The Impact of Fairness on Performance in Congestion Networks

Bryce L. Ferguson and Jason R. Marden

***Abstract*— How do incentives that are designed to improve system performance affect fairness among users? Operators of traffic systems are typically concerned with a spectrum of performance metrics including aggregate congestion levels, fairness, and monetary expenditures. In this work, we focus on the design of taxation mechanisms to optimize both aggregate congestion levels as well as fairness, where fairness is defined as a measure of the difference between the quality of service of similar users in the system. Specifically, we derive the explicit trade-off between the performance guarantee and worst-case unfairness when designing a tolling scheme, showing that improved performance guarantees cause increased unfairness. Additionally, we find this relationship in the settings where the toll designer is informed and uninformed on the populations price sensitivities; by comparing the two, one can observe the value of this information when designing fair and good performing tolls.**

I. Introduction

At the advent of new communication technologies, modern societal infrastructures are becoming increasingly interconnected with the social systems they are designed to serve. With this expanded interconnection, the actions of individual users are becoming increasingly significant in affecting the operation of the system in whole. Though users may make decisions rationally, it is well known that the self-interested decisions of users can negatively impact the overall system performance. This phenomenon occurs in areas such as transportation networks [1], smart power grids [2], ride-sharing systems [3], [4], and resource allocation [5], [6]. A common way of measuring the inefficiency caused by users' selfish decisions is the ratio between the system welfare from users making selfish decisions and the optimal welfare, termed the *price of anarchy* [7].

Often, optimal system performance could be achieved by a central authority dictating each users action; however, it is often not feasible, or ethical, to take away a person's autonomy. Instead, an effective way to improve system performance is to influence users to make decisions that are more inline with the system level objective [8], [9]. One well studied method of influencing users is introducing *incentives* to alter their perceived costs [10]–[15]. Though this change in user behavior can improve the system welfare, it will also cause some users to make decisions that would have otherwise been less desirable. This causes a disparity in the cost incurred by different users, leading to apparent unfairness.

In this work, we seek to understand what is the effect of incentives on the fairness in a system? And, what trade-off exists between designing incentives that give good performance and preserving fairness? Additionally, we ask how does knowledge about the user population affect this trade-off?

We consider these questions in a class of network routing problems in which a mass of traffic must traverse a parrallel network with congestible edges. Solving for a network flow that minimizes the total congestion in the system is trivial; however, as users are prone to take paths that minimize their own observed transit delay, we use the notion of a *Nash flow* to describe the emergent behavior in the system. The primary function of incentives (specifically tolls) is to alter the users costs and reduce the aggregate latency in a Nash flow, and thus reduce the price of anarchy.

The use of tolls to reduce system congestion has been widely studied in network routing problems, including the nominal case where the population is homogeneous and users each respond identically to incentives [10], [11], [16], and the heterogeneous setting where each has an inherent *price sensitivity* that dictates how an incentive is perceived [12]–[14], [17]. In either case, the system designer has the ability to implement tolls to influence users as to improve a variety of performance measures, including aggregate congestion or system-wide fairness; however, there is currently no work in either setting that investigates the relationship between the two objectives.

Previous work has focused on finding flows that minimize congestion [16] or unfairness [18] respectively; to the best of the author's knowledge, no work has considered the multi-criterion problem of considering both fairness and system performance concurrently. In [19], the authors look at algorithms that meet explicit fairness conditions but do not discuss the effect this has on the congestion in the network. The author of [20] considers the unfairness that can occur in a flow that minimizes total congestion, but does not consider possible heterogeneity in users' price sensitivities and the design of tolls, nor do they consider the unfairness of tolls that give more modest performance guarantees.

Though the design of tolls that reduce system congestion have been highly studied, and the fairness of routing schemes has received some focus, no work has yet studied the explicit relationship between the two in incentive design. In this work, we seek to better understand the interrelation between the objectives of designing good performing tolls and designing tolls that preserve fairness. Specifically, we derive bounds on the disparity in users' travel delay when using a tolling scheme that gives optimal performance guarantees.

This research was supported by ONR Grant #N00014-20-1-2359 and AFOSR Grant #FA95550-20-1-0054

B. L. Ferguson (corresponding author) and J. R. Marden are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, {blferguson,jrmarden}@ece.ucsb.edu.

Additionally, we quantify a trade-off in the performance guarantee and the possible unfairness that can occur, showing that tolls which offer superior performance guarantees are subject to causing greater unfairness.

We derive these relationships in both the homogeneous and heterogeneous population setting. In the case where the users are heterogeneous in their price sensitivities, we consider an informed and uninformed system designer, where the system designer is either fully aware of each users price sensitivity or knows only the possible support of users price sensitivities. By comparing the performance-fairness trade-off curve of the informed and uniformed toll designer (shown in Fig. 1) we can see how possessing this information aids a system operator in balancing these two criterion.

II. Preliminaries

A. Network Congestion Games

In this work, we consider network routing problems in which a unit mass of traffic must traverse a parallel network with congestable edges. Let E be a set of directed edges that connect an origin node o to a destination node d, and, on each edge $e \in E$, let $f_e \geq 0$ denote the mass of traffic using that edge to traverse the network. A *feasible flow* $f = \{f_e\}_{e\in E}$ is an assignment of traffic on the network such that $f \in \Delta(E)$, where $\Delta(\cdot)$ denotes the standard probability simplex over the set E such that $\sum_{e\in E} f_e = 1$. To characterize transit delay, each edge $e \in E$ is endowed with a latency function $\ell_e(f_e) = a_e f_e + b_e$ where $a_e \geq 0$ and $b_e \geq 0$ are parameters describing how the delay grows with additional congestion and the constant delay on each edge respectively.

To capture the congestion in the network, in a flow f we characterize the total latency

$$\mathcal{L}(f) = \sum_{e\in E} f_e \cdot \ell_e(f_e),$$

as the aggregate delay seen by the traffic in the network. The objective of a system designer is to minimize the total latency, thus the optimal flow is $f^{\rm opt} \in \arg\min_{f\in\Delta(E)} \mathcal{L}(f)$. A routing problem is denoted $G = (E, \{\ell_e\}_{e\in E})$; the class of such parallel routing networks with finite number of edges is denoted $\mathcal{G}$.

In this work, we consider the setting of selfish routing, where each infinitesimal user in the mass of traffic selects their own path. Though a system designer may desire to minimize the aggregate delay, when users select routes as to minimize their own observed delay the emergent behavior may not align with the system optimal. To address this, a popular method to reduce this inefficiency is to add *incentives* to each edge to alter the users preferences [10], [11], [21], [22]. On an edge $e \in E$ let $\tau_e \in \mathbb{R}_{\geq 0}$ be a fixed toll that acts as a monetary fee to users who choose to travel on that edge. By selecting a *tolling scheme* $\{\tau_e\}_{e\in E}$ a toll designer can incentivize users to shape their emergent behavior. More formally, a user x utilizing an edge $e(x) \in E$ experiences cost

$$J_x(f) = \ell_{e(x)}(f_{e(x)}) + \tau_{e(x)}, \tag{1}$$

describing a tolled, non-atomic congestion game. To capture the emergent behavior in the system, we define a flow f as a Nash flow if for every user $x \in [0,1]$

$$J_x(f) \in \arg\min_{e\in E}\{\ell_{e(x)}(f_{e(x)}) + \tau_{e(x)}\}. \tag{2}$$

A tolled network congestion game is characterized by the network and tolling scheme, and denoted by the tuple $(G, \{\tau_e\}_{e\in E})$. It is shown in [23] that a Nash flow always exists in a game of this form, and the total latency of a Nash flow is unique. To determine the tolling scheme to be used in a specific routing problem, we investigate the use of a *taxation mechanism* $T : \mathcal{G} \to \mathbb{R}^{|E|}_{\geq 0}$ that assigns tolls to each edge.

B. Performance & Fairness

Many works have looked at the use of tolls to reduce congestion in network routing problems [11], [13], [14], [17]; however, what has been studied far less is the fairness (or lack thereof) of these incentive schemes [1], [20].

To quantify the performance of a taxation mechanism by the *price of anarchy*, let $\mathcal{L}^{\rm Nf}(G,T)$ be the total latency in the Nash flow of the congestion game with network G and tolling scheme $T(G)$, and let $\mathcal{L}^{\rm opt}(G)$ be the total latency in the optimal flow in the routing problem G. The price of anarchy quantifies the inefficiency from selfish routing under a given tolling scheme, and is defined as

$$\mathrm{PoA}(G,T) = \frac{\mathcal{L}^{\rm Nf}(G,T)}{\mathcal{L}^{\rm opt}(G)} \geq 1. \tag{3}$$

This definition can be generalized to include families of problem instances,

$$\mathrm{PoA}(\mathcal{G},T) = \sup_{G\in\mathcal{G}} \left\{ \frac{\mathcal{L}^{\rm Nf}(G,T)}{\mathcal{L}^{\rm opt}(G)} \right\}, \tag{4}$$

such that the price of anarchy serves as a worst-case inefficiency over the possible problem instances under a particular tolling scheme T; the price of anarchy can thus be used as a performance guarantee, providing an upper bound on the inefficiency that can occur under a taxation mechanism T. The optimally performing taxation mechanism is therefore the one that minimizes the price of anarchy ratio, i.e., $T^\star \in \arg\min_T(\mathrm{PoA}(\mathcal{G},T))$.

Though taxation mechanisms may exist that improve the performance guarantees by reducing the price of anarchy, an adverse effect is that taxes can cause disparities in the delays users see and lead to unfairness. To quantify how unfair a taxation mechanism can be, we define the *unfairness ratio* as

$$\mathrm{UFR}(G,T) = \frac{\max_{e\in E} \ell_e(f_e^{\rm Nf})}{\min_{e\in E} \ell_e(f_e^{\rm Nf})} \geq 1, \tag{5}$$

where $f^{\rm Nf}$ is the Nash flow in the congestion game G with taxation mechanism T. This metric measures the ratio between what the largest and smallest delay a user sees under a given taxation mechanism. This metric can also be extended to include families of instances,

$$\mathrm{UFR}(\mathcal{G},T) = \sup_{G\in\mathcal{G}} \left\{ \frac{\max_{e\in E} \ell_e(f_e^{\rm Nf})}{\min_{e\in E} \ell_e(f_e^{\rm Nf})} \right\}, \tag{6}$$

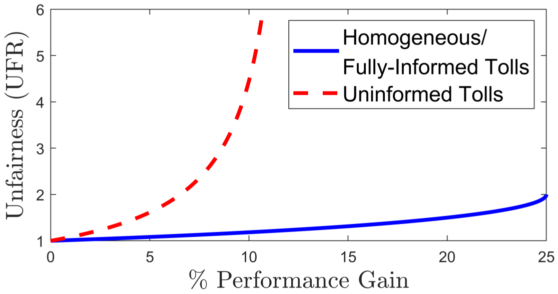


Fig. 1: Trade-off frontier between improvement in performance and unfairness for fully-informed/homogeneous (Theorem 1) and uniformed (Theorem 2) tolls. The horizontal axis measures the percent improvement in the price of anarchy bound caused by using incentives. Improving the performance is subject to increasing unfairness (measured by unfairness ratio) on the vertical axis. The disparity between the two lines quantifies how valuable it is for a system designer to be informed of users price sensitivities. For the uninformed plot, the values $S_{\mathrm{L}} = 1$ and $S_{\mathrm{U}} = 5$ are used.

where the unfairness ratio now represents the worst-case unfairness that can occur while using a taxation mechanism T. In this work, we seek to understand the relationship between performance guarantees and the resulting unfairness.

C. Summary of Contributions

In this work, we investigate the trade-off in performance and fairness when using incentive mechanisms and consider the role of information in designing incentives with good performance that also remain fair. The main question we seek to address is: *how do improved price of anarchy guarantees affect the unfairness that can occur?* We answer this question in two settings: in Section III-A, we derive the relationship between price of anarchy guarantees and unfairness ratio in network congestion games, in Section III-B we derive the same relationship but when the population of users may be heterogeneous in their price sensitivities. We then discuss how the system designer being aware or unaware of the users' price sensitivities affects their ability to design fair and good performing tolls.

III. Main Results

In this work, we investigate what guarantees can be made about fairness in network congestion games while using tolls.

Remark 1: The tolling scheme that optimizes the unfairness ratio defined in (6) is to use no toll at all, i.e., $\tau_e = 0 \ \forall e \in E$, resulting in an unfairness of 1.

This remark can easily be verified by observing that the untolled Nash flow occurs when each user takes a path of minimum cost and thus each experience the same delay and have an unfairness ratio of 1. This simple observation raises questions on how the objective of improving performance and maintaining fairness relate. In the following sections, we will derive bounds on the unfairness that can occur while using optimal tolling schemes as well as determine a trade-off between the price of anarchy and unfairness ratio while designing tolls. Some proofs of theorems and supporting lemmas are omitted for brevity; the proofs in their entirety are available online.

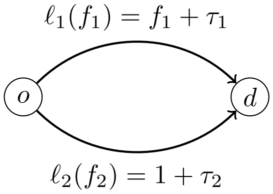


Fig. 2: Two link network congestion game. One edge possesses a linear latency function, the other a constant latency function.

A. Homogeneous Population

In this subsection, we consider the nominal network routing problem where each user is homogeneous in their response to incentives. In this setting, we investigate the relationship between achievable performance guarantees and possible unfairness. Specifically, we provide an upper-bound on unfairness when performance is optimized and characterize the trade-off between performance guarantees and worst-case unfairness, shown in Fig. 1. One endpoint of this trade-off curve is already known: when no tolls are used, the unfairness is minimized and the price of anarchy is no greater than $4/3$ [1], our results prove the other endpoint as well as a characterization of the Pareto-optimal frontier between them.

When designing a tolling scheme, the toll designer may have a target flow that the population of users will reach at equilibrium. We say a flow f is *enforceable* in a network $G \in \mathcal{G}$ if there exists a set of tolls $T(G) = \{\tau_e\}_{e \in E}$ such that f is the Nash flow in this congestion game.

Lemma 1: For any parallel-network $G \in \mathcal{G}$ and feasible flow $f \in \Delta(E)$, there exists a set of tolls $T(G)$ that enforces the flow f.

The proof of Lemma 1 appears in an online appendix.

Although the system designer possesses the ability to enforce any flow, improving performance often comes at the cost of increasing unfairness. Consider the following example where designing tolls to minimize the total latency in the system causes noticeable unfairness.

Example 1: Let G be a two link parallel network with edge latency functions $\ell_1(f_1) = f_1$ and $\ell_2(f_2) = 1$ as shown in Fig. 2. When no tolls are used, the Nash flow that emerges is of all users utilizing the first edge; this gives a price of anarchy of $\mathrm{PoA}(G, \emptyset) = 4/3$. By choosing the tolls $T(G) = \{\tau_1 = 1/2, \tau_2 = 0\}$, the system designer can enforce the flow $f = (1/2, 1/2)$ which minimizes the total latency, i.e., $\mathrm{PoA}(G, T(G)) = 1$. Though this flow optimizes performance, the users utilizing the second edge see a delay of $\ell_2(1/2) = 1$, while the users on the first edge see a delay of only $\ell_1(1/2) = 1/2$, giving an unfairness of

$$\mathrm{UFR}(G, T(G)) = 2.$$

In this example, tolls were introduced that reduced the system cost by 25% at the cost of causing some users to see twice the delay of others. This motivates the question: how unfair are optimal incentives?

In Proposition 1 we give an upper-bound on the unfairness that can be observed under optimal tolls.

Proposition 1: For an affine-latency, parallel network $G \in \mathcal{G}$ if $T^{\star}(G)$ enforces the system optimal flow, i.e.

$\mathrm{PoA}(G, T^\star(G)) = 1$, then the unfairness is upper-bounded by

$$\mathrm{UFR}(G, T^\star) \leq 2. \tag{7}$$

This gives the second end point on the trade-off curve between improving performance and worst-case unfairness shown in Fig. 1. The authors of [20] use a similar metric for unfairness and find a bound that evaluates to the same as (7) but for a broader class of networks. We offer a proof in an online appendix of the bound with the introduced unfairness metric to show it holds in the new setting.

Observing that an optimal toll can cause significant unfairness, we consider that a system designer may be willing to provide more modest performance guarantees if it results in less unfairness in the system. In Theorem 1 we characterize the trade-off between performance guarantees and worst-case unfairness for a toll designer.

Theorem 1: In the class of parallel-network, affine-latency congestion games $\mathcal{G}$ with a taxation mechanism T to guarantee an unfairness of

$$\mathrm{UFR}(G, T(G)) \leq \beta \quad \forall\, G \in \mathcal{G}, \tag{8}$$

for $\beta \in [1, 2]$ then the performance guarantee satisfies

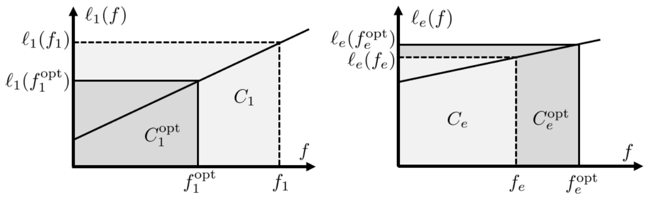
$$\mathrm{PoA}(\mathcal{G}, T) \geq \frac{1}{\beta - \beta^2/4}, \tag{9}$$

and this bound is essentially tight.

This result can be interpreted as a toll designer giving a guarantee on the level of unfairness in the system causing some limit on the achievable performance guarantee, or conversely, by improving the performance guarantee, more unfairness can occur. This trade-off can be seen in Fig. 1.

Proof: We will show that guaranteeing network flows with price of anarchy less than $1/(\beta - \beta^2/4)$ for $\beta \in [1, 2]$, the worst case unfairness that can occur can be made no less than β. We show this by first giving an instance where each of (9) and (8) are tight and neither price of anarchy or unfairness can be reduced without increasing the other; we then show this is the best bound.

Consider a two link network with latency functions $\ell_1(f_1) = f_1$ and $\ell_2(f_2) = \beta$ where $\beta \in [1, 2]$. Let $T(G)$ be an incentive mechanism that enforces the flow $f_1 = 1 - \epsilon$ and $f_2 = \epsilon$ the unfairness ratio is $\mathrm{UFR}(G, T(G)) = \beta$ and the price of anarchy is $\mathrm{PoA}(G, T(G)) = \frac{(1-\epsilon)^2 + \beta\epsilon}{\beta - \beta^2/4}$. By letting $\epsilon \to 0$ we get that (9) and (8) are tight in the limit.



(a) More flow than optimal

(b) Less flow than optimal

Fig. 3: Geometric interpretation of total latency on an edge. The area of a rectangle C_e is $\ell_e(f_e) \cdot f_e$ which is the total latency associated with that edge; the rectangles C_e^{opt} represents the same but for the optimal flow. The price of anarchy can be computed by comparing the total area of the rectangle $\{C_e\}_{e \in E}$ and $\{C_e^{\mathrm{opt}}\}_{e \in E}$

To show there is no better bound, we need to identify network flows that cannot be changed to improve the price of anarchy without making the unfairness ratio worse (i.e., the flow is Pareto-optimal). Let f be a flow in the network G. Because the unfairness ratio only considers two edges (the largest and smallest latency) the aforementioned network flows are those that have the edge with least latency e_1 with flow $f_1 > f_1^{\mathrm{opt}}$ and the remaining flow routed as to minimize the total latency on the remaining edges[1] (reducing f_1 will improve the price of anarchy but increase unfairness). The following geometric interpretation will help in proving the claim: in Fig. 3a, the rectangle C_1 with width f_1 and height $\ell_1(f_1)$ represents the system cost on that edge in the flow f; similarly, the rectangle C_1^{opt} represents the system cost on the edge in the optimal flow. Fig. 3b shows an arbitrary edge $e \neq e_1$ and accompanying rectangles C_e and C_e^{opt} associated with system cost.

We now look at what the largest unfairness can be in Pareto-optimal network flows with the same price of anarchy. The price of anarchy can be measured by the ratio of the total area of rectangle $\{C_e\}_{e \in E}$ and the total area of rectangle $\{C_e^{\mathrm{opt}}\}_{e \in E}$, i.e.,

$$\mathrm{PoA}(G, T(G)) = \frac{\sum_{e \in E} C_e}{\sum_{e \in E} C_e^{\mathrm{opt}}}. \tag{10}$$

The unfairness is measured by the ratio of the height of the tallest rectangle among $\{C_e\}_{e \in E \setminus e_1}$ and rectangle C_1. When ℓ_1 is linear and the remaining edge latency functions are constant, the height of rectangle C_1 can be smallest relative to any other rectangle. Further, the unfairness is maximized when the, now constant, highest utilized latency $\ell_n(f_n) = \beta$ is largest; this can be increased until the edge would no longer be utilized in the flow f. Noting that, if multiple edges have constant latency only one will be used in a Pareto-optimal flow, we can see that the Pareto-optimal frontier must be realized by a two link network with a linear latency function and constant that, in a flow f has arbitrarily-little utilization of the second, constant-latency edge. This precisely describes the instance used in the first part of this proof. ■

B. Heterogeneous Population

In the previous subsection, the fairness of tolling schemes was investigated in the nominal network congestion game setting. However, in many settings, each user may not respond identically to the same monetary incentive [12]–[14], [17]. To account for this, we will investigate the same questions of fairness in tolling with a *heterogeneous population*. Let $s_x > 0$ denote the *price sensitivity* of a user $x \in [0, 1]$ that represents how user x relates the cost of experiencing delay to monetary costs and is the reciprocal to value of time. More formally, a user x utilizing an edge $e(x) \in E$ experiences cost

$$J_x(f) = \ell_{e(x)}(f_{e(x)}) + s_x \tau_{e(x)}. \tag{11}$$

[1] each of the remaining edges will have flow $f_e < f_e^{\mathrm{opt}}$ for all $e \neq e_1$

A population of users is defined by a price-sensitivity distribution $s : [0,1] \to \mathbb{R}_{>0}$; the emergent behavior of a population is still captured by the Nash flow, where no user (of any sensitivity) has incentive to unilateraly deviate from their current action. A flow f is a Nash flow if for every user $x \in [0,1]$

$$J_x(f) \in \arg\min_{e\in E}\{\ell_{e(x)}(f_{e(x)}) + s_x \tau_{e(x)}\}. \tag{12}$$

A network congestion game with heterogeneous population is characterized by the network, user sensitivity distribution, and tolling scheme, and denoted by the tuple $(G, s, \{\tau_e\}_{e\in E})$.

In this work, we consider two settings for the toll designer: the first, where the system designer is fully aware of each users price sensitivity, and second, where the system designer knows only the possible support of users price sensitivities, referred to as the fully informed and uninformed settings respectively. Let $\mathcal{S} = \{s : [0,1] \to [S_{\rm L}, S_{\rm U}]\}$ be the set of price sensitivity distributions for populations with price sensitivities between a known lower and upper bound $S_{\rm L}$ and $S_{\rm U}$. In the fully informed setting, the system designer knows the exact price sensitivity distribution s, while in the uninformed setting, they know only that it comes from the set $\mathcal{S}$.

When the toll designer is fully informed, they may implement a taxation mechanism $T : \mathcal{G} \times \mathcal{S} \to \mathbb{R}^{|E|}_{\geq 0}$ that assigns a set of tolls for a given network-population pair, i.e., $T(G,s) = \{\tau_e\}_{e\in E}$. When the toll designer is unaware of the users price sensitivities, a taxation mechanism is denoted $T : \mathcal{G} \to \mathbb{R}^{|E|}_{\geq 0}$, where the tolls are now chosen agnostic of the sensitivity distribution, i.e., $T(G) = \{\tau_e\}_{e\in E}$ for any $s \in \mathcal{S}$. When it is clear from context, the taxation mechanism will simply be denoted T. In this section, we derive the explicit trade-off between fairness and performance for a toll designer with heterogeneous users and discuss the role of information in implementing tolls that offer improved performance and their respective effect on fairness. This trade-off can be seen in Fig. 1.

1) Fully Informed Tolling: In the homogeneous setting, there was no disparity in users' responses and the system designer was able to select tolls that enforced any feasible flow. When the users are heterogeneous, the system designer is faced with additional challenges in designing tolls; however, when the system designer is fully informed on each users price sensitivity, they have similar capabilities in designing incentives as in the homogeneous setting.

Remark 2: For any parallel-network $G \in \mathcal{G}$, user sensitivity distribution $s \in \mathcal{S}$, and feasible flow $f \in \Delta(E)$, there exists a fully informed tolling mechanism $T(G,s)$ that enforces the flow f. Thus, a fully informed toll designer is afforded the same fairness-performance capabilities as in the homogeneous setting.

This can be seen from the results of [13] and observing that every flow $f \in \Delta(E)$ is a minimal flow for any price sensitivity distribution $s \in \mathcal{S}$ when the network G is a parallel network; making any flow f enforceable.

This shows that having the knowledge of the users price sensitivities allows the toll designer to mitigate any negative implications of a heterogeneous population while designing good and fair tolls. In Fig. 1 the fully informed line thus represents the trade-off in performance and fairness to a toll designer that knows the exact price sensitivity distribution and a toll designer in the homogeneous setting. In the next section, we will investigate how the lack of this information affects this trade-off.

2) Uninformed Tolling: When a toll designer is uninformed on the exact price sensitivity of each user in the system, they are not afforded the luxury of designing tolls that enforce a desired flow; rather, an uninformed toll designer must design *robust tolls* that reduce worst-case inefficiency that can occur from a possible population of users. Though the information of user sensitivities is valuable to a toll designer, encouraging results exist on the design of incentives that offer improved performance guarantees even in the absence of this information [9], [12], [17]. However, the effect of these robust tolls on the unfairness among users has yet to be well understood. To motivate this, consider the following example:

Example 2: Let G be a two link parallel network with edge latency functions $\ell_1(f_1) = f_1$ and $\ell_2(f_2) = 1$ as shown in Fig. 2. Now consider that each user can have any price sensitivity between 1 and 10, i.e., $s_x \in [1, 10]$ for each user x. It is shown in [9] that an optimal incentive in this network will satisfy $\tau_1 - \tau_2 = 1/11$, accordingly, consider the tolling scheme $T^\star(G) = \{\tau_1 = 1/11,\ \tau_2 = 0\}$. Here, optimal refers to minimizing the worst-case price of anarchy guarantee over possible populations of users, or $T^\star \in \arg\min_T \max_{s\in\mathcal{S}} {\rm PoA}(G, s, T(G))$. When choosing an optimal toll in this problem, the price of anarchy guarantee is ${\rm PoA}(G, s, T^\star) \leq 1.223$ for all $s \in \mathcal{S}$.

Considering the entire population of users have sensitivity 10, the resulting Nash flow will be $f^{\rm Nf} = (1/11, 10/11)$ and users will experience an unfairness of

$$\text{UFR}(G, S_{\rm U}, T^\star) = \frac{\ell_1(1/11) = 1/11}{\ell_2(10/11) = 1} = 11.$$

In this example, the toll designer was able to reduce the untolled price of anarchy guarantee of $4/3$ by almost 10%, even in the absence of specific information about users' price sensitivities. However, this came at the cost of possibly causing some users to see 11 times the delay of others.

In the remainder of this section, we investigate the worst-case unfairness that can occur while using an optimal, robust toll as well as the trade-off between designing fair and good performing tolls when the system designer is uninformed on user sensitivities. In Proposition 2, we provide an upper bound on the unfairness that can occur in a parallel-network congestion game while using the optimal, robust tolls.

Proposition 2: For an affine-latency, parallel network $G \in \mathcal{G}$ with the optimal uninformed taxation mechanism $T^\star(G)$, i.e. $T^\star \in \arg\min_T\{{\rm PoA}(G, \mathcal{S}, T(G))\}$, then the unfairness is upper-bounded by

$$\mathrm{UFR}(G, \mathcal{S}, T^{\star}(G)) \leq \frac{1}{1 - \frac{S_{\mathrm{U}}}{S_{\mathrm{L}}+S_{\mathrm{U}}}}. \qquad (13)$$

The proof of Proposition 2 appears in an online appendix.

This result highlights the potentially large amount of unfairness that can occur among users when an uninformed system designer implements performance improving tolls. As was the case in Section III-A, a system designer may be willing to sacrifice some performance guarantees to reduce the level of unfairness in the system. In Theorem 2, we quantify this trade-off between performance guarantees and unfairness guarantees available to an uninformed system designer.

Theorem 2: In the class of parallel-network, affine-latency congestion games $\mathcal{G}$, using a taxation mechanism T that is unaware of users price sensitivities, to guarantee an unfairness of

$$\mathrm{UFR}(G, s, T(G)) \leq \frac{\beta}{\beta - \frac{S_{\mathrm{L}}}{S_{\mathrm{U}}}(\beta - 1)} \quad \forall\, G \in \mathcal{G},\ s \in \mathcal{S}, \qquad (14)$$

for $\beta \in [1, \frac{S_{\mathrm{U}}}{S_{\mathrm{U}}-S_{\mathrm{L}}}]$, the performance guarantee satisfies

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}, T) \geq \frac{1}{\beta - \beta^2/4}. \qquad (15)$$

As before, this result can be interpreted as a system designer implementing tolls that do not exceed some level of unfairness being constrained in the possible performance guarantee that can be achieved, or the level of possible unfairness that can occur by improving the performance guarantee. The less unfairness guaranteed to occur in the system (represented by (14)), the larger the bound on inefficiency that can occur (represented by (15)). This trade-off can be seen in Fig. 1, along side the trade-off for an informed toll designer. The juxtaposition of the two highlights the importance of knowing user sensitivities in designing tolls that give improvements in performance that also remain fair.

Proof: Consider a two link network G with a linear latency function $\ell_1(f_1) = f_1$ and $\ell_2(f_2) = \beta > 0$; above, these networks have been shown to experience worst-case price of anarchy and unfairness. Consider a set of tolls $T(G) = \{\tau_1, \tau_2\}$ that satisfies $\tau_1 - \tau_2 = \frac{\beta-1}{S_{\mathrm{L}}}$ and define $\Delta_\tau := \tau_1 - \tau_2$. Let f^{L} denote the Nash flow that occurs in the network G with tolls $T(G)$ when the each user has price sensitivity S_{L}; similarly, let f^{H} denote the Nash flow when each user has price sensitivity S_{U}. Following from Lemma 3, the maximum price of anarchy is realized by one of these two flows. With tolls $T(G)$,

$$f_1^{\mathrm{L}} = \beta - S_{\mathrm{L}}\Delta_\tau = 1,$$

from selecting $\Delta_\tau = \frac{\beta-1}{S_{\mathrm{L}}}$, and

$$f_1^{\mathrm{H}} = \beta - S_{\mathrm{U}}\Delta_\tau = \beta - \frac{S_{\mathrm{U}}}{S_{\mathrm{L}}}(\beta - 1).$$

The price of anarchy in each flow is

$$\mathrm{PoA}(G, S_{\mathrm{L}}, T(G)) = \frac{1}{\beta - \beta^2/4} \qquad (16)$$

$$\mathrm{PoA}(G, S_{\mathrm{U}}, T(G)) = \frac{(f_1^{\mathrm{H}})^2 + \beta(1 - f_1^{\mathrm{H}})}{\beta - \beta^2/4}, \qquad (17)$$

where f_1^{H} is defined as above. Observe that for $\beta \in [1, \frac{S_{\mathrm{U}}}{S_{\mathrm{U}}-S_{\mathrm{L}}}]$, the worst-case performance satisfies $\mathrm{PoA}(G, S_{\mathrm{L}}, T(G)) > \mathrm{PoA}(G, S_{\mathrm{U}}, T(G))$.

From Lemma 2, the largest unfairness occurs in the flow f^{H}. In the described instance, the unfairness ratio is

$$\mathrm{UFR}(G, S_{\mathrm{U}}, T(G)) = \frac{\beta}{\beta - \frac{S_{\mathrm{U}}}{S_{\mathrm{L}}}(\beta - 1)}. \qquad (18)$$

The price of anarchy in this instance, from (17), can only be reduced by increasing Δ_τ. However, the worst-case unfairness ratio in (18) is increasing with Δ_τ, thus giving a Pareto-optimal point on the trade-off frontier and proving the claim. ■

IV. Conclusion

In this paper, the relationship between designing tolls to reduce congestion and preserve fairness was studied. It was found that there is an explicit trade-off between the performance guarantee of a tolling mechanism and the unfairness that can occur. By deriving this trade-off for system designers that are fully informed and uninformed on the users price sensitivities respectively, the value of knowing the populations response to incentives when designing good performing and fair incentives was shown. Future work will investigate how partial information can aid a toll designer in trading-off performance and fairness guarantees, as well as how repeated interactions with similar populations can allow a system designer to accrue knowledge of the population and improve both performance and fairness guarantees.

References

[1] T. Roughgarden, "How Bad Is Selfish Routing?" *J. ACM*, vol. 49, no. 2, pp. 236–259, 2002.

[2] M. Alizadeh, Y. Xiao, A. Scaglione, and M. van der Schaar, "Dynamic Incentive Design for Participation in Direct Load Scheduling Programs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 6, pp. 1111–1126, 2014.

[3] A. Kleiner, B. Nebel, and V. A. Ziparo, "A mechanism for dynamic ride sharing based on parallel auctions," in *International Joint Conference on Artificial Intelligence*, 2011, pp. 266–272.

[4] C. Silva, B. F. Wollenberg, and C. Z. Zheng, "Application of mechanism design to electric power markets (republished)," *IEEE Transactions on Power Systems*, vol. 16, no. 4, pp. 862–869, nov 2001.

[5] D. Paccagnan, R. Chandan, and J. R. Marden, "Distributed resource allocation through utility design - Part {I:} optimizing the performance certificates via the price of anarchy," *CoRR*, vol. abs/1807.0, 2018. [Online]. Available: http://arxiv.org/abs/1807.01333

[6] L. Zhang, "The efficiency and fairness of a fixed budget resource allocation game," in *Lecture Notes in Computer Science*, vol. 3580. Springer Verlag, 2005, pp. 485–496.

[7] C. Papadimitriou, "Algorithms, games, and the internet," in *STOC '01: Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 2001, pp. 749–753.

[8] H. Tavafoghi and D. Teneketzis, "Informational incentives for congestion games," in *55th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2017*, vol. 2018-Janua. Institute of Electrical and Electronics Engineers Inc., jan 2018, pp. 1285–1292.

[9] B. L. Ferguson, P. N. Brown, and J. R. Marden, "Carrots or Sticks? The Effectiveness of Subsidies and Tolls in Congestion Games," in *2020 American Control Conference (ACC)*, jul 2020, pp. 1853–1858.

[10] V. Bilò and C. Vinci, "Dynamic taxes for polynomial congestion games," in *EC 2016 - Proceedings of the 2016 ACM Conference on Economics and Computation*. New York, New York, USA: ACM Press, 2016, pp. 839–856.

[11] D. Paccagnan, R. Chandan, B. L. Ferguson, and J. R. Marden, "Incentivizing efficient use of shared infrastructure: Optimal tolls in congestion games," nov 2019. [Online]. Available: http://arxiv.org/abs/1911.09806

[12] P. N. Brown and J. R. Marden, "Optimal Mechanisms for Robust Coordination in Congestion Games," in *54th IEEE Conf. Decision and Control*, Osaka, Japan, 2015, pp. 2283–2288.

[13] L. Fleischer, K. Jain, and M. Mahdian, "Tolls for Heterogeneous Selfish Users in Multicommodity Networks and Generalized Congestion Games," in *Proc. 45th IEEE Symp. on Foundations of Computer Science*, Rome, Italy, 2004, pp. 277–285.

[14] R. Cole, Y. Dodis, and T. Roughgarden, "Pricing network edges for heterogeneous selfish users," in *Proc. of the 35th ACM symp. on Theory of computing*, New York, New York, USA, 2003, pp. 521–530.

[15] P. N. Brown and J. R. Marden, "Can Taxes Improve Congestion on All Networks?" *IEEE Transactions on Control of Network Systems*, 2020.

[16] I. Milchtaich, "Social optimality and cooperation in nonatomic congestion games," *Journal of Economic Theory*, vol. 114, no. 1, pp. 56–87, jan 2004.

[17] B. L. Ferguson, P. N. Brown, and J. R. Marden, "Utilizing Information Optimally to Influence Distributed Network Routing," in *Proceedings of the IEEE Conference on Decision and Control*, vol. 2019-Decem. Institute of Electrical and Electronics Engineers Inc., dec 2019, pp. 5008–5013.

[18] D. Chakrabarty, A. Mehta, V. Nagarajan, and V. Vazirani, "Fairness and optimality in congestion games," in *Proceedings of the ACM Conference on Electronic Commerce*. New York, New York, USA: ACM Press, 2005, pp. 52–57.

[19] J. Kleinberg, Y. Rabani, and E. Tardos, "Fairness in routing and load balancing," in *Annual Symposium on Foundations of Computer Science - Proceedings*. IEEE, 1999, pp. 568–578.

[20] T. Roughgarden, "How Unfair is Optimal Routing?" in *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '02. USA: Society for Industrial and Applied Mathematics, 2002, pp. 203–204.

[21] R. Meir and D. C. Parkes, "When are Marginal Congestion Tolls Optimal?" *ATT@IJCAI*, 2016.

[22] S. A. Bagloee and M. Sarvi, "A modern congestion pricing policy for urban traffic: subsidy plus toll," *Journal of Modern Transportation*, vol. 25, no. 3, pp. 133–149, sep 2017.

[23] A. Mas-Colell, "On a Theorem of Schmeidler," *Mathematical Economics*, vol. 13, pp. 201–206, 1984.

Appendix

In this appendix are several proofs supporting the main results presented in the body of the paper.

Proof of Lemma 1:

A potential function for the homogeneous network congestion game with fixed tolls is

$$\phi(f) = \sum_{e \in E} \frac{a_e}{2} f_e^2 + b_e f_e + \tau_e f_e. \tag{19}$$

This can be verified by observing that the gradient of (19) is equal to the user cost functions,

$$\nabla \phi(f) = \begin{bmatrix} a_1 f_1 + b_1 + \tau_1 \\ \vdots \\ a_n f_n + b_n + \tau_n \end{bmatrix} =: Af + b + \tau, \tag{20}$$

where $A = \mathrm{diag}\{a_1, \ldots, a_n\}$, $b = [b_1, \ldots, b_n]^T$, and $\tau = [\tau_1, \ldots, \tau_n]^T$. For any $f \in \Delta(E)$, picking τ such that $\nabla \phi(f) = 0$ will cause f to be a Nash equilibrium, i.e., $\tau = -Af - b$ enforces f. To maintain that $\tau_e \geq 0$ for each $e \in E$, observe that adding the same toll to each edge does not alter the Nash flow, thus $\tau = -Af - b + ||Af + b||_\infty \mathbb{1}_n$ is a valid tolling scheme that enforces f. ■

Proof of Proposition 1:

In a congestion game with network $G \in \mathcal{G}$ and user sensitivity distribution $s \in \mathcal{S}$, suppose $T^\star(G, s)$ enforces f^{opt} that minimizes the total latency and gives price of anarchy one. In [1] the authors show that in an optimal flow, the marginal cost of each utilized edge is equal, i.e.,

$$2a_1 f_1^{\mathrm{opt}} + b_1 = 2a_2 f_2^{\mathrm{opt}} + b_2 = \ldots = 2a_n f_n^{\mathrm{opt}} + b_n. \tag{21}$$

By ordering the edges by increasing constant terms

$$b_1 \leq b_2 \leq \ldots \leq b_n,$$

we get that

$$a_1 f_1^{\mathrm{opt}} \geq a_2 f_2^{\mathrm{opt}} \geq \ldots \geq a_n f_n^{\mathrm{opt}}.$$

This implies that the latency on each edge in this flow satisfies

$$a_1 f_1^{\mathrm{opt}} + b_1 = a_2 f_2^{\mathrm{opt}} + b_2 = \ldots = a_n f_n^{\mathrm{opt}} + b_n.$$

The unfairness ratio is thus

$$\mathrm{UFR}((G, s, T^\star(G, s)) = \frac{a_n f_n^{\mathrm{opt}} + b_n}{a_1 f_1^{\mathrm{opt}} + b_1}$$

In searching for an upper bound, without loss of generality we can make $b_1 = 0$.[2] Now to find an upper bound

$$\begin{aligned} \mathrm{UFR}((G, s, T^\star(G, s)) &= \frac{a_n f_n^{\mathrm{opt}} + b_n}{a_1 f_1^{\mathrm{opt}}} && (22) \\ &= \frac{2a_n f_n^{\mathrm{opt}} + 2b_n}{2a_1 f_1^{\mathrm{opt}}} && (23) \\ &= \frac{2a_n f_n^{\mathrm{opt}} + b_n + b_n}{2a_n f_n^{\mathrm{opt}} + b_n} && (24) \\ &= 1 + \frac{b_n}{2a_n f_n^{\mathrm{opt}} + b_n} \leq 2. && (25) \end{aligned}$$

where (24) holds from (21). ■

Lemma 2: In a network $G \in \mathcal{G}$ with tolls $T(G) = \{\tau_e\}_{e \in E}$, the largest unfairness over sensitivity distributions $s \in \mathcal{S}$ occurs from a population of users each having price sensitivity S_{U}.

Proof: Let f be the Nash flow in a network G with tolls $T(G) = \{\tau_e\}_{e \in E}$ for a population with sensitivity distribution $s \in \mathcal{S}$. Let $\underline{e}$ denote the edge with lowest latency in f and $\overline{e}$ denote the edge with largest latency that is utilized by a non-zero mass of traffic. Pick $S' \in [S_{\mathrm{L}}, S_{\mathrm{U}}]$ to be the sensitivity of some user that utilizes the edge $\overline{e}$. By the definition of a Nash flow,

$$a_{\underline{e}} f_{\underline{e}} + b_{\underline{e}} + S' \tau_{\underline{e}} \geq a_{\overline{e}} f_{\overline{e}} + b_{\overline{e}} + S' \tau_{\overline{e}},$$

[2] Consider subtracting b_1 from each latency function and see the Nash flow, optimal flow, and price of anarchy do not change.

and, by simply rearranging terms,

$$S'(\tau_{\underline{e}} - \tau_{\overline{e}}) \geq a_{\overline{e}} f_{\overline{e}} + b_{\overline{e}} - (a_{\underline{e}} f_{\underline{e}} + b_{\underline{e}}). \tag{26}$$

Consequentially, $\tau_{\underline{e}} \geq \tau_{\overline{e}}$ as the right-hand side of the above inequality is non-negative. Further, the right-hand side of (26) represents the unfairness in the network and is upper bounded by the left-hand side; this upper-bound is largest when $S' = S_{\mathrm{U}}$ and (26) holds with equality when every user has sensitivity S_{U}. ■

Lemma 3: In a two link network $G \in \mathcal{G}_2$, the optimal, uninformed tolls $T^{\star}(G)$ equates the total latency that occurs from Nash flows caused by the population where each user has sensitivity S_{L} and the population where each user has sensitivity S_{U}.

Proof: In a two link network G, with tolls τ_1 and τ_2, and sensitivity distribution s, let S^{ind} denote the sensitivity that equates $a_1 f_1 + b_1 + S^{\mathrm{ind}}\tau_1 = a_2 f_2 + b_2 + S^{\mathrm{ind}}\tau_2$ where f is the Nash flow caused by population s [3]. The Nash flow, from the population with sensitivity distribution s will be the same as the Nash flow from a homogeneous population where each user has sensitivity S^{ind}, as the same flow f will satisfy the equilibrium condition $a_1 f_1 + b_1 + S^{\mathrm{ind}}\tau_1 = a_2 f_2 + b_2 + S^{\mathrm{ind}}\tau_2$. Therefore, a search for worst-case sensitivity distributions is equivalent to a search over homogeneous distributions. The total latency $\mathcal{L}(f) = \sum_{e \in E} a_e f_e^2 + b_e = a_1 f_1^2 + b_1 f_1 + a_2(1 - f_1)^2 + b_2(1 - f_1)$ is quadratic in f_1, and $f_1 = \frac{a_2 + b_2 - b_1 + S^{\mathrm{ind}}(\tau_2 - \tau_1)}{a_1 + a_2}$ is an affine mapping of S^{ind}; thus, the total latency is convex in S^{ind}. The maximum of a convex function occurs at the boundary, which here are the homogeneous distributions where every user has S_{L} or S_{U} respectively. By defining $\Delta_\tau = \tau_2 - \tau_1$, it can be seen that total latency under each of the aforementioned distributions is convex in Δ_τ; further, the total latency of the all S_{L} distribution is decreasing with Δ_τ while total latency of all S_{U} is increasing with Δ_τ. The tolls that minimize the price of anarchy thus equates the total latency of these two distributions. ■

Proof of Proposition 2:

As shown in the proof of Theorem 1, the largest unfairness occurs in two link networks where the first edge has a linear latency function and the second edge a constant latency function. Without loss of generality, we will define these latency functions as $\ell_1(f_1) = f_1$ and $\ell_2(f_2) = \beta$ where $\beta \in \mathbb{R}_{>0}$. The optimal flow in this network is $f_1 = \beta/2$ if $\beta \in [0, 2]$, while the untolled Nash flow is $f_1 = \beta$ if $\beta \in [0, 1]$. Accordingly, any toll that does not increase the price of anarchy will satisfy $\tau_1 \geq \tau_2$. Let f_1^{L} be the flow from a population where each user has sensitivity S_{L}; similarly, let f_1^{H} be the flow from a population where each user has sensitivity S_{U}. From Lemma 3, these are the two maximal flows for total latency, and, from Lemma 2, f_1^{H} is the most unfair flow.

We prove the claim by contradiction. Assume that unfairness in a worst-case instance satisfies $\beta/f_1^{\mathrm{H}} > \frac{1}{1 - \frac{S_{\mathrm{U}}}{S_{\mathrm{L}} + S_{\mathrm{U}}}}$ By rearranging

$$\beta \frac{S_{\mathrm{L}}}{S_{\mathrm{L}} + S_{\mathrm{U}}} > f_1^{\mathrm{H}}, \tag{27}$$

$$\beta \frac{-S_{\mathrm{U}}}{S_{\mathrm{L}} + S_{\mathrm{U}}} > f_1^{\mathrm{H}} - \beta = S_{\mathrm{U}}(\tau_2 - \tau_1), \tag{28}$$

where the equality in (28) holds from the definition of f_1^{H} being a Nash flow satisfying $f_1^{\mathrm{H}} + S_{\mathrm{U}}\tau_1 = \beta + S_{\mathrm{U}}\tau_2$. From Lemma 2, an optimal toll in a two link network equates the total latency of the all S_{L} population and the all S_{U} population; this toll can found algebraically in this network to be $\tau_1 = \frac{\beta}{S_{\mathrm{L}} + S_{\mathrm{U}}}$, $\tau_2 = 0$. Substituting the optimal toll into (28) gives

$$\beta < (S_{\mathrm{L}} + S_{\mathrm{U}}) \frac{\beta}{S_{\mathrm{L}} + S_{\mathrm{U}}} = \beta. \tag{29}$$

This completes the contradiction. ■

[3] if $S^{\mathrm{ind}} > S_{\mathrm{U}}$ then instead choose S_{U}, similarly if $S^{\mathrm{ind}} < S_{\mathrm{L}}$, choose S_{L}.