The Robustness of Marginal-Cost Taxes in Affine Congestion Games

Philip N. Brown, *Student Member, IEEE,* and Jason R. Marden, *Member, IEEE*

Abstract—**The network routing literature contains many results showing that tolls can be used to improve the efficiency of network traffic routing. These results typically require toll-designers to have an exact characterization of the network and user population. We relax this strict informational dependence and present a simple setting in which scaled marginal-cost tolls can be guaranteed to provide significant efficiency improvements over the un-tolled case, even if the toll-sensitivities of the users are unknown.**

I. Introduction

It is widely known that uninfluenced social systems can exhibit suboptimal system-level performance. Characterizing this inefficiency, which is broadly referred to as the *price of anarchy* [2], is a highly active research area in many disciplines including network resource allocation [3], distributed control [4], traffic congestion [5], and others. This inefficiency has prompted new research questions geared at influencing social behavior to improve system performance [1], [6]–[8].

In this paper we focus on the design of an incentive mechanism for influencing social behavior for a simple class of congestion games. Specifically, we consider a routing problem where a unit mass of traffic needs to be routed across a parallel network consisting of edges with affine latency functions. Finding the flow that minimizes the total latency in the network is straightforward if a system-planner has direct control over all routing decisions. However, in social systems such as road traffic routing, individual users make local routing decisions in response to their personal objectives. Accordingly, we model the routing problem as a non-atomic congestion game where the unit of traffic can be viewed as a continuum of users, each controlling an infinitesimally-small amount of traffic and seeking to minimize its own experienced latency. Here, we adopt the popular viewpoint that a *Nash flow* characterizes the emergent collective behavior in such systems.

The pioneering work of Arthur Pigou in [9] demonstrated that the total latency associated with Nash flows could be substantially worse than the optimal total latency [10]. In fact, for affine-cost networks, a Nash flow can have a total latency up to 33% higher than the optimal total latency [11]; that is, the price of anarchy is $4/3$. With these inefficiencies in mind, researchers have focused on the use of monetary taxes to influence the underlying Nash flows. Typically, the efficacy of a taxation methodology is gauged by analyzing the Nash flow for a new routing game where the self-interested users seek to minimize a linear combination of latency and monetary tax. This has been a rich field of study, and many researchers have provided positive results, particularly in cases when the underlying system is characterized perfectly [12]–[14]. For example, given a complete characterization of network topology, latency functions, user demands, and user sensitivities, a system-designer can levy taxes which induce exactly-optimal Nash flows. Another example of network-dependent tolls can be found in [15], where the authors investigate the impact of tolls in affine-cost, parallel networks, and present a taxation mechanism that improves efficiency compared with un-tolled levels, even when the total traffic rate is unknown. There has been little research on the robustness of network-dependent tolls to unexpected changes in network topology or latency functions; a notable exception to this can be found in [16].

In contrast to the above network-dependent results, another important avenue of research is what we term the "network-agnostic" approach. In this approach, the system-planner assigns tolls to each network edge that depend only on the congestion properties of *that particular edge*; tolls cannot depend on the overall network topology. The most common example of this is known as a *marginal-cost* toll, a particular style of flow-varying toll which is known to induce optimal Nash flows without requiring the designer to have knowledge of the specific network topology [17], [18]. In [19], the authors study efficiency guarantees resulting from "restricted" marginal-cost tolls, i.e., marginal-cost tolls which saturate at a given upper bound. Unfortunately, marginal-cost tolls have largely only been studied in cases in which all network users share a common toll-sensitivity (or value-of-time). These tolls' robustness to variations or mischaracterizatons of user sensitivity is heretofore unknown.

In this paper, we recognize that the applicability of a given taxation mechanism hinges not only on its performance guarantees, but also on its robustness to variations or mis-characterizations of the underlying system. Our main contribution is to identify the optimal scaled marginal-cost taxation mechanism in terms of its robustness to mis-characterizations of user sensitivities, and we derive tight efficiency guarantees for this optimal scaled marginal-cost taxation mechanism that hold for any number of network links or distribution of user tax-sensitivities.

II. Model and Related Work

Consider a routing problem in which a unit mass of traffic needs to be routed across a parallel network consisting of a source node, a destination node, and a set of edges E connecting the source to the destination. A *feasible flow* over the network is characterized by a collection of edge flows

P. Brown is with the Department of Electrical, Computer, and Energy Engineering, The University of Colorado at Boulder, CO, 80302 USA e-mail: philip.brown@colorado.edu.

J. R. Marden is with the Department of Electrical, Computer, and Energy Engineering, The University of Colorado at Boulder, CO, 80302 USA e-mail: jason.marden@colorado.edu.

This research was supported by AFOSR grant #FA9550-12-1-0359, ONR grant #N00014-12-1-0643, and NSF grant #ECCS-1351866.

The conference version of this paper appeared in [1].

$f = \{f_e\}_{e\in E} \in \Delta(E)$ where $f_e \geq 0$ denotes the flow on edge e and $\Delta(E)$ denotes the simplex over the set E; i.e., $\sum_{e\in E} f_e = 1$. To characterize transit delay, each edge $e \in E$ is associated with a specific affine latency function of the form

$$\ell_e(f_e) = a_e f_e + b_e, \tag{1}$$

where $a_e \geq 0$ and $b_e \geq 0$ are edge-specific constants. We measure the the efficiency of a flow f by the *total latency*, given by

$$\mathcal{L}(f) = \sum_{e\in E} f_e \cdot \ell_e(f_e), \tag{2}$$

and we denote the flow that minimizes the total latency by $f^* \in \arg\min_{f\in\Delta(E)} \mathcal{L}(f)$. We specify a particular parallel network by the tuple $G = (E, \{\ell_e\}_{e\in E})$, and write the set of all parallel networks as $\mathcal{G}$.

In this paper we study taxation mechanisms for influencing the emergent collective behavior resulting from self-interested price-sensitive users. To that end, we model the above routing problem as a non-atomic congestion game where each edge $e \in E$ is assigned a flow-dependent taxation function $\tau_e : \mathbb{R}^+ \to \mathbb{R}^+$ and each user $x \in [0,1]$ has a taxation sensitivity $s_x \in [S_\mathrm{L}, S_\mathrm{U}] \subseteq \mathbb{R}^+$ where $S_\mathrm{U} \geq S_\mathrm{L} > 0$ denote upper and lower sensitivity bounds, respectively. Given a flow f, the cost that user x experiences for using edge $\tilde{e} \in E$ is of the form

$$J_x(f) = \ell_{\tilde{e}}(f_{\tilde{e}}) + s_x \tau_{\tilde{e}}(f_{\tilde{e}}). \tag{3}$$

We call a flow f a *Nash flow* if for all users $x \in [0,1]$ we have

$$J_x(f) = \min_{e\in E} \{\ell_e(f_e) + s_x \tau_e(f_e)\}. \tag{4}$$

It is well-known that a Nash flow exists for any non-atomic congestion game of the above form [20].

We study network-agnostic *taxation mechanisms*, in which a system-designer essentially commits to a taxation function for each potential network edge, and any network realization merely employs a subset of these pre-defined taxation functions. Simply put, an edge's taxation function is independent of any *other* edge's congestion properties or location in the network. A commonly-studied network-agnostic taxation mechanism is the marginal-cost (or Pigovian) taxation mechanism, which is of the following form: for any edge e with latency function (1), the associated marginal-cost taxation function is

$$\tau_e^{\mathrm{mc}}(f_e) = f_e \cdot \frac{d}{df_e}\ell_e(f_e) = a_e f_e, \ \forall f_e \geq 0. \tag{5}$$

In [17] the author shows that for any $G \in \mathcal{G}$, irrespective of the underlying network structure, Nash flows resulting from marginal-cost taxes are optimal, provided that all users share a common known sensitivity.

III. Our contributions

In this paper, we study the efficacy of a network-agnostic taxation mechanism for situations in which both the number of links and the users' price-sensitivities are unknown or time-varying. We study tolls of the following form: for any scalar coefficient $\kappa \geq 0$, the scaled marginal-cost taxation mechanism, denoted by $\boldsymbol{\tau}^{\mathrm{smc}}(\kappa)$, assigns taxation functions

$$\tau_e^{\mathrm{smc}}(f_e;\kappa) = \kappa \cdot f_e \cdot \frac{d}{df_e}\ell_e(f_e) = \kappa a_e f_e, \ \forall f_e \geq 0.$$

To formalize a notion of worst-case efficiency guarantees, we define the set of possible sensitivity distributions for the users as $\mathcal{S} = \{s : [0,1] \to [S_\mathrm{L}, S_\mathrm{U}]\}$. Let $\mathcal{L}^*(G)$ denote the total latency associated with the optimal flow, and $\mathcal{L}^{\mathrm{nf}}(G, s, \tau)$ denote the total latency associated with the Nash flow resulting from taxation functions τ and sensitivity distribution $s \in \mathcal{S}$.

We define the price of anarchy of the scaled marginal-cost taxation mechanism with respect to both uncertainity in the underlying network and and the users' price-sensitivity, i.e.,

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}, \boldsymbol{\tau}^{\mathrm{smc}}(\kappa)) = \sup_{s\in\mathcal{S}, G\in\mathcal{G}} \left\{ \frac{\mathcal{L}^{\mathrm{nf}}(G, s, \boldsymbol{\tau}^{\mathrm{smc}}(\kappa))}{\mathcal{L}^*(G)} \right\} \geq 1. \tag{6}$$

Our main contribution is identifying how the choice of κ impacts the above price of anarchy, and we identify the optimal κ and the resulting efficiency guarantees.

Theorem 1. *For any network $G \in \mathcal{G}$ with flow on all edges in an un-tolled Nash flow*[1]*, and any $s \in \mathcal{S}$, any scaled marginal-cost taxation mechanism reduces the total latency of any Nash flow when compared to the total latency of any Nash flow associated with the un-tolled case, i.e., for any $\kappa > 0$*

$$\mathcal{L}^{\mathrm{nf}}(G, s, \boldsymbol{\tau}^{\mathrm{smc}}(\kappa)) < \mathcal{L}^{\mathrm{nf}}(G, s, \emptyset).^2 \tag{7}$$

Furthermore, the unique optimal scaled marginal-cost tolling mechanism uses the scale factor

$$\kappa^* = \frac{1}{\sqrt{S_\mathrm{L} S_\mathrm{U}}} = \underset{\kappa\geq 0}{\arg\min} \ \{\mathrm{PoA}(\mathcal{G}, \mathcal{S}, \boldsymbol{\tau}^{\mathrm{smc}}(\kappa))\}. \tag{8}$$

Finally, the price of anarchy resulting from the optimal scaled marginal-cost taxation mechanism is

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}, \boldsymbol{\tau}^{\mathrm{smc}}(\kappa^*)) = \frac{4}{3}\left(1 - \frac{\sqrt{S_L/S_U}}{\left(1+\sqrt{S_L/S_U}\right)^2}\right). \tag{9}$$

Note that the optimal scale factor κ^* is independent of the number of network links and the agent sensitivity distribution[3], so tolls can be computed locally at each edge without requiring global network information. This low information-dependence places our work in contrast to many existing results, e.g. [12], that can guarantee higher efficiencies only at the expense of strict informational requirements. See Figure 1 for plots of the price of anarchy with respect to various parameters.

[1] This is essentially a regularity condition which prevents the creation of badly-designed networks with artificially-high efficiency losses: For example, consider a network which includes an edge e that has a constant latency function, i.e., $\ell_e(f_e) = b_e$, where b_e is sufficiently large so that $f_e^{\mathrm{ne}} = 0$ in the resulting un-tolled Nash flow. For such scenarios, levying tolls on the alternative edges could cause highly-sensitive users to deviate to edge e, thereby causing large network inefficiencies. Note that if such an un-used (and accordingly inefficient) edge does exist, we may levy a very large toll on it (effectively removing it from the network) and obtain our desired well-behaved situation.

[2] If the un-tolled Nash flow for a particular network is optimal, any Nash flow resulting from marginal-cost tolls is also optimal. Thus, all results in the paper assume that $\mathcal{L}^{\mathrm{nf}}(G, s, \emptyset) > \mathcal{L}^*(G)$.

[3] This price of anarchy bound is actually also robust to increases in the total mass of traffic flowing through the network; see Claim 1.1.1.

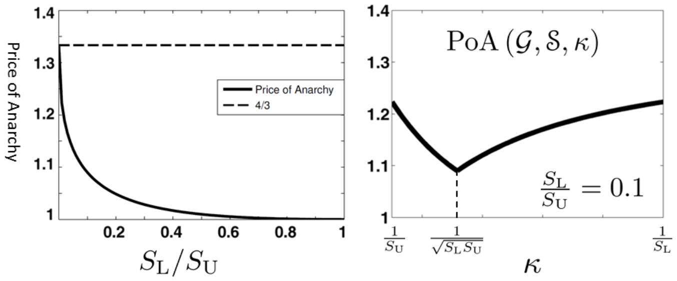

Fig. 1. Left: An illustration of the price of anarchy bound from Theorem 1, with optimal toll scalar $\kappa = (S_{\rm L}S_{\rm U})^{-1/2}$. Since the bound depends only on $S_{\rm L}/S_{\rm U}$, this plot neatly expresses the effect of model uncertainty on toll effectiveness. As expected, we inherit the canonical price of anarchy of $4/3$ when $S_{\rm L}/S_{\rm U} = 0$ (i.e., we may be unable to influence behavior at all). At the other extreme, when $S_{\rm L}/S_{\rm U} = 1$ (i.e., we know sensitivities perfectly) we inherit the canonical price of anarchy of 1. Our result continuously bridges the gap between the two extremes. Right: The price of anarchy (with a fixed ratio of $S_{\rm L}/S_{\rm U} = 0.1$) with respect to toll scalar κ. Note that the price of anarchy is minimized at the inverse of the geometric mean of $S_{\rm L}$ and $S_{\rm U}$.

Theorem 1 Proof: We begin with some notation before delving into the proof of Theorem 1. Throughout, it will often be convenient to focus on special classes of sensitivity distributions. To that end, let $\mathcal{S}_m \subseteq \mathcal{S}$ denote the set of user sensitivity functions that have a range consisting of at most m sensitivity values, i.e., $\left|\cup_{x\in[0,1]} s_x\right| \le m$.

Let $\mathcal{F}(G, \mathcal{S}, \boldsymbol{\tau}) \subset \mathbb{R}^n$ denote the set of Nash flows associated with all routing games $(G, s, \boldsymbol{\tau})$ where $s \in \mathcal{S}$. Note that we are representing Nash flows anonymously: a particular $f^{\rm nf} \in \mathcal{F}(G, \mathcal{S}, \boldsymbol{\tau})$ describes merely *how many* agents are on each edge, not *which* agents are on each edge. For brevity, we often express $\boldsymbol{\tau}^{\rm smc}(\kappa)$ as merely κ.

The proof of Theorem 1 involves proving that the scaling coefficient $\kappa \ge 0$ that mininizes the price of anarchy for heterogeneous populations can be determined by analyzing the scaling coefficient that minimizes the price of anarchy for homogeneous populations, a much smaller class of games. This reduction then facilitates a straightforward computation of the optimal coefficient. The complete proofs of all supporting lemmas can be found in the Appendix.

First, Lemma 1.1 proves that a Nash flow on an n-link network for any heterogeneous population can be represented as a Nash flow for a population with only $(n-1)$ sensitivities. Thus, populations with infinite sensitivities inherit the efficiency guarantees of those with $(n-1)$ sensitivities.

Lemma 1.1. *For any network $G \in \mathcal{G}$ consisting of n links, with $n \ge 2$, and $\kappa \ge 0$,*

$$\mathcal{F}(G, \mathcal{S}, \kappa) = \mathcal{F}(G, \mathcal{S}_{n-1}, \kappa). \tag{10}$$

Second, Lemma 1.2 shows that we may further refine our search to the set of homogeneous sensitivity distributions. In particular, when $\kappa \le \frac{1}{\sqrt{S_{\rm L}S_{\rm U}}}$, the worst-case total latency is realized by Nash flows for a homogeneous population with sensitivity $S_{\rm L}$.

Lemma 1.2. *Let $\kappa \le \frac{1}{\sqrt{S_{\rm L}S_{\rm U}}}$. Then for any $G \in \mathcal{G}$,*

$$\max_{s\in\mathcal{S}} \mathcal{L}^{\rm nf}(G, s, \kappa) = \mathcal{L}^{\rm nf}(G, S_{\rm L}, \kappa). \tag{11}$$

Finally, Lemma 1.3 gives the unique optimal value of κ for homogeneous populations; heterogeneous populations ultimately inherit this optimal result.

Lemma 1.3. *For all $G \in \mathcal{G}$, and for all $\kappa \ne \frac{1}{\sqrt{S_{\rm L}S_{\rm U}}} = \kappa^*$,*

$$\max_{s\in\mathcal{S}_1} \mathcal{L}^{\rm nf}\left(G, s, \kappa^*\right) < \max_{s\in\mathcal{S}_1} \mathcal{L}^{\rm nf}\left(G, s, \kappa\right). \tag{12}$$

Finally, the price of anarchy of $\boldsymbol{\tau}^{\rm smc}(\kappa^*)$ for homogeneous populations is given by (9).

Proof of Theorem 1. We combine the inequalities on the price of anarchy proved in each lemma. Lemma 1.1 implies that

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}, \kappa^*) = \mathrm{PoA}(\mathcal{G}, \mathcal{S}_{n-1}, \kappa^*). \tag{13}$$

Lemma 1.2 implies that

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}_{n-1}, \kappa^*) = \mathrm{PoA}(\mathcal{G}, \mathcal{S}_1, \kappa^*) \tag{14}$$

and the worst-case total latency with $\kappa = \kappa^*$ is better than the un-tolled total latency. By Lemma 1.3, we have that for any $\kappa \ne \kappa^*$,

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}_1, \kappa^*) < \mathrm{PoA}(\mathcal{G}, \mathcal{S}_1, \kappa). \tag{15}$$

Since $\mathcal{S}_1 \subseteq \mathcal{S}$, it is clear that for any κ,

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}_1, \kappa) \le \mathrm{PoA}(\mathcal{G}, \mathcal{S}, \kappa). \tag{16}$$

Combining inequalities (13), (14), (15), and (16), we have that for any $\kappa \ne \kappa^*$,

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}, \kappa^*) < \mathrm{PoA}(\mathcal{G}, \mathcal{S}, \kappa).$$

Thus, (9) is valid for heterogeneous populations as well. □

IV. Conclusions

In this paper, we proved tight bounds on the efficiency losses in affine-cost parallel-network congestion games due to the scaled marginal-cost taxation mechanism.

It is worth noting that the optimal scaled marginal-cost taxation mechanism, i.e., $\boldsymbol{\tau}^{\rm smc}(\kappa^*)$, is not necessarily the optimal taxation mechanism over the entire space of network-agnostic taxation mechanisms; nonetheless, for any network and user sensitivities, the taxation mechanism $\boldsymbol{\tau}^{\rm smc}(\kappa^*)$ always provides improvements in the efficiency of the resulting Nash flows when compared to the untolled case.

Clearly, there are many open questions in the area of robustness to unknown price-sensitivity; future results will extend the analysis to settings with asymmetric action sets and more general cost functions. As we study larger and larger classes of games, we plan to characterize the tradeoffs between the quality of the system-designer's information and the resulting efficiency guarantees that are possible.

It is an ongoing research effort to determine how the results contained in this paper can be generalized; we are actively studying what extensions are possible both in terms of general network topologies and nonlinear latency functions. It is often the case that nonlinear latency functions can exacerbate inefficiencies; it is as yet unknown what role this will play in questions of robustness.

References

[1] P. N. Brown and J. R. Marden, "Social Coordination in Unknown Price-Sensitive Populations," in *52nd IEEE Conference on Decision and Control*, pp. 1168 – 1173, 2013.

[2] C. Papadimitriou, "Algorithms, Games, and the Internet," in *Proc. of the 28th International Colloquium on Automata, Languages and Programming*, 2001.

[3] R. Johari and J. N. Tsitsiklis, "Efficiency Loss in a Network Resource Allocation Game," *Mathematics of Operations Research*, vol. 29, pp. 407–435, Aug. 2004.

[4] J. R. Marden and J. S. Shamma, "Game Theory and Distributed Control," in *Handbook of Game Theory Vol. 4* (H. Young and S. Zamir, eds.), Elsevier Science, 2014.

[5] G. Piliouras, E. Nikolova, and J. S. Shamma, "Risk sensitivity of price of anarchy under uncertainty," in *Proceedings of the fourteenth ACM conference on Electronic commerce - EC '13*, vol. 9, (New York, New York, USA), pp. 715–732, ACM Press, 2013.

[6] J. R. Marden and A. Wierman, "Distributed Welfare Games," *Operations Research*, vol. 61, pp. 155–168, 2013.

[7] R. Gopalakrishnan, J. R. Marden, and A. Wierman, "Potential games are necessary to ensure pure nash equilibria in cost sharing games," *Proceedings of the fourteenth ACM conference on Electronic commerce - EC '13*, pp. 563–564, 2013.

[8] W. Sandholm, "Negative Externalities and Evolutionary Implementation," *The Review of Economic Studies*, vol. 72, pp. 885–915, July 2005.

[9] A. C. Pigou, *The Economics of Welfare*. 1920.

[10] D. Braess, A. Nagurney, and T. Wakolbinger, "On a Paradox of Traffic Planning," *Transportation Science*, vol. 39, pp. 446–450, Nov. 2005.

[11] T. Roughgarden, "How Bad Is Selfish Routing?," *Journal of the ACM JACM (2002)*, vol. 49, no. 2, pp. 236–259, 2002.

[12] 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), pp. 521–530, ACM Press, 2003.

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

[14] G. Karakostas and S. Kolliopoulos, "Edge pricing of multicommodity networks for heterogeneous selfish users," *45th Annual IEEE Symposium on Foundations of Computer Science*, pp. 268–276, 2004.

[15] G. Christodoulou, K. Mehlhorn, and E. Pyrga, "Improving the price of Anarchy for Selfish Routing via coordination mechanisms," *Algorithmica*, vol. 69, no. 3, pp. 619–640, 2014.

[16] U. Bhaskar, K. Ligett, L. Schulman, and C. Swamy, "Achieving Target Equilibria in Network Routing Games without Knowing the Latency Functions," in *IEEE FOCS*, pp. 31–40, 2014.

[17] M. Beckman, C. McGuire, and C. B. Winsten, "Studies in the Economics of Transportation," 1956.

[18] W. Sandholm, "Evolutionary Implementation and Congestion Pricing," *The Review of Economic Studies*, vol. 69, no. 3, pp. 667–689, 2002.

[19] V. Bonifaci, M. Salek, and G. Schäfer, "Efficiency of restricted tolls in non-atomic network routing games," in *Symposium on Algorithmic Game Theory*, vol. 6982 LNCS, pp. 302–313, 2011.

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

Appendix
Proofs of Lemmas 1.1 and 1.2

A. Notation and Terminology

We assume that a network has $n \geq 2$ edges. Throughout the proof, we represent latency function parameters in matrix form: $A \in \mathbb{R}^{n \times n}$ is defined as the diagonal matrix with diagonal elements $(a_1, a_2, \ldots, a_n)$, and column vector $b \in \mathbb{R}^n$ contains all the constant coefficients from the edge latency functions. Without loss of generality, we assume that A has at least $(n-1)$ non-zero entries and that the edges are indexed such that b is arranged in ascending order, i.e., $b_i \leq b_j$ for all $i < j$. Using this notation, we write a flow $f \in \mathbb{R}^n$ as a column vector, so the vector of edge latencies $\ell(f) \in \mathbb{R}^n$ is $\ell(f) = Af + b$, and the total latency $\mathcal{L}(f)$ is given by

$$\mathcal{L}(f) = f^T A f + f^T b. \tag{17}$$

We write $\mathbf{0}$ and $\mathbf{1}$ to denote all-zeros and all-ones column vectors, respectively, and I to denote the identity matrix.

We express the edge set as $E = \{e_1, e_2, \ldots, e_n\}$, and write the latency function of edge e_i as $\ell_i(f_i) = a_i f_i + b_i$.

We often make use of a special Nash flow for a discrete distribution: we call a Nash flow in which every user is indifferent between at least two edges a *minimally-indifferent* Nash flow. We write the set of minimally-indifferent Nash flows for $\mathcal{S}_m$ for a given taxation mechanism $\boldsymbol{\tau}$ as $\mathcal{F}^{\rm mi}(\mathcal{G}, \mathcal{S}_m, \boldsymbol{\tau})$. Note that on a network with n links, there are at most $(n-1)$ sensitivity types in a minimally-indifferent Nash flow.

B. Proof of Lemma 1.1 and Associated Claims

This proof involves first proving two intermediate claims. In Claim 1.1.1 we show that if every link has positive flow in an un-tolled Nash flow, then under $\boldsymbol{\tau}^{\rm smc}(\kappa)$, every link in that network will have positive flow in a Nash flow induced by any finite $\kappa > 0$.

Claims 1.1.1 and 1.1.2 use the following definition: for Nash flow $f^{\rm nf} \in \mathcal{F}(G, \mathcal{S}, \kappa)$, for each edge $e_i \in E$, define s_i^- and s_i^+ by the following:

$$s_i^- = \inf_{x \in [0,1]} \left\{ s_x : \text{agent } x \text{ uses edge } e_i \text{ in flow } f^{\rm nf} \right\}, \tag{18}$$

$$s_i^+ = \sup_{x \in [0,1]} \left\{ s_x : \text{agent } x \text{ uses edge } e_i \text{ in flow } f^{\rm nf} \right\}. \tag{19}$$

For a particular Nash flow, s_i^- and s_i^+ represent the lowest and highest sensitivities of any agent on edge e_i, respectively.

Claim 1.1.1. *For any network $G \in \mathcal{G}$, let $f^{\rm nf} \in \mathcal{F}(G, \mathcal{S}, \kappa)$ for any $\kappa \geq 0$. Then $f^{\rm nf}$ has positive flow on every edge.*

Proof. To avoid trivialities, we assume that a positive mass of users have non-zero sensitivity. In an un-tolled Nash flow f, $\forall e_i, e_j \in E$, it must be that $a_i f_i + b_i = a_j f_j + b_j$. Suppose there is a tolled Nash flow $f^{\rm t} \in \mathcal{F}(G, \mathcal{S}, \kappa)$ for $\kappa > 0$ in which some edge e_k has $f_k^{\rm t} = 0$. Thus, for every edge e_i,

$$(1 + s_i^+ \kappa) a_i f_i^{\rm t} + b_i \leq b_k \leq a_i f_i + b_i. \tag{20}$$

Simplifying (20) and summing over edges, we obtain $\sum_{i=1}^n f_i^{\rm t} \leq \sum_{i=1}^n (f_i)/(1 + s_i^+ \kappa)$. Since at least one s_i^+ is strictly positive, this implies that $\sum_{i=1}^n f_i^{\rm t} < \sum_{i=1}^n f_i$, but this would mean that the tolled flow has less total traffic than the original un-tolled flow, a contradiction. □

Next, in Claim 1.1.2 we show that under scaled marginal-cost tolls, heterogeneous users sort themselves onto the links in a predictable order.

Claim 1.1.2. *Scaled marginal-cost tolls induce an ordering on the edges of a network: for any sensitivity distribution $s \in \mathcal{S}$ and toll scale factor $\kappa > 0$, given any two edges $e_i \in E$ and $e_j \in E$ for which $b_i \leq b_j$, the following conditions hold in a Nash flow $f^{\rm nf} \in \mathcal{F}(G, s, \kappa)$: (i) $a_i f_i^{\rm nf} \geq a_j f_j^{\rm nf}$, and (ii) $s_i^+ \leq s_{i+1}^-$.*

Proof. Consider edges e_i and e_{i+1} in network G. By hypothesis, $b_i \leq b_{i+1}$. Consider a Nash flow $f^{\mathrm{nf}} \in \mathcal{F}(G, s, \kappa)$ with $\kappa \geq 0$ and $s \in \mathcal{S}$. By Claim 1.1.1, $f^{\mathrm{nf}}_{i+1} > 0$. Take any user $x \in [0, 1]$ on edge e_{i+1}. Since this is a Nash flow, user x must (weakly) prefer edge e_{i+1} to edge e_i. Since each edge tolling function is $\tau_e(f_e) = a_e f_e$,

$$(1 + \kappa s_x)(a_i f_i^{\mathrm{nf}} - a_{i+1} f_{i+1}^{\mathrm{nf}}) \geq b_{i+1} - b_i \geq 0.$$

Thus, $a_i f_i^{\mathrm{nf}} \geq a_{i+1} f_{i+1}^{\mathrm{nf}} \geq 0$, for all i, establishing the first conclusion. A user with sensitivity s_{i+1}^{-} would also (weakly) prefer edge e_{i+1} to edge e_i:

$$(1 + \kappa s_{i+1}^{-}) a_{i+1} f_{i+1}^{\mathrm{nf}} + b_{i+1} \leq (1 + \kappa s_{i+1}^{-}) a_i f_i^{\mathrm{nf}} + b_i. \quad (21)$$

Since $a_{i+1} f_{i+1}^{\mathrm{nf}} \leq a_i f_i^{\mathrm{nf}}$, then for any $s > s_{i+1}^{-}$,

$$(1 + \kappa s) a_{i+1} f_{i+1}^{\mathrm{nf}} + b_{i+1} \leq (1 + \kappa s) a_i f_i^{\mathrm{nf}} + b_i.$$

Here, we find that any agent with higher sensitivity $s > s_{i+1}^{-}$ (weakly) prefers edge e_{i+1} to edge e_i, which implies that $s \geq s_i^{+}$; in other words, no agent using edge e_{i+1} has a lower sensitivity than any agent using edge e_i, or $s_i^{+} \leq s_{i+1}^{-}$, establishing the second conclusion.[4] □

To complete the proof, we exploit this ordering to construct a minimally-indifferent Nash flow from a Nash flow for any arbitrary sensitivity distribution, thus showing that worst-case behavior for arbitrary populations can always be realized by populations with a finite number of user sensitivities.

Consider edge e_i in Nash flow $f^{\mathrm{nf}} \in \mathcal{F}(G, s, \kappa)$; by Claim 1.1.2, $s_i^{+} \leq s_{i+1}^{-}$. We may rearrange (21) (and the opposite inequality for s_i^{+}) to obtain

$$\frac{b_{i+1} - b_i}{1 + \kappa s_{i+1}^{-}} \leq a_i f_i^{\mathrm{nf}} - a_{i+1} f_{i+1}^{\mathrm{nf}} \leq \frac{b_{i+1} - b_i}{1 + \kappa s_i^{+}}.$$

Now, for each $i \leq (n-1)$, let s_i be the solution to

$$a_i f_i^{\mathrm{nf}} - a_{i+1} f_{i+1}^{\mathrm{nf}} = \frac{b_{i+1} - b_i}{1 + \kappa s_i}. \quad (22)$$

Note that every $s_i \in [s_i^{+}, s_{i+1}^{-}]$ and that $s_i \leq s_{i+1}$. Now, construct a population of agents[5] in which $\forall i \in \{2, \ldots, n-2\}$, $(f_i^{\mathrm{nf}} + f_{i+1}^{\mathrm{nf}})/2$ agents have a sensitivity of s_i; $\left(f_1^{\mathrm{nf}} + f_2^{\mathrm{nf}}/2\right)$ agents have sensitivity s_1, and $\left(f_{n-1}^{\mathrm{nf}}/2 + f_n^{\mathrm{nf}}\right)$ agents have sensitivity s_{n-1}. Then $f^{\mathrm{nf}} \in \mathcal{F}^{\mathrm{mi}}(G, \mathcal{S}_{n-1}, \kappa)$; i.e., it is a minimally-indifferent Nash flow for the newly-constructed population containing $(n-1)$ sensitivity types. That is, for each s_i, the following is true:

$$(1 + \kappa s_i) a_i f_i + b_i = (1 + \kappa s_i) a_{i+1} f_{i+1} + b_{i+1}.$$

Since for any $f^{\mathrm{nf}} \in \mathcal{F}(G, \mathcal{S}, \kappa)$ we have shown that $f^{\mathrm{nf}} \in \mathcal{F}^{\mathrm{mi}}(G, \mathcal{S}_{n-1}, \kappa)$, it must be true that $\mathcal{F}(G, \mathcal{S}, \kappa) \subseteq \mathcal{F}^{\mathrm{mi}}(G, \mathcal{S}_{n-1}, \kappa)$. The opposite inclusion is obvious, since $\mathcal{S}_{n-1} \subseteq \mathcal{S}$, and the desired result is immediate. □

For the remainder of the proof of Theorem 1, we will assume without loss of generality that all Nash flows are minimally-indifferent.

C. *Proof of Lemma 1.2*

This proof hinges on a change of variables which allows us to linearly parameterize the set of all Nash flows on a network by a set of $(n-1)$ sensitivity values.

For any $G \in \mathcal{G}$, recall that any minimally-indifferent Nash flow $f^{\mathrm{nf}} \in \mathcal{F}^{\mathrm{mi}}(G, \mathcal{S}_{n-1}, \kappa)$ satisfies equation (22) for each pair of adjacent edges. Note that the expression is linear in f^{nf}, but nonlinear in $\{s_i\}$. However, if we define a new variable $z_i = \frac{1}{1+\kappa s_i}$, and let $z = (z_1, \ldots, z_{n-1})^T$, we can write (22) as a linear expression in both f^{nf} and z.

Recall also that $\sum_{i=1}^{n} f_i^{\mathrm{nf}} = 1$. Combining this with the $(n-1)$ equations obtained from (22), we can write the resulting n-dimensional linear system in matrix form as

$$P f^{\mathrm{nf}} = r + Qz \quad (23)$$

where $P \in \mathbb{R}^{n \times n}$ and $Q \in \mathbb{R}^{n \times n-1}$ are constant matrices depending only on G, and $r \in \mathbb{R}^{n \times 1}$ is the unit vector with 1 as the n-th element.

It can easily be verified that P must be full-rank, so we can write a Nash flow as a function of z by inverting P and defining

$$f^{\mathrm{nf}}(z) = R + Mz, \quad (24)$$

where $R \in \mathbb{R}^{n}$ and $M \in \mathbb{R}^{n \times n-1}$ are defined as

$$R = P^{-1} r, \qquad M = P^{-1} Q. \quad (25)$$

The following observations will be helpful to our proof:

Observation 1.2.1. *The matrices M and R possess the following properties for any $G \in \mathcal{G}$:*

$$\mathbf{1}^{\mathrm{T}} M = \mathbf{0}^{\mathrm{T}}, \quad (26)$$

$$\mathbf{1}^{\mathrm{T}} R = 1, \quad (27)$$

$$AR \in \mathrm{sp}\left\{\mathbf{1}\right\}, \quad (28)$$

$$M^T A M \mathbf{1} = -M^T b. \quad (29)$$

Observation 1.2.2. *The total latency $\mathcal{L}\left(f^{\mathrm{nf}}(z)\right)$ is given by the following convex quadratic form in z, which we simply write as a function of z:*

$$\mathcal{L}^{\mathrm{nf}}(z) = z^T M^T A M z + z^T M^T b + L_R, \quad (30)$$

where $L_R = R^T A R + b^T R$ is the total latency associated with the flow that results from $\kappa \to \infty$. Furthermore, L_R is also equal to the zero-toll Nash flow total latency:

$$\mathcal{L}^{\mathrm{nf}}(G, s, 0) = L_R. \quad (31)$$

Proof of Observation 1.2.1. These facts follow algebraically from the fact that by definition, for any $z \in \mathbb{R}^{n-1}$, $f^{\mathrm{nf}}(z)$ satisfies (23). □

Proof of Observation 1.2.2. We simply substitute $f^{\mathrm{nf}}(z)$ (that is, equation (24)) into (17) to obtain

$$\mathcal{L}(f^{\mathrm{nf}}(z)) = R^T A R + b^T R + z^T M^T A M z + b^T M z + 2 R^T A M z.$$

Consider the last term, $2R^T AMz$. By (28) in Observation 1.2.1, $\exists \alpha \in \mathbb{R}$ such that $R^T A = \alpha \mathbf{1}^{\mathrm{T}}$, and by (26), $\mathbf{1}^{\mathrm{T}} M = \mathbf{0}^T$, so $2R^T AMz = 0$. Simplifying, we obtain

$$\mathcal{L}^{\mathrm{nf}}(z) = z^T M^T A M z + z^T M^T b + L_R,$$

[4] Note that if $b_i = b_{i+1}$, all agents are indifferent between edges e_i and e_{i+1} in any Nash flow, so from the standpoint of edge-ordering, these two edges would behave as a single edge.

[5] This construction is not unique; there are infinitely-many ways to assign mass to the various sensitivity types.

where we let $\mathcal{L}^{\mathrm{nf}}(z) = \mathcal{L}(f^{\mathrm{nf}}(z))$ for brevity. Since A is positive semidefinite, $\mathcal{L}^{\mathrm{nf}}(z)$ is convex in z. Finally, note that that for $\kappa = 0$, $z = \mathbf{1}$. Thus, $f^{\mathrm{nf}}(\mathbf{1})$ represents the zero-toll Nash flow on G for any user sensitivity distribution. By (29) in Observation 1.2.1, we know that $M^T AM\mathbf{1} = -M^T b$, so the zero-toll total latency is given by $\mathcal{L}^{\mathrm{nf}}(\mathbf{1}) = L_R$. □

By focusing on minimally-indifferent Nash flows, we may use (24) to parameterize the set of all Nash flows for any network.

1) Characterizing the set of Nash flows: To formalize our definition of $f^{\mathrm{nf}}(z)$ (given in (24)), for any $S_{\mathrm{L}} \leq S_{\mathrm{U}}$ and $\kappa \geq 0$, we define the convex, bounded polytope $Z \subset \mathbb{R}^{n-1}$ as the set of solutions $\{z \in \mathbb{R}^{n-1}\}$ to the following inequalities:

$$\frac{1}{1+\kappa S_{\mathrm{L}}} \geq z_1 \geq \cdots \geq z_i \geq z_{i+1} \geq \cdots \geq z_{n-1} \geq \frac{1}{1+\kappa S_{\mathrm{U}}}. \tag{32}$$

By construction, this polytope Z is the domain of $f^{\mathrm{nf}}(z)$. In fact, Z is diffeomorphic to $\mathcal{F}(G, \mathcal{S}, \kappa)$: It is clear from (23) that any Nash flow can be written as $f^{\mathrm{nf}}(z) = R + Mz$ for some choice of z. Furthermore, for a given $\kappa > 0$, any $z \in Z$ uniquely defines a set of sensitivities $\{s_i\}_{i=1}^{n-1}$ according to the expression $z_i = \frac{1}{1+s_i\kappa}$, and the resulting sensitivities are ordered so they uniquely define a minimally-indifferent Nash flow on G. Thus, $f^{\mathrm{nf}}(z)$ is a continuous bijection between Z and $\mathcal{F}(G, \mathcal{S}, \kappa)$.

To complete the proof of Lemma 1.2, we argue by the convexity of Z and the properties of $\mathcal{L}^{\mathrm{nf}}(z)$ that when $\kappa \leq \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}}$ (i.e., tolls are low) the worst Nash flow is one in which all agents share the same low sensitivity.

Since Z is a bounded convex polytope, by convexity $\mathcal{L}^{\mathrm{nf}}(z)$ must take its maximum at a vertex of Z; it is straightforward to show that a vertex of Z corresponds to a Nash flow in which every agent lies at one of the extreme ends of the sensitivity range. This means that for any routing game, there are exactly two homogeneous vertices: one each for S_{L} and S_{U}, and $(n-2)$ heterogeneous vertices at which some agents have sensitivity S_{L} and the rest have S_{U}.

2) Homogeneous vertices represent worst-case Nash flows: Let z_v represent such a heterogeneous vertex; path-ordering dictates that it must be of this form: $z_v = [z_{\mathrm{L}}, \ldots, z_{\mathrm{L}}, z_{\mathrm{U}}, \ldots, z_{\mathrm{U}}]^T$. Thus, if we write the i-th column of M as μ_i, and let $\mu_{\mathrm{L}} = \sum_{i=1}^{\ell-1} \mu_i$ and $\mu_{\mathrm{U}} = \sum_{i=\ell}^{n-1} \mu_i$ (where ℓ is the lowest-index link being used by agents with sensitivity S_{U}), Mz_v can be expressed as $Mz_v = z_{\mathrm{L}}\mu_{\mathrm{L}} + z_{\mathrm{U}}\mu_{\mathrm{U}}$. Through a series of tedious algebraic steps relying on Observation 1.2.1, it can be shown that

$$\mu_{\mathrm{U}}^T (A\mu_{\mathrm{U}} + b) \leq 0. \tag{33}$$

Now, we wish to compute the difference $\mathcal{L}^{\mathrm{nf}}(z_{\mathrm{L}} \cdot \mathbf{1}) - \mathcal{L}^{\mathrm{nf}}(z_v)$. It can be shown that this difference is given by the expression

$$(z_{\mathrm{L}} - z_{\mathrm{U}})\, \mu_{\mathrm{U}}^T \left[(z_{\mathrm{L}} + z_{\mathrm{U}} - 1)\, A\mu_{\mathrm{U}} + (1 - 2z_{\mathrm{L}})\,(A\mu_{\mathrm{U}} + b)\right]. \tag{34}$$

When $\kappa \leq \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}}$, it is true that $z_{\mathrm{L}} \geq z_{\mathrm{U}}$, that $z_{\mathrm{L}} + z_{\mathrm{U}} - 1 \geq 0$, and that $1 - 2z_{\mathrm{L}} \leq 0$. A is positive semidefinite, so $\mu_{\mathrm{U}}^T A\mu_{\mathrm{U}} \geq 0$, and (33) shows that the expression in (34) must always be non-negative: $\mathcal{L}^{\mathrm{nf}}(z_{\mathrm{L}} \cdot \mathbf{1}) - \mathcal{L}^{\mathrm{nf}}(z_v) \geq 0$.

Since $(z_{\mathrm{L}} \cdot \mathbf{1})$ corresponds to the homogeneous sensitivity distribution in which every agent has a sensitivity of S_{L}, this shows that the total latency of a heterogeneous Nash flow can never be worse than that of a low-sensitivity homogeneous Nash flow if $\kappa \leq \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}}$:

$$\max_{s \in \mathcal{S}} \mathcal{L}^{\mathrm{nf}}(G, s, \kappa) = \mathcal{L}^{\mathrm{nf}}(G, S_{\mathrm{L}}, \kappa).$$

Thus, for $\kappa \leq \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}}$, the worst-case Nash total latency for any population is realized by a population containing only one type, completing the proof. □

D. Proof of Lemma 1.3

The proof of Lemma 1.3 is straightforward; we show that for homogeneous populations with sensitivity s and scale factor $\kappa > 0$, the expression for the total latency is a 2nd-order rational function in $(s\kappa)$. This function possesses monotonicity properties that lead directly to the desired result.

For homogeneous $s \in \mathcal{S}_1$, every element of z is equal since every agent has the same sensitivity; i.e., for $s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]$ and $\kappa \geq 0$, $z = \frac{1}{1+s\kappa} \cdot \mathbf{1}$. By substituting this into (30), if we write $\Theta = -\mathbf{1}^{\mathrm{T}} b^T M = \mathbf{1}^{\mathrm{T}} M^T AM\mathbf{1} \geq 0$ (see Observation 1.2.1), we may explicitly write the total latency of a homogeneous Nash flow as

$$\begin{aligned} \mathcal{L}^{\mathrm{nf}}(G, s, \kappa) &= L_R + \frac{\mathbf{1}^{\mathrm{T}} M^T AM\mathbf{1}}{(1+s\kappa)^2} + \frac{b^T M\mathbf{1}}{1+s\kappa} \\ &= L_R - \frac{s\kappa}{(1+s\kappa)^2}\Theta. \end{aligned} \tag{35}$$

It is easy to verify that the minimum of (35) occurs whenever $\kappa = 1/s$, and is equal to $L_R - \Theta/4$. Furthermore, partial derivatives of (35) show that the worst-case total latency is *minimized* for some unique κ^* such that $\mathcal{L}^{\mathrm{nf}}(G, S_{\mathrm{L}}, \kappa^*) = \mathcal{L}^{\mathrm{nf}}(G, S_{\mathrm{U}}, \kappa^*)$. It can easily be verified from (35) that the solution to this equation is

$$\kappa^* = \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}}. \tag{36}$$

The partial derivatives of (35) with respect to κ also show that for any $\kappa \neq \kappa^*$,

$$\max_{s \in \mathcal{S}_1} \mathcal{L}^{\mathrm{nf}}(G, s, \kappa^*) < \max_{s \in \mathcal{S}_1} \mathcal{L}^{\mathrm{nf}}(G, s, \kappa).$$

Now we compute the price of anarchy resulting from tolls as defined in (36). Since we know that an un-tolled latency can never be more than $4/3$ times an optimal latency, from (35) we can write

$$\frac{\mathcal{L}^{\mathrm{nf}}(G, s, 0)}{\mathcal{L}^*(G)} = \frac{L_R}{L_R - \frac{1}{4}\Theta} \leq \frac{4}{3}. \tag{37}$$

This implies that $\Theta \leq L_R$, and it follows algebraically that for κ^* as defined in (36), $s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]$, and $\mathcal{G}$, the expression for the price of anarchy is given by (9). □