability can be further improved through parallel computation. Moreover, following an approach of "divide and conquer," several decomposition methods could possibly also be leveraged as we move to larger networks; the difficulty lies in how to reasonably "merge" results derived for subnetworks so as to obtain the final result for the whole network.

Our ongoing work includes extending the PoA analysis and reduction framework from single-class to multiclass transportation networks. We have recently obtained results for the multiclass user-centric inverse problem [41], which paves the way for data-driven PoA estimation in these networks. We are also considering alternative models/methods to improve the accuracy in the PoA evaluation. In addition, it is of interest to consider jointly estimating/adjusting the OD demand matrices and recovering the travel latency cost functions.

Acknowledgements

The authors would like to thank the Boston Region Metropolitan Planning Organization, and S. Peterson in particular, for supplying the EMA data and providing them invaluable clarifications throughout their work. ■

REFERENCES

[1] "World's Population Increasingly Urban With More Than Half Living in Urban Areas," *Report on World Urbanization Prospects, United Nations Department of Economic and Social Affairs*, Jul. 2014. [Online]. Available: http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html

[2] D. Shrank, T. Lomax, and B. Eisele, "The 2011 urban mobility report," Texas A&M Transp. Inst., Tech. Rep., 2011. [Online]. Available: https://nacto.org/docs/usdg/2011_urban_mobility_report_schrank.pdf

[3] INRIX, "Economic and environmental of traffic congestion in Europe and the U.S.," 2015. [Online]. Available: http://www.inrix.com/economic-environment-cost-congestion/

[4] H. Youn, M. T. Gastner, and H. Jeong, "Price of anarchy in transportation networks: Efficiency and optimality control," *Phys. Rev. Lett.*, vol. 101, no. 12, pp. 128701/1–128701/4, 2008.

[5] J. G. Wardrop, "Some theoretical aspects of road traffic research," *Proc. Inst. Civil Eng.*, vol. 1, pp. 325–378, 1952.

[6] P. Patriksson, *The Traffic Assignment Problem: Models and Methods*. Utrecht, The Netherlands, 1994.

[7] D. Branston, "Link capacity functions: A review," *Transp. Res.*, vol. 10, no. 4, pp. 223–236, 1976.

[8] T. Abrahamsson, "Estimation of origin-destination matrices using traffic counts—A literature survey," IIASA, Tech. Rep. IR-98-021, May 1998.

[9] S. Bera and K. V. K. Rao, "Estimation of origin-destination matrix from traffic counts: The state of the art," *Eur. Transp.*, vol. 49, pp. 2–23, 2011.

[10] D. Bertsimas, V. Gupta, and I. C. Paschalidis, "Data-driven estimation in equilibrium using inverse optimization," *Math. Programm.*, pp. 1–39, 2015.

[11] H. Spiess, "A gradient approach for the OD matrix adjustment problem," Centre de Recherche sur les Transports, Univ. de Montréa, Montreal, QC, Canada, Tech. Rep. 693, 1990.

[12] J. T. Lundgren and A. Peterson, "A heuristic for the bilevel origin–destination-matrix estimation problem," *Transp. Res. B, Methodol.*, vol. 42, no. 4, pp. 339–354, 2008.

[13] S. Pourazarm, C. G. Cassandras, and T. Wang, "Optimal routing and charging of energy-limited vehicles in traffic networks," *Int. J. Robust Nonlinear Control*, vol. 26, no. 6, pp. 1325–1350, 2016.

[14] J. Alonso, "Autonomous vehicle control systems for safe crossroads," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 6, pp. 1095–1110, 2011.

[15] J. Lee, B. B. Park, K. Malakorn, and J. J. So, "Sustainability assessments of cooperative vehicle intersection control at an urban corridor," *Transp. Res. C, Emerg. Technol.*, vol. 32, pp. 193–206, 2013.

[16] Y. J. Zhang, A. A. Malikopoulos, and C. G. Cassandras, "Optimal control and coordination of connected and automated vehicles at urban traffic intersections," in *Proc. Amer. Control Conf. (ACC)*, 2016, pp. 6227–6232.

[17] J. Rios-Torres and A. A. Malikopoulos, "A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1066–1077, May 2017.

[18] S. C. Dafermos and F. T. Sparrow, "The traffic assignment problem for a general network," *J. Res. Nat. Bureau Standards B*, vol. 73, no. 2, pp. 91–118, 1969.

[19] L. J. LeBlanc, E. K. Morlok, and W. P. Pierskalla, "An efficient approach to solving the road network equilibrium traffic assignment problem," *Transp. Res.*, vol. 9, no. 5, pp. 309–318, 1975.

[20] S. C. Dafermos, "The traffic assignment problem for multiclass-user transportation networks," *Transp. Sci.*, vol. 6, no. 1, pp. 73–87, 1972.

[21] A. Nagurney, "A multiclass, multicriteria traffic network equilibrium model," *Math. Comput. Model.*, vol. 32, nos. 3–4, pp. 393–411, 2000.

[22] S. Ryu, A. Chen, and K. Choi, "Solving the stochastic multi-class traffic assignment problem with asymmetric interactions, route overlapping, and vehicle restrictions," *J. Adv. Transp.*, vol. 52, no. 2, pp. 255–270, 2015.

[23] T. L. Friesz, J. Luque, R. L. Tobin, and B.-W. Wie, "Dynamic network traffic assignment considered as a continuous time optimal control problem," *Oper. Res.*, vol. 37, no. 6, pp. 893–901, 1989.

[24] B. N. Janson, "Dynamic traffic assignment for urban road networks," *Transp. Res. B, Methodol.*, vol. 25, no. 2, pp. 143–161, 1991.

[25] S. Nguyen, *Estimating Origin Destination Matrices From Observed Flows*. Amsterdam, The Netherlands: Elsevier, 1984.

[26] M. L. Hazelton, "Estimation of origin–destination matrices from link flows on uncongested networks," *Transp. Res. B, Methodol.*, vol. 34, no. 7, pp. 549–566, 2000.

[27] H. Yang, Q. Meng, and M. G. Bell, "Simultaneous estimation of the origin-destination matrices and travel-cost coefficient for congested networks in a stochastic user equilibrium," *Transp. Sci.*, vol. 35, no. 2, pp. 107–123, 2001.

[28] H. Yang, "Sensitivity analysis for the elastic-demand network equilibrium problem with applications," *Transp. Res. B, Methodol.*, vol. 31, no. 1, pp. 55–70, 1997.

[29] M. Patriksson, "Sensitivity analysis of traffic equilibria," *Transp. Sci.*, vol. 38, no. 3, pp. 258–281, 2004.

[30] J. Zhang, S. Pourazarm, C. G. Cassandras, and I. C. Paschalidis, "The price of anarchy in transportation networks by estimating user cost functions from actual traffic data," in *Proc. IEEE 55th Conf. Decision Control (CDC)*, Dec. 2016, pp. 789–794.

[31] J. Zhang, S. Pourazarm, C. G. Cassandras, and I. C. Paschalidis, "Data-driven estimation of origin-destination demand and user cost functions for the optimization of transportation networks," in *Proc. 20th IFAC World Congr.*, Toulouse, France, Jul. 2017, pp. 9680–9685.

[32] B. Monnot, F. Benita, and G. Piliouras, "How bad is selfish routing in practice?," 2017. [Online]. Available: https://arxiv.org/abs/1703.01599

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[34] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, 2000.

[35] J. Zhang, S. Pourazarm, C. G. Cassandras, and I. C. Paschalidis (2017). *InverseVIsTraffic*. [Online]. Available: https://github.com/jingzbu/InverseVIsTraffic

[36] Y. Noriega and M. A. Florian, "Algorithmic approaches for asymmetric multi-class network equilibrium problems with different class delay relationships," 2007. [Online]. Available: https://www.cirrelt.ca/DocumentsTravail/CIRRELT-2007-30.pdf

[37] H. Bar-Gera (2017). *Transportation Networks for Research*. [Online]. Available: https://github.com/bstabler/TransportationNetworks

[38] A. Wächter and C. Laird (2016). *Interior Point OPTimizer*. [Online]. Available: https://en.wikipedia.org/wiki/IPOPT

[39] S. J. Wright, *Primal-Dual Interior-Point Methods*. Philadelphia, PA, USA: SIAM, 1997.

[40] M. Aftabuzzaman, "Measuring traffic congestion-a critical review," *Austral. Transp. Res. Forum*, vol. 30, no. 1, 2007.

[41] J. Zhang and I. C. Paschalidis, " Data-driven estimation of travel latency cost functions via inverse optimization in multi-class transportation networks," in *Proc. 56th IEEE Conf. Decision Control*, Melbourne, Vic., Australia, Dec. 2017.

ABOUT THE AUTHORS

Jing Zhang (Student Member, IEEE) received the B.S. degree in mathematics from Wuhan University, Wuhan, Hubei, China, in 2010, the M.S. degree in complex systems and control from the Institute of Systems Science, Chinese Academy of Sciences, Beijing, China, in 2013, and the Ph.D. degree in systems engineering from Boston University, Boston, MA, USA, in 2017.

He is currently a Research Scientist at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. His research interests include anomaly detection, optimization, machine learning, and time series analysis.

Dr. Zhang is a recipient of the Boston Area Research Initiative (BARI) Research Seed Grant Award (Spring 2017). He placed second within the New England states in the 2017 Net Impact & Toyota Next Generation Mobility Challenge.

Sepideh Pourazarm (Student Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in electrical engineering/control systems from the K.N. Toosi University of Technology, Tehran, Iran, in 2004 and 2007, respectively, and the Ph.D. degree in systems engineering from Boston University, Boston, MA, USA, in 2017.

From 2007 to 2011, she was an Instrumentation and Control Engineer in the oil and gas industry in Tehran, Iran. She currently works as a Data Scientist in Nielsen Corp., Needham, MA. Her research experiences are in the areas of optimization algorithms, time-series analysis, machine learning, and modeling and control of energy-aware systems with applications to wireless sensor networks, transportation networks, and energy-limited vehicles.

Dr. Pourazarm was a recipient of several awards, including the 2017 College of Engendering Societal Impact Dissertation Award and the 2017

Systems Engineering Outstanding Dissertation of the Year Award at Boston University, a 2016 ACM's Student Research Competition Travel Award, a 2015 Honorable Mention of the Center for Information and Systems Engineering Graduate Research Symposium, a 2015 NSF Data Science Workshop Travel Award, and a 2014 IEEE ICNSC Best Student Paper Finalist Award.

Christos G. Cassandras (Fellow, IEEE) received the B.S. degree from Yale University, New Haven, CT, USA, in 1977, the M.S.E.E. degree from Stanford University, Stanford, CA, USA, in 1978, and the M.S. and Ph.D. degrees from Harvard University, Cambridge, MA, USA, in 1979 and 1982, respectively.

He was with ITP Boston, Inc., Cambridge, MA, USA, from 1982 to 1984, where he was involved in the design of automated manufacturing systems. From 1984 to 1996, he was a faculty member at the Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst, MA, USA. He is currently a Distinguished Professor of Engineering with Boston University, Brookline, MA, USA, the Head of the Division of Systems Engineering, and a Professor of Electrical and Computer Engineering. He specializes in the areas of discrete event and hybrid systems, cooperative control, stochastic optimization, and computer simulation, with applications to computer and sensor networks, manufacturing systems, and transportation systems. He has authored about 400 refereed papers in these areas, and six books.

Dr. Cassandras is a member of Phi Beta Kappa and Tau Beta Pi. He is also a Fellow of the International Federation of Automatic Control (IFAC). He is a recipient of several awards, including the 2011 IEEE Control Systems Technology Award, the 2006 Distinguished Member Award of the IEEE Control Systems Society, the 1999 Harold Chestnut Prize (IFAC Best Control Engineering Textbook), a 2011 prize and a 2014 prize for the IBM/IEEE Smarter Planet Challenge competition, the 2014 Engineering Distinguished Scholar Award at Boston University, several honorary professorships, a 1991 Lilly Fellowship, and a 2012 Kern Fellowship. He was the Editor-in-Chief of the IEEE Transactions on Automatic Control from 1998 to 2009. He serves on several editorial boards and has been a Guest Editor for various journals. He was the President of the IEEE Control Systems Society in 2012.

Ioannis Ch. Paschalidis (Fellow, IEEE) received the M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1993 and 1996, respectively.

In September 1996, he joined Boston University, Boston, MA, USA, where he has been ever since. He is a Professor at Boston University with appointments in the Department of Electrical and Computer Engineering, the Division of Systems Engineering, and the Department of Biomedical Engineering. He is the Director of the Center for Information and Systems Engineering (CISE). He has held visiting appointments with MIT and Columbia University, New York, NY, USA. His current research interests lie in the fields of systems and control, networking, applied probability, optimization, operations research, computational biology, and medical informatics.

Dr. Paschalidis is a recipient of the NSF CAREER award (2000), several best paper and best algorithmic performance awards, and a 2014 IBM/IEEE Smarter Planet Challenge Award. He was an invited participant at the 2002 Frontiers of Engineering Symposium, organized by the U.S. National Academy of Engineering and the 2014 U.S. National Academies Keck Futures Initiative (NAFKI) Conference. He is the inaugural Editor-in-Chief of the IEEE Transactions on Control of Network Systems.