



Bridging the user equilibrium and the system optimum in static traffic assignment: a review

Valentina Morandi¹

Received: 22 September 2021 / Revised: 13 October 2022 / Accepted: 19 March 2023 /
Published online: 3 April 2023
© The Author(s) 2023

Abstract

Solving the road congestion problem is one of the most pressing issues in modern cities since it causes time wasting, pollution, higher industrial costs and huge road maintenance costs. Advances in ITS technologies and the advent of autonomous vehicles are changing mobility dramatically. They enable the implementation of a coordination mechanism, called coordinated traffic assignment, among the sat-nav devices aiming at assigning paths to drivers to eliminate congestion and to reduce the total travel time in traffic networks. Among possible congestion avoidance methods, coordinated traffic assignment is a valuable choice since it does not involve huge investments to expand the road network. Traffic assignments are traditionally devoted to two main perspectives on which the well-known Wardropian principles are inspired: the user equilibrium and the system optimum. User equilibrium is a user-driven traffic assignment in which each user chooses the most convenient path selfishly. It guarantees that fairness among users is respected since, when the equilibrium is reached, all users sharing the same origin and destination will experience the same travel time. The main drawback in a user equilibrium is that the system total travel time is not minimized and, hence, the so-called Price of Anarchy is paid. On the other hand, the system optimum is an efficient system-wide traffic assignment in which drivers are routed on the network in such a way the total travel time is minimized, but users might experience travel times that are higher than the other users travelling from the same origin to the same destination, affecting the compliance. Thus, drawbacks in implementing one of the two assignments can be overcome by hybridizing the two approaches, aiming at bridging users' fairness to system-wide efficiency. In the last decades, a significant number of attempts have been done to bridge fairness among users and system efficiency in traffic assignments. The survey reviews the state-of-the-art of these trade-off approaches.

✉ Valentina Morandi
valentina.morandi1@unibs.it

¹ Economics and Business Department, University of Brescia, Brescia, Italy

Keywords Hybrid traffic assignment · Price of anarchy · Congestion avoiding · Fair assignment

Mathematics Subject Classification 90-02

1 Introduction

Road congestion is becoming a serious and one of the most urgent problems in metropolitan areas, where the traffic demand is steadily growing. Congestion is a significant burden in terms of wasted time, pollution, industrial costs and road maintenance. Hence, alleviating traffic volumes will become more and more urgent as the population grows. Travellers complain about traffic congestion because it adds a delay to their travel times that can be used for other activities. On the industry side, delays reduce productivity and, consequently, increase the operating costs. Congestion can influence a lot of economic decisions because it affects the choice of the living place, the working place and the travelling mode for most of the population living in urban areas. In addition, congestion continues to increase because of the continuous growth of the population and the increased motorization ratio. According to a new report by INRIX and Centre for Economics and Business Research (CEBR), the annual cost of traffic congestion and gridlocks on individual households and national economies in the U.S., U.K., France and Germany will rise to \$293 billion dollars in 2030 with a cumulative value of congestion cost, from now to 2030, near \$4.4 trillion. According to U.N., the transportation plays a central role in sustainable development. In fact, in 1997 the U.N. estimated that, over the next twenty years, transportation would be expected to be the major driving force behind a growing world demand for energy. Hence, the U.N. decided to focus on sustainable transportation including it in the 11th Sustainable Development Goal, i.e. making cities and human settlements inclusive, safe, resilient and sustainable. Sustainable transportation can be applied in the context of infrastructure, public transport systems, goods delivery networks, affordability, efficiency and convenience of transportation, as well as improving urban air quality and health, and reduce greenhouse gas emissions.

As the demand for transport increases, the traffic planner could tackle the traffic congestion issue in two ways: enlarging infrastructures or optimizing the resources at hand. Enlarging infrastructures means, in most cases, that the road network has to be extended to accommodate the growing demand. This is not always convenient in terms of costs/benefits trade-off and/or feasible because of environmental constraints. In fact, the idea of enlarging the existing infrastructure or to build new ones should be the last resort in the light of being sustainable and to follow the circular economy principles. So spotlights have been moved on how to use efficiently existing infrastructures and, if possible, to find new sustainable way to cleverly use them. In the last year, many attempts aiming at regulating congestion have been done using Intelligent Transportation Systems (ITS), such as ramp metering, reversible lanes, limited access roads, bus lanes, carpooling lanes, express toll lanes, congestion pricing mechanisms, tradable credit schemes, variable message signs, etc., or coordinating traffic assignment. Reviews on congestion reduction methods can be found in Papageorgiou and

Kotsialos (2000), Peeta et al. (2000), Sichitiu and Kihl (2008) and Luo and Hubaux (2004). Although all proposed attempts had an impact on road congestion, some of them are willing to be used and exploited in the future because of their synergy with new technologies, such as autonomous vehicles and next generation sat-nav devices with, eventually, the advent of 5 G. Actual sat-nav devices has only partially solved the congestion problem since the provided information is a snapshot of the current traffic condition on which the sat-nav decisions are based. Without any knowledge of the future behaviour of drivers, the result is simply a shift of the congestion in other parts of the city network. This is because even the most efficient sat-nav devices consider only the actual information of traffic on the road networks, without considering the impact of the simultaneous choices on the traffic patterns. In fact, as assessed by Klein et al. (2018), using real-time sat-nav devices, the resulting traffic pattern is likely to be close to usual inefficient equilibrium rather than near to a system optimal traffic pattern. Hence, one of the final pieces of the jigsaw will be having information about the will and the behaviour of each driver in terms of path to be followed, paired with timing. At the same time, information on the network status, which is implemented mainly using road-side sensors, is continuously changing and better and reliable communication systems are needed. In this sense, latest developments in autonomous vehicles and vehicle to vehicle communications are paving the way to the coordination among vehicles. People traveling with autonomous vehicles will declare their intent before starting the journey and vehicle will evaluate the best choice among feasible paths in order to reach the destination. Moreover, the vehicle is most likely connected with road-side sensors, with other vehicles and communicating the tour planning to a central authority is easy. According to these new technologies, having a central coordination mechanism implementing a fair and efficient traffic assignment will become soon the ruler of the roost among ITS approaches in eliminating traffic congestion and, more in general, of reducing the total travel time in congested road networks. In fact, the traffic coordination has been acknowledged in Speranza (2018) as one of the most prominent trends in transportation and logistic.

A centralized coordination system may optimize the network performance and paths may be assigned to vehicles according to an optimal assignment. However, traffic coordination can be easily applied to current road networks only if individual needs are taken into account. It is well known that a centralized system optimizing network performance that assigns paths to user without any consideration about fairness among users will tremendously affect the users' compliance to the system. Thus, coordinated traffic assignment on real road networks has to be efficient from the system perspective but also fair from the users' point of view.

Traffic assignments are traditionally divided into two main approaches inspired by the well-known Wardropian principles: the user equilibrium and the system optimum. User equilibrium is a user-driven traffic assignment in which each user chooses the most convenient path selfishly. It guarantees that fairness among users is respected since, when the equilibrium is reached, all users sharing the same origin and destination will experience the same travel time. The main drawback of implementing the user equilibrium is that the total travel time is not minimized. In fact, the inefficiencies produced by the user equilibrium are well known in literature under the name of "price of anarchy", i.e. the price the system is willing to pay to let users choose the

route on their own. On the other hand, the system optimum is a system-wide traffic assignment in which drivers are routed on the network in such a way the total travel time is minimized. Unfortunately, users might experience travel times that are higher than the other users travelling from the same origin to the same destination. This is because the focus is only on reducing the system travel time. As assessed in Klein et al. (2018), the system optimum is the most efficient assignment while being “unstable” since it is unfair and users could not comply with the guidance prescriptions. Since there are drawbacks in using one of the two main approaches, in the last years several attempts to bridge the users’ fairness with an efficient traffic assignment have been developed.

To this aim, in this survey the literature bridging the two different perspectives will be explained and deeply discussed, along with many open research questions that can be tackled in the immediate future. According to Sheffi (1985), the rush hour time windows could be modeled as a continuous and constant demand for transportation and, hence, static traffic assignment models are the right tool to tackle the congestion problem. The natural extension is the dynamic traffic assignment problem that could be used with different demand patterns and provides more detailed information about traffic flows. Although analyzing dynamic traffic assignment models is out of the scope of the paper, we will refer to Peeta and Ziliaskopoulos (2001), Boyce et al. (2001) and Saw et al. (2015) for comprehensive reviews.

Before going through the latest and the most important developments in fair and efficient traffic assignments models, the survey will go through the survey methodology in Sect. 2, through the concept of road congestion and how it can be measured in Sect. 3. Then, in Sect. 4, the most common models used in traffic assignment optimization are shown. The two former sections introduce, to a neophyte in the field, to the main concepts needed to understand and implement traffic assignment models. Then, in Sect. 5, the state of the art of approaches bridging the user equilibrium and the system optimum will be thoroughly discussed. Finally, in Sect. 7, conclusions and ideas for future research will be provided.

2 Survey methodology

In this survey, we focus on studies where traffic assignment models are formulated to address both the issue of efficiency and fairness in congested road networks.

With this aim, contributes to the literature have been searched through the main scholar databases for operations research, transportation and game theory such as Scholar, Scopus and Web of Science. Keywords used for the search are: price of anarchy, Braess’ paradox, fair traffic assignment, efficient traffic assignment, constrained system optimum, bounded rational user equilibrium, congestion charging, congestion tolls.

Starting from the results of the former keywords, a preliminary set of relevant publications has been selected. Then, references therein have been analyzed searching for articles that were missing in the first search phase. Selection has been conducted with the aim to focus only on traffic assignment models and, in particular, in measuring and/or reducing the gap between traditional user-driven or centralized approaches.

Only contributions with a strong modelling flavour have been selected for this review. In order to keep the number of relevant publications at a reasonable level, only journal publications, books and seminal proceedings have been selected. Journal publications and book chapters cover the majority of the citations, i.e. the 88 and the 3.5% respectively, while conference proceedings represent the 6% of the total number of publication. A few technical reports (2.5%) have been selected for their importance. The distribution over journals is wide since many aspects of the traffic assignment models have been analyzed. However, research articles on the modelling part are mainly published in specialized journals, such as *Transportation Science*, *Transportation Research* (especially Part B and C), *European Journal of Operations Research* and *Transportation Research Record*, which cover more the 40% of the contained citations. Research items are mostly presented in a chronological order from older ones to new advances in the specific field. At the end, the survey accounts for a final set of 117 selected publications.

3 Congestion avoidance and measures

Traffic congestion is the result of the imbalance between the network capacity and the demand for transportation. According to Falcocchio and Levinson (2015), congestion in transportation occurs when the number of vehicles travelling on a road segment reaches unacceptable levels of discomfort or delay. Congestion phenomena are divided into two main categories: the **recurring** and the **non-recurring** congestion. According to Falcocchio and Levinson (2015) and Stopher (2004), the recurring congestion is the delay that travellers regularly experiences during certain periods of time (for example, the rush hour or morning commute). The non-recurring congestion is a delay due to not predictable events that disrupt the traffic flow such as car breakdowns, crashes, works in progress and bad weather conditions. When dealing with recurrent congestion, collecting information is crucial, as assessed by Ben-Elia et al. (2013). In Ben-Elia et al. (2013), an experiment with different level of information accuracy is carried out and the negative effect of low information levels is demonstrated. However, even with full information provided, in case of bottlenecks it is necessary to reconsider network design features.

How the congestion is measured? When does the congestion appear? The traffic congestion can be detected comparing the actual speed with a theoretic free-flow speed, i.e. the maximum speed allowed on a road segment, or comparing the amount of vehicles on a certain road segment with a threshold defining the maximum amount of vehicles for which the road segment is considered congestion-free. Congestion intensity measures are many and they allow understanding the level of discomfort that can be experienced on a certain road network. A well-known congestion intensity measure is the congestion delay rate, i.e. the difference between actual travel time rate and free-flow travel time rate (min/km), i.e. the travel time under maximum allowed speed. The USA Transportation Research Board (Ryus et al. (2011)) uses the experienced speed in order to classify roads with respect to the Level of Service (LoS), i.e. grades from A (free-flow) to F (forced breakdown congestion). Another intensity measure is the travel time index that is the ratio between the free-flow speed and the experienced

speed. The main intensity measure specifically used in traffic assignment problems is the road congestion. The road congestion is obtained as the ratio of the number of vehicles travelling on the road segment (arc in traffic assignment literature) and the capacity of the road. The road capacity has not to be seen as a strong bound on the number of vehicles that can flow on the road segment but, rather, it has to be seen as a threshold from which the travel time on the network will start to increase significantly.

Congestion can be measured also from the users' side. A measure for the congestion experienced by a user is the so-called users' unfairness. This measure is evaluated on the whole user's path, and it is defined as the relative difference between the experienced travel time and the free-flow travel time. It depends on the free-flow speed and on the experienced speed, but also on the trip length. According to Falcocchio and Levinson (2015), longer trips are impacted more by congestion with respect to shorter trip so, considering only the experienced speed, the measure could be misleading.

In all the considered measures, the evaluation of the experienced travel time on arch road segment is crucial. How to model the travel time on each road, considering congestion effects? In Stopher (2004), it has been pointed out that the congestion is a phenomenon that occurs when the demand exceeds the road capacity. Considering this definition, the underlying assumption is that the experienced travel time and the experienced speed depend on the demand travelling on that road segment. The relationship between travel time and demand is usually expressed by a so-called latency function, where the travel time is a non-linear function of the congestion level expressed as percentage of the capacity saturated by the demand. Related concepts and most used latency functions will be described in 4.

Congestion detection mechanisms are out of the scope of this survey, but it is worthy to mention that it is a big issue for traffic planners. In recent years, many devices have been developed in order to derive vehicle speed, safety distance between vehicles and other congestion parameters. Main methods are RFID sensors, CCTV cameras and vehicle to vehicle communications.

4 Traffic assignments and the price for anarchy

According to Patriksson (2015), transportation planning is usually divided into five steps: goal definition, base year inventory, model analysis, travel forecast and network evaluation. The goal definition step is related to find an agreement on goals and objectives. In the base year inventory step, all the data related to the network and demand patterns has to be collected. In the model analysis phase the relationship between measured quantities (traffic flows and road congestion, for instance) is searched for. Model analysis is the result of four different phases: trip generation, trip distribution, modal split and traffic assignment. Trip generation consists in finding the number of trips that originate and terminate in different zone of the studied area. Usually this phase is carried out considering socio-economic, geographic and land use features and the different zones are categorized by main purpose as work, leisure or shopping area. In trip distribution phase some formulas to predict the demand of travellers from an origin zone to a destination zone have to be developed. To an origin zone and to a destination zone is usually associated an OD pair with the demand of transportation

from its origin to its destination. Usually, the demand of travellers is a function of an attractive parameter for each zone. Modal split is a phase in which we determine the mean of transport used by each traveller. The number of travellers that choose a particular mean of transport depends mainly on travel cost in terms of monetary cost or travel time, but, sometimes, socio-economic factors also affect the choice. Traffic assignment is devoted to assign the demand from an origin to a destination to routes in transportation network. This phase is particularly relevant because an estimate of traffic demand and travel time is returned. Once the model analysis has been done, a travel forecast is produced using data collected in the goal definition step and re-calibrated with the results of the model analysis step. Finally, the network evaluation is a phase in which alternative transportation network and facilities benefits are evaluated and compared. In this literature review we will focus only on the traffic assignment phase of the model analysis step.

Traffic assignment is a method that assigns the demand of the OD pair to trips on a transportation network. As input of traffic assignment, the OD matrix, representing all the OD pairs with demands, is required along with the network representation (usually a capacitated network). The output is an estimate of the traffic flows on each link and, consequently, an estimation of the travel time on each link. The first attempts to tackle the traffic assignment problem were during a just after the World War II. The first proposed assignment was the so-called *all-or-nothing assignment* as proposed in Campbell (1950). Since the main assumption of the all-or-nothing assignment is that the travel time does not depend on the flow in the links, all the demand of an OD pair is assigned totally to the shortest path for that OD pair.

After the traffic research community realised that the all-or-nothing assignment was not realistic, they tried to take into account congestion effects in routing vehicles. The result is the so-called *latency or link performance function*, i.e. a function in which link travel time depends on the number of vehicles using the link. According to Sheffi (1985), in a traffic assignment problem the set of constraints specifies that the demand of all the OD pair has to be satisfied, the flows has to be non-negative, the road segment utilization is the sum of all the flows traversing that road segment. Each road segment $a \in A$ is associated with its latency function $t_a = t_a(x_a)$ where x_a represent the total flow of vehicles (or the entering rate) on the road segment a . The latency function is usually assumed convex and non-decreasing. In the literature, several latency functions have been proposed. A survey on the used latency function in the literature is proposed in Branston (1976). Below the most used latency functions are listed:

- $t = t_0 e^{\frac{x}{c}}$
- $t = t_0 \alpha^{\beta \frac{x}{c}}$ where α and β are parameters.
- $t = t_0 [1 + \alpha (\frac{x}{c})^\beta]$ where α and β are parameters (BPR).
- $t = \begin{cases} \frac{d}{S_0} & x \geq \delta \\ \frac{d}{S(x)} & x \leq \delta \end{cases}$ where d is the distance, S_0 is the free-flow speed and $S(x)$ the speed experienced with flow greater than δ . δ should be considered as the congestion threshold.

The most used latency function in literature is the one proposed by the Bureau of Public Roads (BPR) with $\alpha = 0.15, \beta = 4$ and c is the road segment capacity.

The objective function of the traffic assignment depends on the kind of equilibrium the problem is trying to achieve. The path traversing time depends on the number of vehicles that are flowing through the road segments belonging to the path. The latency on a path is usually defined as the sum over all the road segments in the considered path of all the road segment latency function value under current conditions.

In the wake of capturing congestion effects in traffic assignment models, in 1952 the two Wardrop principles on flow distribution have been stated (see Wardrop (1952a) and Wardrop (1952b) for details). The first one is called *user equilibrium*, and it is based on the assumption that all users are in equilibrium, i.e. no one is willing to change its own route since there are no faster routes on the network. The second is called *system optimum* and it is based on the assumption that the total travel time is minimized and all drivers comply with the guidance prescriptions. In Beckmann et al. (1956), the mathematical models for the traffic assignment have been developed in forms of a convex non-linear optimization problem with linear constraints. Subsequently, in Frank and Wolfe (1956), an iterative algorithm to solve the quadratic optimization problems has been presented. More precisely, the solution algorithm is a combination of the Frank-Wolfe algorithm and traffic assignment models and it leads to a powerful and effective method for solving the traffic assignment problem alternating the all-or-nothing assignment with a line search approach. This method is nowadays used for the solution of many traffic assignment models.

4.1 User equilibrium and system optimum

When all drivers individually decide the route they will use in travelling from origin to destination, there are no drivers that can unilaterally choose another route because all used route from an origin to a destination are characterized by the same average travel time. This is because each driver decides to use the least duration path and, at the end, all routes have the same travelling time. This equilibrium situation is called *user equilibrium*. The user equilibrium have some underlying assumptions: the drivers have complete information about the available paths and the network flows are stable over time. According to Sheffi (1985), the user equilibrium model is the following:

$$\min \sum_{(ij) \in A} \int_0^{x_{ij}} t_{ij}(\omega) d\omega$$

$$x_{ij} = \sum_{c \in C} \sum_{k \in K_c} a_{ij}^{kc} y_c^k \quad \forall (i, j) \in A \quad (1)$$

$$d_c = \sum_{k \in K_c} y_c^k \quad \forall c \in C \quad (2)$$

$$x_{ij} \geq 0 \quad \forall (i, j) \in A \quad (3)$$

$$y_c^k \geq 0 \quad \forall c \in C \quad \forall k \in K_c. \quad (4)$$

The objective function is the sum over all road segments of the integral between 0 and the link flow of the road segment latency function. This expression has no

economical meaning, but it is only a mathematical construction, as stated in Sheffi (1985). The road network is represented through a graph $G = (V, A)$ where A is the set of road segments and V is the set of intersection points. Variables x_{ij} represent the flow on each road segment (i, j) while y_c^k represent the flow of the OD pair c on path k . Constraints (1) bounds the arc flows x_{ij} to the flow of the paths passing through the road segment (i, j) using the incidence matrix a_{ij}^{kc} . Constraints (2) guarantee to the demand satisfaction. In Sheffi (1985), proof of existence and uniqueness of the user equilibrium are provided. It is provided also a proof of the correspondence between the user equilibrium definition and the proposed model. In Lujak et al. (2015), it is also proved that this equilibrium corresponds to a Nash equilibrium in a game with a large amount of players. The stochastic version of the user equilibrium, in which the drivers are assumed to have incomplete information and to be not completely rational, is provided in Sheffi (1985). Well-known modifications are expected-utility-theory based models Mirchandani and Soroush (1987), the travel time budget model Lo et al. (2006), the late arrival penalty model Watling (2006) and the prospect-based UE model Xu et al. (2011) e and Avineri (2006). Fuzzy drivers decisions, which fuzziness is due to their perceived travel time, is provided in Ramazani et al. (2011), in Miralinaghi et al. (2016) and references therein.

When all drivers act together in such a way the total travel time is minimized, we are facing a *system optimum* solution. System optimum occurs when the sum of the latency experienced by all the users is minimized. According to Sheffi (1985), the system optimum model is the following:

$$\min \sum_{(ij) \in A} x_{ij} t_{ij}(x)$$

$$x_{ij} = \sum_{c \in C} \sum_{k \in K_c} a_{ij}^{kc} y_c^k \quad \forall (i, j) \in A \tag{5}$$

$$d_c = \sum_{k \in K_c} y_c^k \quad \forall c \in C \tag{6}$$

$$x_{ij} \geq 0 \quad \forall (i, j) \in A \tag{7}$$

$$y_c^k \geq 0 \quad \forall c \in C \quad \forall k \in K_c. \tag{8}$$

The objective function is the sum over all road segments of the road segment latency function multiplied by the flow on the road segment. Constraint set and variables are the same of the user equilibrium formulation.

4.2 The price of anarchy: the gap between UE and SO

Achieving the user equilibrium does not imply that the total travel time is minimized as in the system optimum. In fact, in most cases, as shown in Harks et al. (2015) and in Jahn et al. (2005), the inefficiency of the equilibrium can be measured. This measure is the **price of anarchy** and it is defined as the worst-case ratio of the cost of an equilibrium (in terms of total travel time of all the drivers) over the cost under a system-optimum. An example of the price of anarchy was first provided by Pigou.

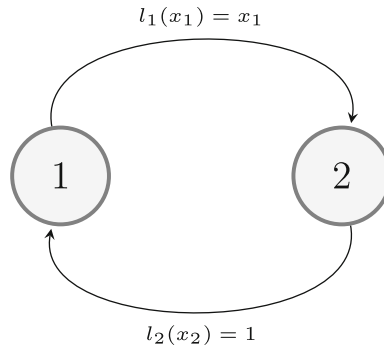


Fig. 1 Pigou's example

The Pigou's example network is depicted in Fig. 1 and example description follows. One unit of traffic wants to travel from Node 1 to Node 2 and travel times on the two road segments, upper and lower road segment, are depicted in Fig. 1 according to the functions $l_1(x_1)$ depending on flow x_1 and $l_2(x_2)$ depending on flow x_2 . It is easy to see that the travel time from Node 1 to Node 2 on upper arc depends linearly on the flow x while the travel time on the lower road segment is always equal to 1, regardless the flow sent on the road segment. According to the UE definition, the flow x has to be split on the two road segments, with flow x_1 and x_2 , in such a way no portion of the flow is envious, i.e. there are no better paths in terms of travel time on the network. The UE equilibrium is attained when the entire flow is route on the upper road segment and the total travel time is $x_1 l_1(x_1) + x_2 l_2(x_2) = x_1^2 + x_2 \cdot 1 = 1^2 + 0 = 1$. On the other hand, spitting the flow x in two equal parts, i.e. $x_1 = x_2 = \frac{1}{2}$, the total travel time is $x_1 l_1(x_1) + x_2 l_2(x_2) = x_1^2 + x_2 \cdot 1 = \frac{1}{2}^2 + \frac{1}{2} \cdot 1 = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} < 1$. In fact, this flow assignment is the so-called SO. Note that this assignment is not valid as UE since travellers on the lower road segment are experiencing a travel time which is much higher than the one on the upper road segment, in fact doubled. The price of anarchy in the Pigou's example is $\frac{1}{\frac{3}{4}} = \frac{4}{3}$.

Literature on the price of anarchy is wide and bounds has been found for affine and non-negative coefficient polynomial road segment latency functions. Considering an instance with latency function l drawn from a family L of non-decreasing continuous functions, the price of anarchy is bounded from above by $\alpha(L)$, i.e. $\sum_a x_a^{UE} t_a(x^{UE}) \leq \alpha(L) \sum_a x_a^{SO} t_a(x^{SO})$. In a single commodity case and for linear functions, the price of anarchy is $\alpha(L) = \frac{4}{3}$, as assessed in Roughgarden and Tardos (2002) where bounds for other function families has been also derived. In fact, this is the bound attained in the Pigou's example. In general, for generic function families, the price of anarchy is bounded to be $\alpha(L) = \Theta\left(\frac{p}{\ln p}\right)$, as assessed in Correa et al. (2007). In O'Hare et al. (2016), a thorough study on how the demand magnitude impacts on price of anarchy has been conducted. They identified four empirical rules that leads to an increase of the price of anarchy and numerical evidences show that the price of anarchy follows a power law decay for large demands. Most recent findings on the magnitude of the price of anarchy show its dependency on the flow magnitude, as in Correa et al. (2019) and

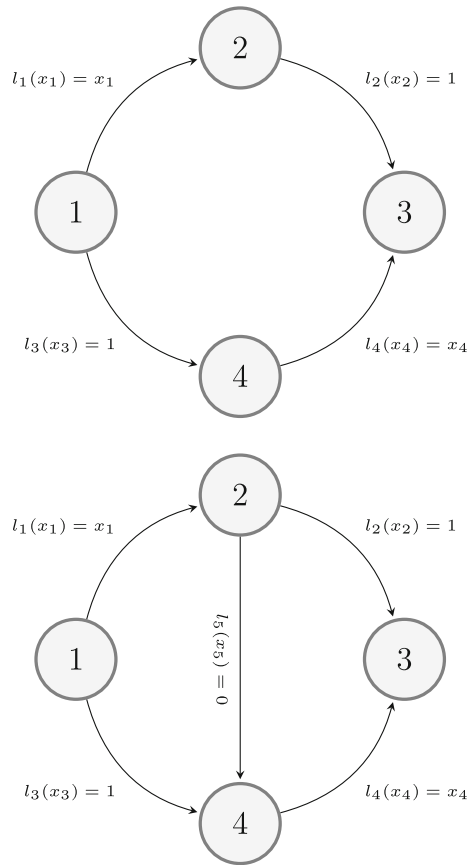
in Colini-Baldeschi et al. (2020). The former focuses on bound for the price of anarchy while the latter confirm that the price of anarchy follows a power law with respect to the magnitude of the flow and they stated that, given polynomial latency functions, it can be explicitly evaluated. A micro-simulation framework, embedding features of traffic such as reaction time, acceleration, deceleration, aggressiveness and many others, is provided in Belov et al. (2021). The results show that the price of anarchy embedding such features can be much higher than the theoretical one. A thorough discussion on the impact of network topology on the price of anarchy is provided in Roughgarden (2003) where the independence of the worst-case price of anarchy from the network topology is shown.

The **maximum latency price of anarchy** is an alternative way to measure the price of selfish routing. In Lin et al. (2011), the price of selfish routing with respect to the maximum latency experienced by a user is studied. In other words, the user equilibrium total travel time is compared to the total travel time of a min-max latency model. The min-max latency model is a model that minimizes the maximum latency over all experienced paths under the same constraints of the user equilibrium. In Bayram et al. (2015a) the maximum latency is used as a measure of unfairness. Bounds on maximum latency price of anarchy bounds have been also derived. In Correa et al. (2007), it is proved that, even for linear latency functions, the maximum latency price of anarchy can be unbounded.

A further evidence about the price of anarchy and its relationship with the underlying network is given by the **Braess's paradox**. In 1968, Braess proposed an example in which the system optimum does not equate with the best overall selfish flow through a network. The Braess's paradox is stated as follows: "For each point of a road network, let there be given the number of cars starting from it, and the destination of the cars. Under these conditions, one wishes to estimate the distribution of traffic flow. Whether one street is preferable to another depends not only on the quality of the road, but also on the magnitude of the flow. If every driver takes the path that looks most favourable to him, the resultant running times need not be minimal. Furthermore, it is indicated by an example that an extension of the road network may cause a redistribution of the traffic that results in longer individual running times".

The Braess networks before and after the network expansion are, respectively, depicted in Fig. 2. One unit of flow has to be routed from Node 1 to Node 3. The network after differs from the network before only by having added a new road segment from Node 2 to Node 4 with constant latency function equal to 0. This change may appear as irrelevant since the travel time on road segment is equal to zero. However, the situation dramatically change. In fact, the UE assignment on the before network is halved in the two feasible paths, i.e. 1-2-3 and 1-4-3. This means that the resulting flows are $x_1 = x_2 = x_3 = x_4 = \frac{1}{2}$ and the total travel time is $x_1l_1(x_1) + x_2l_2(x_2) + x_3l_3(x_3) + x_4l_4(x_4) = \frac{1}{2}1 + \frac{1}{2}1 + \frac{1}{2}1 + \frac{1}{2} + \frac{1}{2} = \frac{1}{4} + \frac{1}{2} + \frac{1}{2} + \frac{1}{4} = \frac{3}{2}$. Interesting enough, this assignment corresponds to the system optimum and, hence, the network before expansion is not affected by the price of anarchy. However, the network after the expansion does. In fact, there exists a new path from Node 1 to Node 2 that results to be cheaper in terms of travel time and so flows are no more in equilibrium. The UE, in this case, is attained when the entire flow is routed on path 1-2-4-3 with an

Fig. 2 Braess's example—
Network before and after



enormous increase in terms of total travel time, $x_1 l_1(x_1) + x_2 l_2(x_2) + x_3 l_3(x_3) + x_4 l_4(x_4) + x_5 l_5(x_5) = 1 + 0 + 0 + 1 + 0 = 2$. Note that the system optimum on the network after expansion remains the one we had in the network before expansion.

Examples and real world instances of Braess's paradox are shown in Sheffi (1985) and in Youn et al. (2008) where such situations have been detected in big cities as London and New York.

The price of anarchy induced by Braess's paradox is the so-called **Braess's ratio**. Let $L_i(G)$ is the common latency of the i -th OD pair on a graph G . Let $H \subseteq G$ a subgraph obtained removing road segments from G paying attention in having at least a path for each OD pair. Braess's ratio is: $\beta(G) = \max_{H \subseteq G} \min_{i=1, \dots, k} \frac{L_i(G)}{L_i(H)}$. In Lin et al. (2011), it is shown that the maximum latency price of anarchy is an upper bound for the Braess's ratio. In Roughgarden (2006), a wide study on how to remove Braess' paradox phenomena from networks is provided. They proved that there are no approximation algorithms under a precision threshold to detect and eliminate the paradox and, thus, an efficient detectability is not possible.

Braess' paradox has been also studied with flows over time in Macko et al. (2013) and according to the traffic assignment with flows over time proposed in Koch and

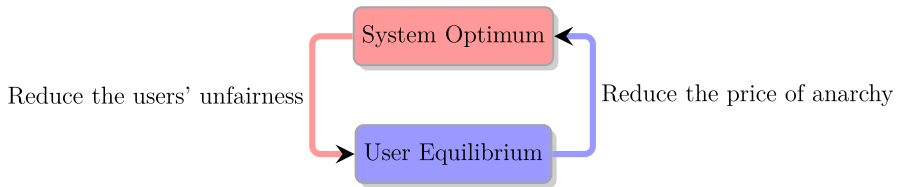


Fig. 3 Bridging UE/SO scheme

Skutella (2011). They compared the stationary flow Braess' paradox, as the provided example, with the one over time and they have provided examples in which the Braess' paradox appears only when the flow over time is considered. In Akamatsu and Heydecker (2003), the Braess' paradox over time is also studied examining different example networks and queuing patterns in which the paradox is unavoidable.

Research directions and further insights on the price of anarchy can be found in Roughgarden (2008).

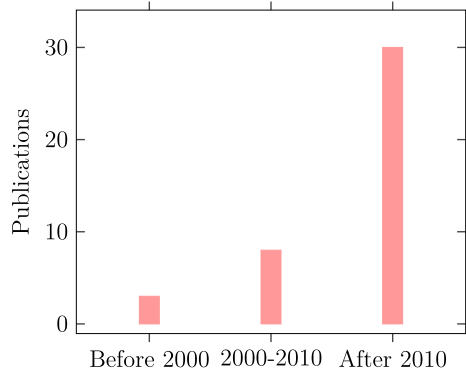
5 Bridging the user equilibrium and the system optimum: methodologies and related approaches

As pointed out in the introduction, models bridging the user equilibrium and the system optimum have received great attention during the last years, as assessed by numbers in Fig. 4, where it is clearly shown that the number of publications in the last ten years is considerably increased. The approaches bridging the user equilibrium and the system optimum are mostly divided into two branches: introducing users' fairness constraints in the system optimum and relaxing the requirements of the user equilibrium. As shown in Fig. 3, from the system optimum we can achieve the user equilibrium by imposing a certain level of fairness among users. On the other hand, the system optimum could be achieved from the user equilibrium by relaxing the fairness requirements imposed by the pure user equilibrium. This review accounts for these distinctions by, firstly, showing the most used fairness and efficiency measures in 5.1. Then, the state-of-the-art of system optimal models with users' constraints is presented in Sect. 5.2 and the state-of-the-art of relaxed user equilibrium models in Sect. 5.3. Section 5.4 contains other attempts aiming at inducing a social optimum from a user equilibrium such as the game theory concept of Stackelberg routing, dedicated lanes and tradable credit schemes and congestion charging mechanisms.

5.1 Fairness and efficiency measures

The concept of fairness in traffic assignment were firstly introduced in Jahn et al. (2005) to measure the efficiency of seeking near system optimum traffic patterns without losing the fairness property. With this aim, they introduce several notions of unfairness of a solution such as the loaded unfairness, the normal unfairness, the UE unfairness and the free-flow unfairness. These measures compare the travel times resulting from the used assigned with off-line travel time measures. The loaded unfairness is the

Fig. 4 Number of contributions in bridging the UE and SO through the years



maximum over travellers of the ratio of experienced travel time to the experienced travel time of the fastest traveler for the same OD pair. The normal unfairness is the maximum over travellers of the ratio of a path travel time and the shortest path travel time for the same OD pair, both under free-flow conditions. The UE unfairness is the maximum over travellers of the ratio of experienced travel time to the travel time for the same OD pair in a user equilibrium. The Free-flow unfairness is the maximum over travellers of the ratio of experienced travel time to the fastest path travel time for the same OD pair under free-flow conditions. These measures are also used within evacuation models to measure shelter proximity and evacuation rapidity, as in Bayram et al. (2015b).

5.2 System optimal traffic assignment with users' constraints

The system optimal traffic assignment with users' constraints is a system-optimum traffic assignment problem in which constraints on experienced unfairness among users are introduced in the formulation. These constraints are called side constraints and allow to take into account user-friendly additional restrictions. In fact, the system optimum is difficult to implement for real-world networks because it could be very unfair with a subset of users. As for the user equilibrium, system optimal traffic assignment with users' constraints pays a price of anarchy which is, in general, strongly reduced with respect to the user equilibrium one depending on the tightness of the side constraints. On the other hand, the system optimum could lead to unacceptably long paths for some drivers while one with side constraints can help in reducing the users' unfairness and in enhancing the compliance. To have an idea of potential savings with cooperative policies, results on the upper and lower bound on the total travel time are shown in Feldmann et al. (2003).

When side constraints are referred to users' travel time and/or path length, the approach is called constrained system optimum and it is the most prolific research area in the field of system optimal traffic assignment with users' constraints. The idea underlying the constrained system optimum is to propose a little sacrifice in terms of length or travel time to some drivers in order to improve congestion on the whole network. In constrained system optimum, for each OD pairs, a feasible paths

set is generated containing only those paths that are not longer/slower than a fixed percentage of the shortest/fastest path for the OD pair. In order to measure how fast a path is, the generating path algorithm uses the so-called normal length, which is an a priori estimate of the real travel time. The path set will be constructed taking only those paths that have normal length shorter than the OD pair shortest path normal length multiplied by a percentage. Usually the road segment length, the road segment free-flow travel time and the road segment travel time under user equilibrium are used as normal length measure.

To the best of our knowledge the first attempt to tackle a constrained system optimum has been developed in Jahn et al. (2000). In this work traffic flows are routed through a road network in such a way the total road usage is minimized while proposing to users only those paths that are not too long in terms of geographical length, as in Möhring (2013). The proposed model is formulated as a non-linear multi OD-pair flow problem. They use, as latency function, the Davidson's function $t_a(x) = t_a^{freeflow} + \frac{\alpha x_a}{u'_a - x_a}$ where α is a tuning parameter and u'_a a parameter chosen in such a way $u'_a > u_a$. As solution method, they propose the Frank-Wolfe algorithm where, in order to search a feasible direction, a linearization of the non-linear problem is used. Eligible paths are generated using a column generation technique. They pointed out that, even in the linearized version, the problem of finding flows on a network, in such a way the total travel time is minimized and the followed paths are not longer than a threshold, is known to be NP-hard.

The impact of choosing a constrained system optimum traffic assignment is widely explained in Schulz and Stier-Moses (2006) where a theoretical work on the efficiency and fairness is proposed. They measure its efficiency by comparing the output of the constrained system optimum with the best solution without guidance and to the user equilibrium while the unfairness is measured comparing travel times of different users. They measure and prove upper bounds of unfairness and efficiency of the constrained system optimum considering different classes of latency functions (affine, non-decreasing differentiable, etc.). One interesting features of this work is that they compare results using, as normal length, either the free-flow travel time and the travel time under user equilibrium conditions. They pointed out that the use of the travel time under user equilibrium conditions as normal length is more reliable a priori estimate of the travel times since they depend also on the traffic flow that intends to travel on the network. According to this modelling choice, Jahn et al. (2005) proposes a constrained system optimum model and methodology that involves using the most commonly used latency function provided from the USA Bureau of Public Road, $t_a(x) = t_a^{UE} [1 + 0.15(\frac{x}{u_a})^4]$ where t_a^{UE} represents the road segment travel time under user equilibrium conditions. They propose as measure of the unfairness the comparison between experienced travel times with the best travel times, with the free-flow travel times and with the travel time under user equilibrium. As methodology, they propose a variant of the Frank-Wolfe algorithm with a column generation technique in the linearized sub-problem. They provide a wide computational study in which they test the model on seven real road networks where demands are generated using estimations of real data. They show that using a constrained system optimum the unfairness experienced by users (considering all the unfairness measures) is small.

In Correa et al. (2007), a variant of the constrained system optimum with the minimization of the maximum latency is proposed. They propose results on the unfairness and they compare results obtained using either the maximum latency and the total latency as objective value. They show that a flow optimal for the total latency is near-optimal with respect to the minimum maximum latency and it is quite fair and also minimizing with respect to maximum latency produce an optimal solution that is within a constant factor with respect to the optimal solution produced with the total latency. One of the reasons motivating this work is the study of the bottlenecks where a minimum maximum latency level has to be guaranteed in order to avoid the typical phenomena related to bottleneck congestion. Theoretical bounds on the objection function value using different latency functions are derived in Schulz and Stier-Moses (2006) for the traditional constrained system optimum while theoretical bounds when the maximum latency is minimized are provided in Feldmann et al. (2003).

The first attempt to use a linear programming model to solve the constrained system optimum traffic assignment problem is presented in Angelelli et al. (2016). The proposed approach is hierarchical. First, a linear programming model is run in order to lower the maximum congestion level on the network and, then, a second linear programming model is run to route drivers on fair paths without exceeding the maximum congestion level found in the first model. They show that it is always possible to lower the maximum congestion level with a lower level of overall experienced unfairness. In details, the second model minimizes the total travel time on selected paths while keeping the network non-congested, if possible, or at its minimum congestion level, otherwise. The set of eligible paths is generated a priori as for Jahn et al. (2005). As the number of paths generated a priori is, in the worst case, exponential in the instance size, in Angelelli et al. (2018) a column generation heuristic algorithm is proposed. They show that the algorithm returns a solution which is very near to the optimal one while having very short computational time even on big networks. In Angelelli et al. (2020a), a linear programming model with a traffic-dependent latency function is presented. More specifically, the model makes use of a piecewise linear approximation of the convex latency function. Here, the BPR latency function is embedded in the linear programming formulation. Thorough computational results assess the potentiality of using the linear formulation that allows to take advantage of extremely powerful linear commercial solvers. The linear programming model is able to provide solution in big road networks and, for very big road networks, two heuristic algorithms obtaining excellent results are provided.

Models analyzed so far are devoted to the minimization of the total travel time and to the minimization of the maximum latency. Both perspectives are of interest, but there could be heavily congested road segments, that are not the worst case in terms of road segment congestion level, neglected by the former objective and maybe they could be the goals a traffic planner aim to achieve. Considering only the maximum latency, the model will not consider the average value of the additional travel time on the road segment and the total travel time could be very bad. On the other end, minimizing the total travel time is totally blind from the point of view of the variability among different road segments. To this aim, in Angelelli et al. (2019), a constrained system optimum model able to control the right tail of the distribution of congestion on road segments. They show that the obtained assignment produces almost the same total

travel time of the traditional constrained system optimum model while guaranteeing a very good level of fairness in the spreading of congestion over the road segments of the network.

The classical constrained system optimum formulations aim at minimizing the total travel time while guaranteeing a given fairness level among users but, since eligible paths are generated a priori, the level of the experienced unfairness could turn out to be higher than the imposed level. Even though in most cases, these eligible paths are the ones involved in the final solution, the choice is not made on the basis of the real flow and, hence, some useful paths can be missed. In order to overcome the drawbacks of the current state-of-the-art traffic assignment models, in Angelelli et al. (2020b) two constrained system optimum formulations are provided where the path selection is embedded into the formulations and, thus, the real experienced unfairness is directly controlled inside the model. In Angelelli et al. (2020b) the benefits achieved with the new modelling choice are shown and explained also through formal properties.

Constrained system optimum has been also applied to pedestrian flows in urban areas in DalSasso and Morandi (2021). The aim is to route pedestrians to minimize the system goal while routing them on paths that no longer than a certain percentage of the shortest path. Given the different nature of the problem at hand, the system objective focuses on reduce gathering phenomena both on roads and crossroads.

In Li and Zhao (2008) and in Zhenlong and Xiaohua (2008) a double objective related to the constrained system optimum achievement is proposed. In fact, it is not a constrained system optimum, but a trade-off between system optimum and user equilibrium called integrated-equilibrium routing problem. It calculates the system total travel time under SO, T_{SO} , and the travel time for each user under user equilibrium t_{UE} . Then, the first objective Z_1 is the classic system optimum one and the second objective Z_2 is the classic user equilibrium one. Constraints set is as usual plus a constraint on the system optimum $Z_1 \leq T_{SO} + \epsilon$ and a set of constraints, one for each user, $Z_2 \leq t_{UE} + \xi$ where ϵ and ξ are functions respectively of T_{SO} and t_{UE} .

Attempts to bridge the system optimum with the user equilibrium have been done also in field of agent-based models and/or simulations. In Lujak et al. (2015), an agent-based model is proposed that uses a new set of constraints in which the unfairness is bounded by a no-envy criteria between users. In Lujak et al. (2015), the concept of normalized mean path duration is introduced as the geometric mean of the flow on a path multiplied by the number of driver using that path for each commodity. In mathematical notation, it is $\gamma_c = \sqrt[|P_c|]{\prod_{p \in P_c} f^p x^p}$ where c is the commodity, P_c its path set, f^p is the latency function of path p and x^p is the flow on path p . The no-envy criteria for each commodity is: $\gamma_c \geq \gamma_{c'}^\alpha$ with $0 \leq \alpha \leq 1$ and $\forall c \in C, c' \in C, c' \neq c$. Other examples can be found in Levy et al. (2017), in Klein et al. (2018) and in Levy et al. (2018) where other agent-based models on route-choice games are presented. Among agent-based simulations, a social routing framework has been proposed in Van Essen et al. (2019) that aims at attaining a pseudo-system optimum traffic assignment through assigning path to users so as to minimize the total travel time plus the marginal cost for compliant users that are routed on paths that are sub-optimal with respect to the fastest path on the network. They divided the possible scenarios into selfish, social and mixed scenarios based on the fraction of the demand

which is willing to follow the pro-social instructions and, hence, to follow path that are different from the fastest one. They found out that savings in terms of total travel time are remarkable. In fact, as assessed by Djavadian et al. (2014), some drivers are more willing to follow the guidance instruction in order to enhance the system benefit. Hence, they divided the set of drivers accordingly.

Besides traffic assignments models, there exist in literature other lines of research that can be easily associated with system optimum with users' constraints. In the following, some examples are provided. The constrained system optimum, as proposed in Jahn et al. (2005), has been successful used also for managing traffic under emergencies and/or natural catastrophes. A constrained system optimum approach has been presented also in Bayram et al. (2015a), where the model is applied to shelters location and evacuation planning in disasters' management. The approach first assigns users to shelters and, then, users are assigned to the shortest path to their shelter with a given degree of tolerance. The set of feasible paths is determined as in Jahn et al. (2005). A second order cone programming technique is used to efficiently solve the problem. In Bayram et al. (2015a), the total evacuation time is minimized under optimal location of the shelters. Since, also in evacuation problems, the traversing time depends on how many vehicles are standing on a road segment, they proposed to apply a constrained system optimum model to the problem at study. A stochastic version of the problem with different scenarios is tackled in Bayram and Yaman (2018a) while a Bender's decomposition approach is proposed in Bayram and Yaman (2018b). In Bayram (2016), a very comprehensive review on traffic assignment models for evacuation planning is also provided. In Yuan et al. (2019) a constrained system optimum is proposed to evacuate areas in a secure and stable way. They define the evacuation time as the time needed from the start of the emergence to reach the evacuee secure area and they bound the evacuation time of all users to be within a tolerance factor of the fastest user evacuation time.

Paths selection is not always related to the path length or duration. In Cortés et al. (2013) a set of diversification constraints is provided. These constraints allow to demand to be sent on at least a predefined number of paths or a number of arc-disjointed paths, i.e. paths without road segments in common. One particular difference between this work and the others is that the cost function is concave. That is because the problem regards supply chain management and the curve represents scale economies effect in transportation problems (See references in Cortés et al. (2013) for details). An optimal iterative algorithm based on the Kuhn-Tucker optimality conditions is provided.

On the network design side, the issue of fairness, here called equity, is considered when new infrastructures have to be constructed and the consequent effect on traffic flows have to be evaluated; see Yang and H. Bell (1998), Patriksson (2008), Liu and Wang (2015), Meng et al. (2001), Yang and Zhang (2002) and Guo and Yang (2009) for details and references therein.

Another field of application are the communication networks where the constrained system optimum model is also widely used. We report an example. In Holmberg and Yuan (2003) the main issue is to avoid paths with high dispatching delays. The time delay is calculated by summing up the estimated link delay of each road segment that belongs to the considered path. For this reason a limit on the cost per unit of flow on each path is calculated. These limits can also include distortion on the network and link

failure. A system optimum objective function is used and the constraints set considers only paths which weight is less or equal to the bound due to link delays on that path. The result is a classical constrained system-optimum. Another little difference is in the forcing constraints, i.e. the ones that say if a path is used or not. Since binary variables are used to recognize if a path is used or not, a relaxation is proposed. The problem is solved with a column generation technique.

Table 1 Quick reference to literature on system optimal assignments with users' constraints

Topic	Reference	Approach
CSO in traffic assignment	Jahn et al. (2000)	CSO with paths constrained in length
	Feldmann et al. (2003)	Theoretical bounds minimizing the maximum latency
	Jahn et al. (2005)	CSO with paths constrained using the user equilibrium travel time
	Schulz and Stier-Moses (2006)	bounds for the general CSO
	Correa et al. (2007)	Variant of the CSO minimizing the maximum latency
	Angelelli et al. (2016)	Linear hierarchical CSO with paths constrained in length
	Angelelli et al. (2018)	Fast large-scale heuristic for Angelelli et al. (2016)
	Angelelli et al. (2019)	Linear CSO minimizing congestion on the most congested road segments
	Angelelli et al. (2020a)	Linear CSO with UE constrained paths and large-scale heuristic
Angelelli et al. (2020b)	CSO bounding real unfairness experienced with loaded road segments	
Integrated equilibrium	Li and Zhao (2008)	Integrated equilibrium routing problem
	Cortés et al. (2013)	CSO models for large-scale evacuation problems
Agent-based social routing	Lujak et al. (2015)	Agent based CSO embedding no envy criteria
	Van Essen et al. (2019)	Agent based CSO adding the unfairness marginal cost
CSO in emergency evacuation	Bayram et al. (2015a)	CSO in shelter location and evacuation problems
	Bayram (2016)	CSO models for large-scale evacuation problems
	Bayram and Yaman (2018a)	CSO in shelter location and evacuation problems under uncertainty
	Yuan et al. (2019)	CSO in evacuation strategy optimization

Literature on system optimal assignments with users' constraints is summarized in Table 1.

5.3 Relaxations of the user equilibrium

The relaxation of the user equilibrium conditions has been studied through years from several perspectives. The relaxation of the user equilibrium conditions goes mainly through limiting the set of feasible path from an origin to destination to those that no longer than the fastest path on the network (constrained user equilibrium) and through the definition of an indifference band that refer to users' behaviour in perceiving differences among assigned paths, i.e. the so-called bounded rational user equilibrium.

The constrained user equilibrium (briefly CUE) is a user equilibrium in which the path set is limited to the paths that meet some restrictions on length, travel time or other parameters. An example of constrained user equilibrium is provided in Zhou and Li (2012), where the path set is restricted only to those paths that are shorter than a fixed threshold. They use, in order to compare paths, the path euclidean length and they label as feasible a path only if it is shorter than the shortest path from the origin to the destination multiplied by a scaling factor greater than 1. A path-based formulation of the user equilibrium, in which only feasible paths are allowed, is provided and a column generation technique combined with the Frank-Wolfe algorithm is proposed. They use as latency function the usual one provided by the USA Bureau of Public Road. They propose also a length constrained system optimum formulation with the same path set used for the constrained user equilibrium and they use it in order to formulate an optimal road pricing scheme able to achieve the constrained system optimum solution.

These concepts fully apply when electric vehicles are involved in the traffic assignment. This is because the vehicle battery have its own duration and the trip cannot last more than a fixed threshold. In Jiang et al. (2012), a distance constrained user equilibrium with penalties in the case in which the threshold is exceeded is provided. Further extension related to needs of electric vehicles are provided in Jiang and Xie (2014) and in Jiang et al. (2014).

The bounded rational user equilibrium is the main approach that modifies the user equilibrium. It does not aim directly at reducing the price of anarchy and, hence, bridging the UE to the system optimum. In fact, in bounded rational user equilibrium, drivers follow their own perception, as for the classical user equilibrium. However, the relaxation of the classical user equilibrium, could lead to a reduction in terms of the price of anarchy provided that, among all feasible assignments, the one with minimum total travel time is chosen. The bounded rational user equilibrium was first proposed in Mahmassani and Chang (1987) as a relaxation of the user equilibrium in which a path can be used only if its path traversal time is within a range with respect to the fastest path on the network (see Zhang (2011) for further details). This range is called indifference band and is usually obtained by means of road user behavioral studies or empirical observations. Moreover, this indifference band could be calibrated depending on the OD pair to which is assigned. The concept of indifference band has been successfully embedded in traffic simulations (see Jayakrishnan et al. (1994),

Table 2 Quick reference to literature on relaxations of the user equilibrium

Topic	Reference	Approach
CUE	Zhang (2011)	User equilibrium under different behaviour assumptions
	Zhou and Li (2012)	CUE with paths constrained in length
	Jiang et al. (2012)	CUE with penalties for paths exceeding the length threshold
	Jiang and Xie (2014)	CUE with mixed electric and gasoline vehicles
	Jiang et al. (2014)	CUE with mixed electric and gasoline vehicles and parking slots
BRUE	Mahmassani and Chang (1987)	First bounded rational user equilibrium model
	Lou et al. (2010)	Congestion pricing under bounded rational user equilibrium
	Di and Liu (2016)	Review on the bounded rational user equilibrium
	Ye and Yang (2017)	Adjustment processes for the bounded rational user equilibrium

Mahmassani and Jayakrishnan (1991), Hu and Mahmassani (1997) and Mahmassani and Liu (1999) for details) as a key element to bridge the user equilibrium to a more efficient assignment and, even if users will follow only their own perceptions, empirical evidences of the gain in terms of efficiency were described in Jou et al. (2005), Jou et al. (2010) and later in Di et al. (2017). Mathematical properties and the behaviour of the price of anarchy under bounded rational user equilibrium can be found in Lou et al. (2010). Further mathematical properties are derived in Di et al. (2013), Di et al. (2014) and Di et al. (2016). A complete review on the applications of the bounded rational user equilibrium can be found in Di and Liu (2016). Recent attempts to include the bounded rational user equilibrium into traffic behavioural studies can be found in Ye and Yang (2017). Although the bounded rational user equilibrium ensures a certain level of fairness among users (considering traffic flows), it suffers from two shortcomings: the equilibrium is not unique (see Zhang (2011)) and it only aims at reaching an equilibrium state without minimizing the total travel time. Thus, no guarantee on the reduction of the price of anarchy can be derived. In fact, the user equilibrium solution is itself feasible for a bounded rational user equilibrium. A comprehensive review of bounded rational user equilibrium models in a dynamic setting can be found in Szeto et al. (2015).

Literature on relaxations of the user equilibrium is summarized in Table 2.

5.4 Other methodologies inducing social optima

Decision makers and administrators are always struggling in finding the best practice in terms of reducing congestion with an eye on budget and environmental constraints

that do not allow or partially allow enlarging the current infrastructures. At the same time, they are trying to enhance the use of low emission vehicles and the car sharing. To this aim, we report the state-of-the-art of two well-known approaches inducing social optima: the Stackelberg routing, the use of dedicated lanes and tradable schemes and the use of congestion charging mechanisms. Literature on methodologies inducing social optima is summarized in Table 3.

The Stackelberg routing In order to decrease the price of anarchy, individuals need some external steering in being cooperative, as they cannot identify socially desired alternatives themselves. To that end, travel information can be quite helpful. For instance, the Stackelberg routing, provided in Korilis et al. (1997), assigns a fraction of travellers by a central authority (i.e. leader) as they comply with advice that they received, while the remaining individuals (i.e. followers) choose their route selfishly (see Krichene et al. (2014) for further details). In Krichene et al. (2014), the leader anticipates on the (expected) selfish response in order to improve overall network performance. The Stackelberg routing has been proved to be effective, as shown in Bonifaci et al. (2010), where bounds on price of anarchy have been derived. Several modifications of the pure Stackelberg algorithm have been introduced in the last years as the introduction of tolls (see Swamy (2012) for details) or the imperfect knowledge of the duration on certain road segments (see Bhaskar et al. (2019) for details).

Dedicated lanes and tradable schemes The standard traffic equilibrium models can be bridged to a social optimum or to a side-constrained traffic flow patterns when specific methodologies, as dedicated lanes or tradable credit schemes, are applied. In Song et al. (2015), a pro-social mathematical model able to find the optimal locations of dedicated lanes is investigated aiming at optimize social benefits. They distinguish among high-occupancy vehicle lanes and high-occupancy toll lanes with the aim to reduce the traffic on certain road network areas. They also provide toll rates to be imposed in order to reach the desired result. In Chen et al. (2016) and in Esmaeilzadeh Seilabi et al. (2020), the concept of dedicated lanes has been extended to AVs (autonomous vehicles) and a time-dependent dedicated lanes deployment plan is proposed. The aim is to promote the use of AVs and to minimize the social cost. Beside the concept of dedicated lanes, the tradable credit scheme can be successfully embedded in traffic management. According to Yang and Wang (2011), in tradable credit schemes travelers need to pay credits in order to travel in the network. These credits are determined by a central authority on the basis of the actual demand for transportation and distributed to drivers. This provides to the central authority a mechanism for managing the traveler demand while achieving system-level goals. Travelers can trade credits amongst themselves. According to the authors, since no transfer of wealth will take place between the central authority and the travelers, there could be less societal objection to its implementation in practice. In Wang et al. (2012), the work has been extended to heterogeneous users in which the value of their time differs. A multi-period version has been proposed in Miralinaghi and Peeta (2016) while uncertainty issues are embedded in Shirmohammadi et al. (2013). Tradable credit schemes are also used in promoting low emissions vehicles, as in Miralinaghi and Peeta (2019). In Shirmohammadi and Yin (2016), the maximum queue in a bottleneck is studied when tradable credit schemes are implemented. Other works on tradable credit scheme can

be found in Wu et al. (2012), Miralinaghi and Peeta (2020) and Hosseinasab et al. (2018).

Congestion charging mechanisms Many cities have developed strategies for congestion reduction and to let drivers avoid entering overcrowded, or likely to be, areas. These strategies are mainly the congestion charging (see de Palma and Lindsey (2011) for details) or ramp metering schemes (see Kachroo and Özbay (2011) for details).

Road pricing and congestion charging can be developed in several ways. The first one is called facility-based and it regards tolling roads, bridges and tunnels only on a few facilities. It can be a single point toll or a distance-based toll. Another pricing scheme is called cordons. Tolls on cordons are an area-based charging method in which vehicles pay a toll to cross a cordon in the inbound or outbound direction or both. Another toll scheme is the zonal scheme, in which vehicles pay a fee to enter or exit a zone or to travel inside the zone. Some other schemes can be implemented using tolls proportional to distance. In de Palma and Lindsey (2011), some advice, on which toll scheme can be chosen depending on the case of study, are given and a good review on congestion pricing technologies is provided. Some works distinguish between congestion charging and road pricing. In Stopher (2004) it is considered as congestion charging a situation in which tolls are applied on an area that is most likely congested and, as road pricing, a situation in which tolls are distance-based. The latter is more fair than the former because tolls are spread along the journey in a progressive way, and they depend on how much the travel is long.

According to Lindsney and Verhoef (2001), early literature on congestion pricing was focused on the so-called first-best tolling, i.e. tolls exactly matches the external costs generated by each traveller. The name first-best comes from the fact that tolls are derived according to a first-best optimum in which the whole road network is used at maximum efficiency. The first-best tolling could be used as a theoretical value but, transportation community agree that it is of limited practical relevance. Recently, literature focuses to more realistic form of congestion pricing, the so-called second-best congestion pricing. Second-best tolling include a number of road pricing mechanisms that are dynamic and can be applied only where needed. The second-best tolling mechanisms can take into account many factors such as heterogeneity of users, social and political feasibility, fairness issues, etc. Recently, in De Palma et al. (2005), also a third-best tolling has been derived accounting for the so-called no-queue tolling, i.e. tolls that are imposed only when queues occur. A thorough formal description of the mathematical background of the road pricing theory, along with models and algorithms, is provided in Yang and Huang (2005) and in Lindsney and Verhoef (2001).

6 Criticisms and future research directions

The state-of-the-art of methods bridging the two Wardropian optima is relatively new in the traffic assignment research area, and it has been mostly applied in the static context, mainly because of the difficulty in embedding social and individual preferences within such models. However, the dynamic versions of the Wardropian optima are more

Table 3 Quick reference to literature on Stackelberg routing, dedicated lanes and tradable credit schemes and congestion charging

Topic	Reference	Approach
Stackelberg routing	Korilis et al. (1997)	Stackelberg routing for traffic assignment
	Krichene et al. (2014)	Stackelberg routing with behaviour predictions
	Bonifaci et al. (2010)	Price of anarchy under Stackelberg routing
	Swamy (2012)	Stackelberg routing with road pricing
	Bhaskar et al. (2019)	Stackelberg routing with incomplete information
Dedicated lanes	Song et al. (2015)	Optimal locations of dedicated lanes to reduce congestion
	Chen et al. (2016)	Optimal locations of dedicated lanes also for AVs
	Esmailzadeh Seilabi et al. (2020)	Optimal locations of dedicated lanes and tolling to minimize the total travel time
Tradable credit schemes	Yang and Wang (2011)	Managing network flows with tradable credit schemes with homogeneous drivers
	Wang et al. (2012)	Managing network flows with tradable credit schemes with heterogeneous drivers
	Miralinaghi and Peeta (2016)	Multi-period tradable credit schemes
	Miralinaghi and Peeta (2019)	Tradable credit schemes promoting zero-emissions vehicles
	Shirmohammadi and Yin (2016)	Tradable credit schemes to solve bottleneck congestion
	Shirmohammadi et al. (2013)	Tradable credit schemes with uncertainty issues
	Lindsney and Verhoef (2001)	First-best and second-best toll schemes
Congestion charging	Stopher (2004)	Distance-based toll schemes
	De Palma et al. (2005)	Queue toll schemes
	Yang and Huang (2005)	Foundations of road pricing techniques
	de Palma and Lindsey (2011)	Optimal toll schemes to reduce congestion

suitable to describe real-world situations and, hence, the main research area in which the social welfare and individual preferences should be embedded is surely the dynamic traffic assignment area. The reason is two-fold. On one side, the dynamic version is more suitable to describe the behaviour of the traffic flows when the assumption of a steady-state behaviour of traffic flows is no longer valid. In fact, the steady-state traffic flows assumption holds only during the peak hour on a macro level. Although

this is the period of time in which congestion is more likely to occur, there is a need to move to a less-than-macro level in order to take more reliable decisions. Secondly, including dynamics potentially opens a number of field of applications for the here described “fair” methodologies. One example could be embedding such efficiency and fair balanced considerations in real-time traffic guidance to avoid local bottlenecks just occurred. Most of the time, real-time sat-nav decisions are based on the actual state of the network and, in case of a bottleneck, they may return the same diversion for all drivers in that road segment. This could even worsen the situation, since the congestion is simply shifted to another road segment that could have less capacity. This is the main reason why a fair and efficient coordinated approach is of a great importance.

Another area which currently remains unexplored is the one related to users’ behaviour modelling. Most of the proposed models are based on simplified users’ satisfaction rules, such as having a travel time which is no longer than a certain percentage of the best possible travel time on the network or other simple rules. These are surely of importance for drivers, but some other factors could play a role in users’ satisfaction such as avoiding multiple changes of path (exiting and entering highways multiple times could be annoying) for the entire journey, the speed variability (traveling at a low stable speed is better than alternating queuing and free-flow) and many others. The literature lacks of proper users’ behaviour modelling both in static and dynamic traffic assignment when bridging the two Wardropian assignments.

The state-of-the-art literature mainly focuses on traffic assignment models with some nice examples of applications in evacuation planning. However, the field of application could be much broader. A very few examples of fair and efficient traffic assignment models can be found in the pedestrian routing field. In fact, in very crowded events or downtown areas, a lot can be done in order to improve the users’ experience when walking/visiting. Users’ could accept to follow a certain footpath as long as it is comfortable to walk in. In that sense, a latency function could also be provided for visiting a certain location along the way and requires dynamic modelling as the time factor is crucial. Another application is related to the big metropolitan train/tube stations in order to avoid bottlenecks in exit and enter gates. Furthermore, fair and efficient traffic assignment models could be used to suggest routes for commuters in multi-modal public transportation networks. It is well-known that usually there are many ways to commute from an origin to a destination, and many are almost equivalent for commuters. By coordinating commuters through equivalent paths, the central planner could achieve a balanced commuters’ distribution to avoid overcrowded means of transports.

Last but not least, these concepts could be of a great interests also for the logistic world. In a world in which we expect a parcel to be shipped within a day, balancing the system efficiency (in terms of time travelled by parcels) and the users’ satisfaction (receiving the parcel within reasonable time) becomes urgent and urgent in order to be competitive with the giants of the sector. In that case, efficiency could be measured in many ways but mostly in terms of global lead time while users’ satisfaction could be measured with respect to the best option available on the market.

The final outcome of the survey is, on one hand, the lack of literature aiming at bridging the user equilibrium and the system optimum in close-by research fields such as the dynamic and real-time time traffic assignment models and the users’ behaviour

modelling pointing out that there is huge room for future developments. On the other hand, many of the ideas presented in the literature can be successfully applied to other fields that are not simply vehicular traffic assignments. The review is designed be an useful tool for PhD students, researchers and practitioners that want to have a overview of techniques to balance the system welfare and users' compliance in transportation problems (vehicular networks, pedestrians, logistics, etc.). Moreover, it highlights that the field is getting attentions in the last years, especially with the advent of new information technologies and smart infrastructures.

7 Conclusions

The survey shows the potentiality of bridging the most well-known Wardrop's principles to efficiently model traffic assignment problems. This exciting research area has become even more interesting during the last years in which the literature has grown a lot. The proposed literature is surely valuable to traffic planners and it will become more and more interesting with the advent of autonomous vehicles and the advances in information technology. Many research questions remain open in many branches of the literature and fields of application can be widened to multidisciplinary approaches such as behavioural aspects and developments in ITS technologies. As a concluding remark, the survey opens the ideas implemented into traffic assignments to a wider audience. In the era of sharing economy, finding a way to satisfy the users while optimizing the system is no longer only a research question but a need. To this aim, the scope of the survey is also to provide tools to be applied in many other fields. In fact, the concept of balancing efficiency and fairness could be of interest for other research communities, such as pedestrian flows community, public transportation policies in delay management, simulations on infrastructures to be built, optimizing movements in big logistics hub, optimizing visitors trajectories in big over-crowded events and many others.

Acknowledgements The author expresses her appreciation for the comments and suggestions of anonymous reviewers which allowed them to substantially improve a previous version of this paper.

Funding Open access funding provided by Università degli Studi di Brescia within the CRUI-CARE Agreement.

Conflict of interest The author declares that she has no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akamatsu T, Heydecker B (2003) Detecting dynamic traffic assignment capacity paradoxes in saturated networks. *Transp Sci* 37:123–138
- Angelelli E, Arsik I, Morandi V, Savelsbergh M, Speranza MG (2016) Proactive route guidance to avoid congestion. *Transp Res Part B Methodol* 94:1–21
- Angelelli E, Morandi V, Speranza MG (2018) Congestion avoiding heuristic path generation for the proactive route guidance. *Comput Oper Res* 99:234–248
- Angelelli E, Morandi V, Speranza MG (2019) A trade-off between average and maximum arc congestion minimization in traffic assignment with user constraints. *Comput Oper Res* 110:88–100
- Angelelli E, Morandi V, Speranza MG (2020) Minimizing the total travel time with limited unfairness in traffic networks. *Comput Oper Res* 123:105016
- Angelelli E, Morandi V, Savelsbergh M, Speranza MG (2020a) System optimal routing of traffic flows with user constraints using linear programming. *Eur J Oper Res*
- Avineri E (2006) The effect of reference point on stochastic network equilibrium. *Transp Sci* 40(4):409–420
- Bayram V (2016) Optimization models for large scale network evacuation planning and management: a literature review. *Surv Oper Res Manag Sci* 21:63–84
- Bayram V, Yaman H (2018) A stochastic programming approach for shelter location and evacuation planning. *RAIRO-Oper Res* 52:779–805
- Bayram V, Yaman H (2018) Shelter location and evacuation route assignment under uncertainty: a benders decomposition approach. *Transp Sci* 52:416–436
- Bayram V, Tansel B, Yaman H (2015) Compromising system and user interests in shelter location and evacuation planning. *Transp Res Part B Methodol* 72:146–163
- Bayram V, Tansel BÇ, Yaman H (2015) Compromising system and user interests in shelter location and evacuation planning. *Transp Res Part B Methodol* 72:146–163
- Beckmann Martin, McGuire CB and Christopher BW (1956) *Studies in the economics of transportation*, Technical Report
- Belov A, Mattas K, Makridis M, Menendez M, Ciuffo B (2021) A microsimulation based analysis of the price of anarchy in traffic routing: the enhanced Braess network case. *J Intell Transp Syst* 1–16
- Ben-Elia E, Di Pace R, Bifulco GN, Shiftan Y (2013) The impact of travel informations accuracy on route-choice. *Transp Res Part C Emerging Technol* 26
- Bhaskar U, Ligett K, Schulman LJ, Swamy C (2019) Achieving target equilibria in network routing games without knowing the latency functions. *Games Econom Behav* 118:533–569
- Bonifaci V, Harks T, Schäfer G (2010) Stackelberg routing in arbitrary networks. *Math Oper Res* 35:330–346
- Boyce D, Lee D-H, Ran B (2001) Analytical models of the dynamic traffic assignment problem. *Netw Spat Econ* 1:377–390
- Branston D (1976) Link capacity functions: a review. *Transp Res* 10:223–236
- Campbell ME (1950) *Route selection and traffic assignment*. Highway Research Board, Washington
- Chen Z, He F, Zhang L, Yin Y (2016) Optimal deployment of autonomous vehicle lanes with endogenous market penetration. *Transp Res Part C Emerging Technol* 72:143–156
- Colini-Baldeschi R, Cominetti R, Mertikopoulos P, Scarsini M (2020) When is selfish routing bad? the price of anarchy in light and heavy traffic. *Oper Res* 68:411–434
- Correa JR, Schulz AS, Stier-Moses NE (2007) Fast, fair, and efficient flows in networks. *Oper Res* 55:215–225
- Correa J, Cristi A, Oosterwijk T (2019) On the price of anarchy for flows over time. pp. 559–577
- Cortés P, Muñozuri J, Guadix J, Onieva L (2013) Optimal algorithm for the demand routing problem in multicommodity flow distribution networks with diversification constraints and concave costs. *Int J Prod Econ* 146:313–324
- DalSasso V, Morandi V (2021) Gathering avoiding centralized pedestrian advice framework: an application for covid-19 outbreak restrictions. In: *Optimization and data science: trends and applications*, pp. 39–49. Springer
- De Palma A, Kilani M, Lindsey R (1932) Comparison of second-best and third-best tolling schemes on a road network. *Transp Res Rec* 89–96:2005
- de Palma A, Lindsey R (2011) Traffic congestion pricing methodologies and technologies. *Transp Res Part C Emerging Technol* 19
- Di X, Liu HX (2016) Boundedly rational route choice behavior: a review of models and methodologies. *Transp Res Part B Methodol* 85:142–179

- Di W, Yin Y, Lawphongpanich S, Yang H (2012) Design of more equitable congestion pricing and tradable credit schemes for multimodal transportation networks. *Transp Res Part B Methodol* 46:1273–1287
- Di X, Liu HX, Pang J-S, Ban XJ (2013) Boundedly rational user equilibria (Brue): mathematical formulation and solution sets. *Transp Res Part B Methodol* 57:300–313
- Di X, He X, Guo X, Liu HX (2014) Braess paradox under the boundedly rational user equilibria. *Transp Res Part B Methodol* 67:86–108
- Di X, Liu HX, Ban XJ (2016) Second best toll pricing within the framework of bounded rationality. *Transp Res Part B Methodol* 83:74–90
- Di X, Liu HX, Zhu S, Levinson DM (2017) Indifference bands for boundedly rational route switching. *Transportation* 44:1169–1194
- Djavadian S, Hoogendoorn RG, Van Arerm B, Chow JYJ (2014) Empirical evaluation of drivers' route choice behavioral responses to social navigation. *Transp Res Rec* 2423:52–60
- Falcochchio John C, Levinson Herbert S (2015) Road traffic congestion: a concise guide. Springer, Cham
- Feldmann R, Gairing M, Lücking T, Monien B, Rode M (2003) Selfish routing in non-cooperative networks: a survey. In: International symposium on mathematical foundations of computer science, pp. 21–45. Springer
- Frank M, Wolfe P (1956) An algorithm for quadratic programming. *Naval Res Logist Q* 3:95–110
- Grazia Speranza M (2018) Trends in transportation and logistics. *Eur J Oper Res* 264:830–836
- Guo X, Yang H (2009) User heterogeneity and bi-criteria system optimum. *Transp Res Part B Methodol* 43:379–390
- Harks T, Kleinert I, Klimm M, Möhring RH (2015) Computing network tolls with support constraints. *Networks*, 65
- Holmberg K, Yuan D (2003) A multicommodity network-flow problem with side constraints on paths solved by column generation. *INFORMS J Comput* 15:42–57
- Hongli X, Lou Y, Yin Y, Zhou J (2011) A prospect-based user equilibrium model with endogenous reference points and its application in congestion pricing. *Transp Res Part B Methodol* 45:311–328
- Hosseini-nasab S-M, Shetab-Boushehri S-N, Hejazi SR, Karimi H (2018) A multi-objective integrated model for selecting, scheduling, and budgeting road construction projects. *Eur J Oper Res* 271(1):262–277
- Jahn O, Möhring RH, Schulz AS, Stier-Moses N (2005) System-optimal routing of traffic flows with user constraints in networks with congestion. *Oper Res* 53:600–616
- Jahn O, Möhring RH, Schulz AS (2000) Optimal routing of traffic flows with length restrictions in networks with congestion. In: Operations research proceedings 1999, pp. 437–442. Springer
- Jayakrishnan R, Mahmassani HS, Hu TY (1987) An evaluation tool for advanced traffic information and management systems in urban networks. *Transp Res Part C Emerging Technol* 2
- Jiang N, Xie C (2014) Computing and analyzing mixed equilibrium network flows with gasoline and electric vehicles. *Comput Aid Civ Infrastruct Eng* 29:626–641
- Jiang N, Xie C, Waller ST (2012) Path-constrained traffic assignment: model and algorithm. *Transp Res Rec* 2283:25–33
- Jiang N, Xie C, Duthie JC, Travis Waller S (2014) A network equilibrium analysis on destination, route and parking choices with mixed gasoline and electric vehicular flows. *EURO J Transp Logist* 3:55–92
- Jou R-C, Lam S-H, Liu Y-H, Chen K-H (2005) Route switching behavior on freeways with the provision of different types of real-time traffic information. *Transp Res Part A Policy Pract* 39:445–461
- Jou R-C, Hensher DA, Liu Y-H, Chiu C-S (2010) Urban commuters' mode-switching behaviour in Taipei, with an application of the bounded rationality principle. *Urban Stud* 47:650–665
- Kachroo P, Özbay KMA (2011) Feedback ramp metering in intelligent transportation systems. Springer, Cham
- Klein I, Levy N, Ben-Elia E (2018) An agent-based model of the emergence of cooperation and a fair and stable system optimum using Atis on a simple road network. *Transp Res Part C Emerging Technol* 86:183–201
- Koch R, Skutella M (2011) Nash equilibria and the price of anarchy for flows over time. *Theory Comput Syst* 49:71–97
- Korilis YA, Lazar AA, Orda A (1997) Achieving network optima using Stackelberg routing strategies. *IEEE/ACM Trans Netw* 5:161–173
- Krichene W, Reilly JD, Amin S, Bayen AM (2014) Stackelberg routing on parallel networks with horizontal queues. *IEEE Trans Autom Control* 59:714–727
- Levy N, Klein I, Ben-Elia E (2018) Emergence of cooperation and a fair system optimum in road networks: a game-theoretic and agent-based modelling approach. *Res Transp Econ* 68:46–55

- Levy N, Klein I, Ben-Elia E (2017) Emergence of cooperation and a fair system optimum in road networks: a game-theoretic and agent-based modelling approach. *Res Transp Econ*
- Lin H, Roughgarden T, Tardos É, Walkover A (2011) Stronger bounds on Braess's paradox and the maximum latency of selfish routing. *SIAM J Discret Math* 25:1667–1686. <https://doi.org/10.1137/090769600>
- Lindsay R, Verhoef E (2001) Traffic congestion and congestion pricing. Emerald Group Publishing Limited, UK
- Liu H, Wang DZW (2015) Global optimization method for network design problem with stochastic user equilibrium. *Transp Res Part B Methodol* 72:20–39
- Li Z, Zhao X (2008) Integrated-equilibrium routing of traffic flows with congestion. In: *World Congress vol 17*, pp. 16065–16070
- Lo HK, Luo XW, Siu BWY (2006) Degradable transport network: travel time budget of travelers with heterogeneous risk aversion. *Transp Res Part B Methodol* 40:792–806
- Lou Y, Yin Y, Lawphongpanich S (2010) Robust congestion pricing under boundedly rational user equilibrium. *Transp Res Part B Methodol* 44:15–28
- Lujak M, Giordani S, Ossowski S (2015) Route guidance: bridging system and user optimization in traffic assignment. *Neurocomputing* 151:449–460
- Luo J, Hubaux J-P (2004) A survey of inter-vehicle communication. *Tech Rep*
- Mahmassani HS, Chang G-L (1987) On boundedly rational user equilibrium in transportation systems. *Transp Sci* 21:89–99
- Mahmassani HS, Jayakrishnan R (1991) System performance and user response under real-time information in a congested traffic corridor. *Transp Res Part A Gen* 25:293–307
- Mahmassani HS, Liu Y-H (1999) Dynamics of commuting decision behaviour under advanced traveller information systems. *Transp Res Part C Emerging Technol* 7:91–107
- Martin M, Kate L, Steskal L (2013) Braess's paradox for flows over time. *Theory Comput Syst* 53:86–106
- Meng Q, Yang H, Bell MGH (2001) An equivalent continuously differentiable model and a locally convergent algorithm for the continuous network design problem. *Transp Res Part B Methodol* 35:83–105
- Miralinaghi M, Peeta S (2016) Multi-period equilibrium modeling planning framework for tradable credit schemes. *Transp Res Part E Logist Transp Rev* 93:177–198
- Miralinaghi M, Peeta S (2019) Promoting zero-emissions vehicles using robust multi-period tradable credit scheme. *Transp Res Part D Transp Environ* 75:265–285
- Miralinaghi M, Peeta S (2020) Design of a multiperiod tradable credit scheme under vehicular emissions caps and traveler heterogeneity in future credit price perception. *J Infrastruct Syst* 26:04020030
- Miralinaghi M, Lou Y, Hsu Y-T, Shabanpour R, Shafahi Y (2016) Multiclass fuzzy user equilibrium with endogenous membership functions and risk-taking behaviors. *J Adv Transp* 50:1716–1734
- Mirchandani P, Soroush H (1987) Generalized traffic equilibrium with probabilistic travel times and perceptions. *Transp Sci* 21:133–152
- Möhring RH (1999) Optimal routing of traffic flows with length restrictions in networks with congestion. In: *Operations research proceedings 1999: selected papers of the symposium on operations research (SOR99)*, Magdeburg, September 1–3, pp. 437. Springer-Verlag
- O'Hare SJ, Connors RD, Watling DP (2016) Mechanisms that govern how the price of anarchy varies with travel demand. *Transp Res Part B Methodol* 84:55–80
- Papageorgiou M, Kotsialos A (2000) Freeway ramp metering: an overview. In: *Proceedings on intelligent transportation systems, 2000*, IEEE, pp. 228–239
- Patriksson M (2008) On the applicability and solution of bilevel optimization models in transportation science: a study on the existence, stability and computation of optimal solutions to stochastic mathematical programs with equilibrium constraints. *Transp Res Part B Methodol* 42:843–860
- Patriksson M (2015) *The traffic assignment problem: models and methods*. Courier Dover Publications, New York
- Peeta S, Ramos J, Pasupathy R (2000) Content of variable message signs and on-line driver behavior. *Transp Res Rec J Transp Res Board* 1750:102–108
- Peeta S, Ziliaskopoulos AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. *Netw Spatial Econ* 1
- Ramazani H, Shafahi Y, Seyedabrishami SE (2011) A fuzzy traffic assignment algorithm based on driver perceived travel time of network links. *Scientia Iranica* 18:190–197
- Roughgarden T (2008) Algorithmic game theory: some greatest hits and future directions. In: Giorgio A, Juhani K, Giancarlo M, and Luke O, (eds), *Fifth Ifip international conference on theoretical computer science—TCS 2008*, pp. 21–42. Springer US

- Roughgarden T (2003) The price of anarchy is independent of the network topology. *J Comput Syst Sci* 67:341–364
- Roughgarden T (2006) On the severity of Braess's paradox: designing networks for selfish users is hard. *J Comput Syst Sci* 72:922–953
- Roughgarden T, Tardos É (2002) How bad is selfish routing?. *J ACM (JACM)*, 49
- Ryus P, Vandehey M, Eleftheriadou L, Dowling RG, Ostrom BK (2011) Highway capacity manual 2010. TR News, pp. 45–48
- Saw K, Katti BK, Joshi G (2015) Literature review of traffic assignment: static and dynamic. *Int J Transp Eng* 2:339–347
- Schulz AS, Stier-Moses NE (2006) Efficiency and fairness of system-optimal routing with user constraints. *Networks* 48:223–234
- Seilabi SE, Tabesh MT, Davatgari A, Miralinaghi M, Labi S (2020) Promoting autonomous vehicles using travel demand and lane management strategies. *Front Built Environ* 6:156
- Sheffi Y (1985) Urban transportation networks: equilibrium analysis with mathematical programming methods. Prentice-Hall, London
- Shirmohammadi N, Yin Y (2016) Tradable credit scheme to control bottleneck queue length. *Transp Res Rec* 2561:53–63
- Shirmohammadi N, Zangui M, Yin Y, Nie Yu (2013) Analysis and design of tradable credit schemes under uncertainty. *Transp Res Rec* 2333:27–36
- Sichitiu M, Kihl M (2008) Inter-vehicle communication systems: a survey. *IEEE Commun Surv Tutor* 88–105
- Song Z, Yin Y, Lawphongpanich S (2015) Optimal deployment of managed lanes in general networks. *Int J Sustain Transp* 9:431–441
- Stopher PR (2004) Reducing road congestion: a reality check. *Transp Policy* 11:117–131
- Swamy C (2012) The effectiveness of Stackelberg strategies and tolls for network congestion games. *ACM Trans Algorith (TALG)* 8:1–19
- Szeto WY, Wang Yi, Han Ke (2015) Bounded rationality in dynamic traffic assignment. In: Bounded rational choice behaviour: applications in transport, pp. 163–187. Emerald Group Publishing Limited
- Ta-Yin H, Mahmassani HS (1997) Day-to-day evolution of network flows under real-time information and reactive signal control. *Transp Res Part C Emerging Technol* 5:51–69
- Van Essen M, Eikenbroek O, Thomas T, Van Berkum E (2019) Travelers' compliance with social routing advice: impacts on road network performance and equity. *IEEE Trans Intell Transp Syst*
- Wang X, Yang H, Zhu D, Li C (2012) Tradable travel credits for congestion management with heterogeneous users. *Transp Res Part E Logist Transp Rev* 48:426–437
- Wardrop JG (1952b) Proceedings of the institute of civil engineers
- Wardrop JG (1952) Road paper. Some theoretical aspects of road traffic research. *Proc Inst Civ Eng* 1:325–362
- Watling D (2006) User equilibrium traffic network assignment with stochastic travel times and late arrival penalty. *Eur J Oper Res* 175:1539–1556
- Yang H, Bell MGH (1998) Models and algorithms for road network design: a review and some new developments. *Transp Rev* 18:257–278
- Yang H, Wang X (2011) Managing network mobility with tradable credits. *Transp Res Part B Methodol* 45:580–594
- Yang H, Zhang X (2002) Multiclass network toll design problem with social and spatial equity constraints. *J Transp Eng* 128:420–428
- Yang H, Huang H-J (2005) Mathematical and economic theory of road pricing
- Ye H, Yang H (2017) Rational behavior adjustment process with boundedly rational user equilibrium. *Transp Sci* 51:968–980
- Youn H, Gastner MT, Jeong H (2008) Price of anarchy in transportation networks: efficiency and optimality control. *Phys Rev Lett* 101
- Yuan H, Wang R, Zhang X, Yueqi H, Zhang F, Zhu T, Liu H (2019) Evacuation strategy optimization study based on system theory. *IEEE Access* 7:111232–111244
- Zhang L (2011) Behavioral foundation of route choice and traffic assignment: comparison of principles of user equilibrium traffic assignment under different behavioral assumptions. *Transp Res Rec* 2254:1–10
- Zhenlong L, Xiaohua Z (2008) Integrated-equilibrium routing of traffic flows with congestion. In: IFAC proceedings vol 41, pp. 16065–16070

Zhou B, Li X (2012) User equilibrium with length constrained users. In: 2012 Fifth international joint conference on computational sciences and optimization, pp. 494–499. IEEE

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.