Title
Robust Methods for Influencing Strategic Behavior

Author
Brown, Philip Nathaniel

Publication Date
2018

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

Robust Methods for Influencing Strategic Behavior

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Philip Nathaniel Brown

Committee in charge:

Professor Jason R. Marden, Chair
Professor João P. Hespanha
Professor Ramtin Pedarsani
Professor Noah Friedkin
Professor Timothy Roughgarden, Stanford University

June 2018

The Dissertation of Philip Nathaniel Brown is approved.

Professor João P. Hespanha

Professor Ramtin Pedarsani

Professor Noah Friedkin

Professor Timothy Roughgarden, Stanford University

Professor Jason R. Marden, Committee Chair

May 2018

Robust Methods for Influencing Strategic Behavior

To Patricia and Genevieve

Acknowledgements

Thanks to everyone who walked with me on this PhD journey! First, to my advisor Jason, you've been the best of mentors. You have superbly modeled graduate school as an apprenticeship: you taught me what it means to be an academic by an expert combination of showing and telling. From the very beginning when you recruited me to CU, your guidance and partnership have been keys to my success. Now that I sit at the end of the road looking back, I cannot imagine a life without the struggles and triumphs that I've had in these past 6 years – and I cannot thank you enough for your role in starting, guiding, and keeping me on this path.

I have been fortunate to be able to immerse myself in two separate sets of faculty at two separate universities, and I cannot possibly hope to name all of the individuals among them who have meant something to me over these years. Many thanks go to my committee: João Hespanha, Ramtin Pedarsani, Noah Friedkin, and Tim Roughgarden – thanks for being part of this aspect of my PhD. Special thanks also to Behrouz Touri: you have always expressed more faith in me than I've felt like I gave you reason to, and this has been incredibly encouraging.

To my labmates from CU and UCSB (you know who you are), you were a huge part of my PhD experience and it wouldn't have been the same without you. It has been inspiring to work with peers who hold themselves to such consistently high standards; I'm confident I wouldn't have been as successful had you not been around.

As my preparation for this journey began long ago, I want to mention my mentors, business partners, and friends Mark Mauss and Travis Danner. My time changing the world (and shoveling chicken fat) alongside you prepared me for graduate school in many important ways. Your work ethic, intelligence, and love of family is always on my mind. To Professor Greg Durgin at Georgia Tech, you were my first academic mentor and a large part of the reason I considered entering graduate school in the first place. To my parents Greg and Nancy, you always gave me an environment in which my learning could thrive, and you have continued to be an inspiration and an encouragement ever since.

To my wife Patricia, I love you. Our entire married life has been centered on me getting my PhD, and now we get to embark on a new adventure together. To say that your constant care and support has been the key to my survival these past years would be the gravest understatement. Thank you for the beauty and the love that you have so freely brought into my life!

Finally, to my daughter Genevieve. You are not yet old enough to understand what it has meant for you that I've been pursuing my PhD, but I hope to teach you this for the rest of your life. May the life of learning that I have embarked upon be a light and inspiration to your life as well, and let the groundwork I have laid always spur you to challenge the limits of your own ambition and curiosity.

Curriculum Vitæ
Philip Nathaniel Brown

Education

2018	Ph.D. in Electrical and Computer Engineering (Expected), University of California, Santa Barbara.
2015	M.S. in Electrical Engineering, University of Colorado at Boulder.
2007	B.S. in Electrical Engineering, Georgia Institute of Technology.

Publications

Journal Publications

1. **P. N. Brown**, H. P. Borowski, and J.R. Marden, "Security Against Impersonation Attacks in Distributed Systems," accepted for publication in *IEEE Transactions on Control of Network Systems*, 2018. (to appear)
2. **P. N. Brown** and J.R. Marden, "Optimal Mechanisms for Robust Coordination in Congestion Games," accepted for publication in *IEEE Transactions on Automatic Control*, 2018. (to appear)
3. **P. N. Brown** and J.R. Marden, "The Robustness of Marginal-Cost Taxes in Affine Congestion Games," in *IEEE Transactions on Automatic Control*, August 2017.
4. **P. N. Brown** and J.R. Marden, "Studies on Robust Social Influence Mechanisms: Incentives for Efficient Network Routing in Uncertain Settings," in *IEEE Control Systems Magazine*, February 2017.

Proceedings of Refereed Conferences

1. **P. N. Brown** and J.R. Marden, "The Benefit of *Perversity* in Taxation Mechanisms for Distributed Routing," *56th IEEE Conference on Decision and Control*, 2017.
2. J.I. Poveda, **P. N. Brown**, J.R. Marden and A.R Teel, "A Class of Distributed Adaptive Pricing Mechanisms for Societal Systems with Limited Information," *56th IEEE Conference on Decision and Control*, 2017. [**Best Student Paper Finalist**]
3. **P. N. Brown** and J.R. Marden, "Studies on Mechanisms for Robust Social Influence," *1st IEEE Conference on Control Technology and Applications*, 2017.
4. **P. N. Brown** and J.R. Marden, "Fundamental Limits of Locally-Computed Incentives in Network Routing," *American Control Conference*, 2017.
5. **P. N. Brown** and J.R. Marden, "Avoiding Perverse Incentives in Affine Congestion Games," *IEEE Conference on Decision and Control*, 2016. [**Best Student Paper Finalist**]
6. **P. N. Brown** and J.R. Marden, "A Study on Price-Discrimination for Robust Social Coordination," *American Control Conference*, 2016.
7. **P. N. Brown** and J.R. Marden, "Optimal Mechanisms for Robust Coordination in Congestion Games," *IEEE Conference on Decision and Control*, 2015. (invited paper)
8. **P. N. Brown** and J.R. Marden, "Social Coordination in Unknown Price-Sensitive Populations," *IEEE Conference on Decision and Control*, 2013.

Patents

1. T. Danner, **P. N. Brown,** and M. Mauss. "Method of Alcoholysis of Fatty Acids and Fatty Acid Glycerides," 2016, U.S. Patent No. 9328054B1. May 3, 2016.

Awards and Honors

1. **2017 Best Student Paper Finalist,** *56th IEEE Conference of Decision and Control*
2. **2017 Outstanding TA Award, ECE Department,** UCSB
3. **2016 Best Student Paper Finalist,** *55th IEEE Conference of Decision and Control*
4. **2016 Researcher of the term,** Center for Control, Dynamical-Systems, and Computation, UCSB
5. **2016 Best talk in Session Award,** American Control Conference, Boston, MA
6. **2012 Chancellor's Fellowship,** University of Colorado.
7. **2007 ECE Senior Scholar Award,** Georgia Tech (highest GPA in final year of study)
8. **2007 J.E. McDaniel Award,** Georgia Tech (highest GPA among graduating co-op students)

Abstract

Robust Methods for Influencing Strategic Behavior

by

Philip Nathaniel Brown

Today's world contains many examples of engineered systems that are tightly coupled with their users and customers. In these settings, the strategic and economic behavior of users and customers can have a significant impact on the performance of the overall system, and it may be desirable for an engineer to devise appropriate methods of incentivizing human behavior to improve system performance. This work seeks to understand the fundamental tradeoffs involved in designing behavior-influencing mechanisms for complex interconnected sociotechnical systems. We study several examples and pose them as problems of game design: a planner chooses appropriate ways to select or modify the utility functions of individual agents in order to promote desired behavior. In social systems these modifications take the form of monetary or other incentives, whereas in multiagent engineered systems the modifications may be algorithmic. Here, we ask questions of sensitivity and robustness: for example, if the quality of information available to the planner changes, how can we quantify the impact of this change on the planner's ability to influence behavior? We propose a simple overarching framework for studying this, and then apply it to three distinct domains: incentives for network routing, distributed control design for multiagent engineered systems, and impersonation attacks in networked systems. We ask the following questions:

- *What features of a behavior-influencing mechanism directly confer robustness?*

  We show weaknesses of several existing methodologies which use pricing for congestion control in transportation networks. In response to these issues, we propose a universal taxation mechanism which can incentivize optimal routing in transportation networks,

requiring no information about network structure or user sensitivities, provided that it can charge sufficiently large prices. This suggests that large prices have more robustness than small ones. We also directly compare flow-varying tolls to fixed tolls, and show that a great deal of robustness can be gained by using a flow-varying approach.

- *How much information does a planner need to be confident that an incentive mechanism will not inadvertently induce pathological behavior?*

  We show that for simple enough transportation networks (symmetric parallel networks are sufficient), a planner can provably avoid perverse incentives by applying a generalized marginal-cost taxation approach. On the other hand, we show that on general networks, perverse incentives are always a risk unless the incentive mechanism is given some information about network structure.

- *How can robust games be designed for multiagent coordination?*

  We investigate a setting of multiagent coordination in which autonomous agents may suffer from unplanned communication loss events; the planner's task is to program agents with a policy (analogous to an incentive mechanism) for updating their utility functions in response to such events. We show that even when the nominal game is well-behaved and the communication loss is between weakly-coupled agents, there exists no utility update policy which can prevent arbitrarily-poor states from emerging. We also investigate a setting in which an adversary attempts to influence a distributed system in a robust way; here, by understanding susceptibility to adversarial influence, we hope to inform the design of more robust network systems.

Contents

Chapter 1

Introduction

Many of today's engineered systems are tightly interconnected with their users, and system performance often depends greatly on user behavior [2]. As a result, the traditional lines between engineering and the social sciences are becoming increasingly blurred. Analytical tools such as game theory are finding new applications in engineering [3, 4] and concepts from control theory are being applied to understand the dynamics of social systems [5–7]. It is now often insufficient to judge an engineered system on its technical merits alone, since strategic user behavior can lead to unpredictable and undesirable results [8]. Of particular importance to this dissertation are socially-integrated engineering problems in which users' social and strategic behavior has a significant impact on overall system performance [9, 10]. These types of systems appear in a variety of contexts in theory and practice: transportation networks [11–13], ridesharing applications [14, 15], supply-chain management [16], and electric power grids [17] are immediate examples. A common problem in these settings is that individual users' incentives may not be aligned with the objectives of the central planner. Thus, in addition to the merely-technical challenges they pose, an engineer may need to consider methods of influencing individual user behavior to effect positive change on aggregate system performance [18, 18–22].

This chapter contains material which is adapted, with permission, from [1], previously published by IEEE Control Systems.

In this context, we are interested in studying methods of indirectly influencing strategic and economic behavior by appropriately modifying the incentives that users face for their various decisions. This incentive-modification may take the form of traditional financial incentives (as considered in Part I), or it may take the more abstract form considered in Part II which we motivate with a problem of the distributed control of a multi-agent system. In any case, every behavior-influencing mechanism requires information about the underlying system and the user population who are to be influenced.

As a simple example, if a system planner desires to price a network resource to encourage efficient network usage, it may be desirable to characterize the sensitivity of the user population to pricing. If this information is difficult to gather or altogether unavailable, the planner may need to rely on crude estimates of user price-sensitivities, and the pricing design must take this uncertainty into account. At worst, a misunderstanding of informational dependencies can lead to "perverse incentives," or incentives that exacerbate the very problems they were intended to solve.

Here, a theory of "robust social influence" is an attractive goal: how can behavior-influencing mechanisms be designed so that they are robust to a variety of mischaracterizations or variations in models of social behavior or models of the underlying system? Some natural questions in this context include:

- How robust are existing behavior-influencing methodologies to variations in underlying system parameters?
- What features of a behavior-influencing mechanism directly confer robustness?
- How much information does a planner need to be confident that an incentive mechanism will not inadvertently induce pathological behavior?
- How can robust games be designed for multiagent coordination?

To address these questions, we begin by asking what elements a theoretical framework for robust incentive design should include. In essence, *what should robust incentives look* *like?* One typically imagines a system planner considering a particular instance of a system and then designing an incentive mechanism for that specific instance. In contrast, the core approach in this dissertation is to envision a system planner creating a *design methodology* that informs the construction of incentive mechanisms *for a family* of socio-technical systems, rather than for a single instance of that family.

1.1 The Robustness Meta-Problem

We now introduce the overarching analytical philosophy of this work. In this dissertation, the system planner is typically envisioned to be choosing rules that govern how a behavior-influencing mechanism is selected as a function of the realized details of a problem instance. If these rules (which we will frequently refer to as a *methodology*) are chosen in such a way that the behavior-influencing mechanism performs well in some sense on problems drawn from the considered family of problems, then the methodology can be said to be *robust.*

Somewhat more formally, let $\mathcal{G}$ represent a class (that is, a set) of problems, and let $\mathcal{T}$ be a class of behavior influencing mechanisms. For example, $\mathcal{G}$ may represent the class of all network traffic routing problems with linear latency functions, and $\mathcal{T}$ could represent the class of pricing mechanisms that assign fixed prices to network links. Each $T \in \mathcal{T}$ can be thought of as a set of rules for modifying problem instances in some admissible way. We write $T(G)$ to denote the problem instance G augmented by the behavior-influencing mechanism T. Using the routing problem example, $T(G)$ could represent the tolls specified by T for the routing problem instance G.

Given G and T, let $W(G, T(G))$ represent some measure of the quality of the emergent behavior resulting from the application of T to G.[1] In the context of routing problems, $W(G, T(G))$ could mean the social welfare of a Nash equilibrium[2] for routing problem G

[1]Here, we explicitly pass G as an argument to W to underscore the fact that in many settings "performance" is measured specifically with respect to the uninfluenced problem.

[2]Nash equilibrium and other technical details are formally defined in Chapter 2.

under the influence of the tolls $T(G)$. In essence, W is an objective function that the planner would like to maximize. Nominally, the planner's problem is to find a $T \in \mathcal{T}$ which maximizes W in worst case over $\mathcal{G}$:

$$\sup_{T\in\mathcal{T}} \inf_{G\in\mathcal{G}} W(G, T(G)). \tag{1.1}$$

The problem specified in (1.1) models design-time uncertainty by the variability among members of $\mathcal{G}$; a "larger" $\mathcal{G}$ would indicate more uncertainty at design-time.

In (1.1), the entire problem instance G is passed to the mechanism T. Taken at face value, this implicitly assumes that the only uncertainty is at design-time and that once the system is in operation, the mechanism T has access to all details of G. In many cases, some (but not all) of a problem instance's details are revealed at run-time, and it may be helpful to have a concise way to model this run-time uncertainty as well. To explicitly model the information available to mechanism T at run-time, let $\mathcal{I}$ denote a projection-like operator on $\mathcal{G}$ which functions as a mask, revealing only certain information about a problem instance to the associated mechanism T. Then the system planner's objective is to select a mechanism T that performs well on problems in $\mathcal{G}$, given the fact that T cannot observe all details of each problem $G \in \mathcal{G}$; we express this meta-objective with the quality measure Q defined as

$$Q_{\mathcal{G}}\left(\mathcal{T}, \mathcal{I}\right) := \sup_{T\in\mathcal{T}} \inf_{G\in\mathcal{G}} W\left(G, T(\mathcal{I}(G))\right). \tag{1.2}$$

One of the benefits of this formalism is that it provides an overarching framework with which a planner can map simple natural-language questions to concise, well-defined mathematical expressions. For example, consider a situation in which a planner wishes to know the value of investing in a new set of methodologies or tools. If the current set of methodologies is specified by $\mathcal{T}$ and the new methodologies are specified by $\mathcal{T}'$, then the value of incorporating the new methodologies is concisely given by

$$Q_{\mathcal{G}}\left(\mathcal{T} \cup \mathcal{T}', \mathcal{I}\right) - Q_{\mathcal{G}}\left(\mathcal{T}, \mathcal{I}\right) \geq 0. \tag{1.3}$$

The value of obtaining new information can be assessed similarly, by comparing $Q_{\mathcal{G}}(\mathcal{T}, \mathcal{I})$ for different specifications of uncertainty mask $\mathcal{I}$.

This dissertation contains a series of analytical studies on problems of a form inspired by the meta-problem (1.2). The overarching goal is not merely to provide a series of isolated answers, but rather to understand the fundamentals of the sensitivity of (1.2) in various contexts. In each of the studies contained herein, the specific questions being asked can be mapped to a problem of the form in (1.2). We will generally not provide this mapping explicitly, but we feel that presenting the meta-problem in this semi-formal way should help the reader to more quickly grasp the motivation and subtleties of the results. The remainder of this chapter is as follows: Section 1.2 gives an informal overview of the traffic routing problems that are studied later in the dissertation, and then Section 1.3 presents an informal listing of the dissertation's main results, along with references to specific theorems and the chapters containing them.

1.2 Selfish Routing as a Testbed for Robust Incentive Design

A classical example of an engineered system whose performance depends heavily on the choices of its users is that of a transportation network; this is captured in the literature by a problem known as a non-atomic routing game [23–25]. The basic problem setup is this: there is a group of travelers who need to be routed through a congestion-sensitive network in a way that minimizes the average travel time. It is typically straightforward to compute an optimal routing profile (also called a *network flow*), but implementing a particular flow could require that a central planner has the ability to force every driver to take a specified route. Unfortunately, it is well-known that if each driver chooses her route in order to individually minimize her own travel time, the resulting aggregate behavior can be substantially less efficient than the centrally-computed optimal flow [8].

For concreteness, here we sketch two example problems; these are among the main canonical routing problems which have received much attention in this domain. Both clearly demonstrate some of the challenges inherent in selfish routing problems, and set the stage for attempts to influence behavior to effect positive change.

1.2.1 Pigou's Example: the Inefficiency of Self-Interest

The first example, depicted in Figure 1.1, illustrates the basic problem that travelers' individual self-interested choices can lead to over-congested network flows. This illustration is generally attributed to the economist Arthur Pigou, and remains a centerpiece of work in this area [26]. The setting is as follows: there is 1 unit of travelers who can choose between two links connecting a source node and destination node. For simplicity, we assume that there are infinitely-many travelers, and that each individual is "small" in the sense that she has no impact on congestion. Each link has a latency function $\ell(f)$ which captures the delay experienced by users of the link as a function of mass of users choosing that link. In our example network, the latency function on link 1 is linear with $\ell_1(f_1) = f_1$, and the latency function on link 2 is constant with $\ell_2(f_2) = 1$. The first link offers a faster journey, provided that it is not chosen by too many users.

The flow on this network which minimizes the average delay is shown on the left in Figure 1.1: the traffic is split evenly between the two links, so that a mass of $1/2$ experiences a latency of $1/2$ on the top link, and the remaining traffic experiences a latency of 1 on the lower link, giving a total latency of $\mathcal{L}(f) = (1/2)^2 + 1/2 = 0.75$. However, this requires half the drivers to choose quite a long route; any individual driver on the link 2 has a compelling incentive to switch to link 1 and arrive at her destination in half the time. Unfortunately, if all drivers choose the route with the lowest latency, they will all crowd on to the upper link 1 and establish the flow depicted on the right in Figure 1.1. This flow is known as a *Nash*

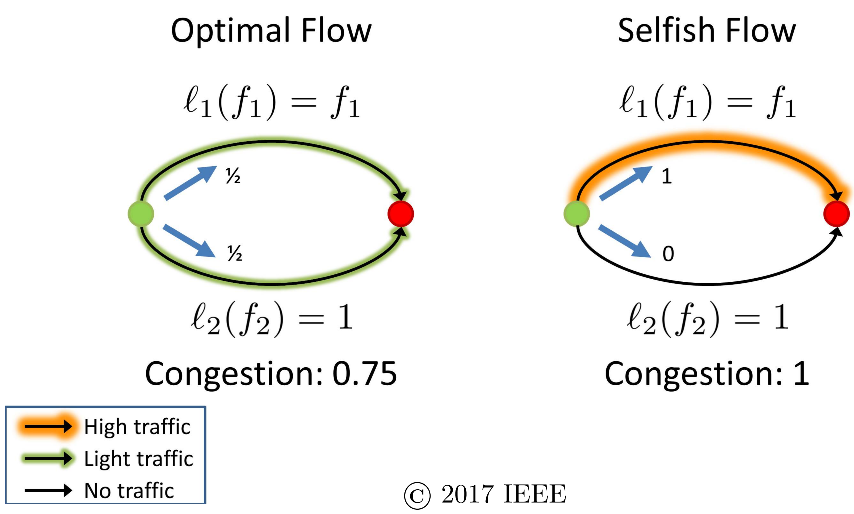

Figure 1.1: Pigou's Network, illustrating the negative effects of selfish behavior. In this network, drivers can choose between the upper, congestion-sensitive link and the lower, constant-latency link. The image on the left depicts a congestion-minimizing routing profile in which the traffic is split evenly between the two links. However, in this optimal flow, agents on the lower link experience a latency of 1, and (individually) could decrease their travel time by switching to the upper link. Unfortunately, this self-interested behavior can have negative consequences for system performance. The image on the right depicts a routing profile arising when every driver chooses the path with lowest delay; here, drivers have crowded on to the upper link, degrading its performance. A central problem is that no driver has an incentive to choose the lower, more efficient path.

flow or *Wardrop equilibrium*, defined as a flow in which no traveler can change paths and strictly decrease her cost. In this Nash flow, the network's total congestion is 1, a factor of 4/3 greater than the optimal total congestion.

1.2.2 Braess's Paradox: the Unintended Consequences of Naïve Influence

A second canonical example known as Braess's Paradox (first noted by Dietrich Braess [27, 28]) illustrates that seemingly-innocuous attempts to influence user behavior can lead to unexpected and perverse consequences. Consider the network depicted in Figure 1.2(a); traffic can choose between two paths, each routing through its own intermediate node. As-is, the total congestion on the network at Nash flow is 1.5, since half the traffic uses the upper path and half uses the lower path. It is also straightforward to see that this flow is optimal.

Suppose now that the system planner adds a single zero-cost link to the network connecting the two intermediate nodes to one another, as depicted in Figure 1.2(b). Now, under the old flow in which users split evenly, any user at node (B) would prefer to take the new zero-cost link rather than continue on the upper path (since this would entail a cost reduction of 1/2). This increases the lower path's congestion, causing more users at (A) to choose the upper path, but those users in turn will choose the new zero-cost link once they arrive at node (B). Ultimately, equilibrium is reached at the routing profile depicted in Figure 1.2(b), with a corresponding total congestion of 2. Here, this behavior-influencing mechanism (augmenting the network with a zero-cost link) backfired and caused a dramatic increase in total congestion [29]. This prompts the question: how should a system planner approach the problem of influencing behavior in a principled way?

1.2.3 Can pricing mitigate Braess's Paradox?

Motivated by the inefficiency resulting from selfish behavior, there has been a great deal of research on the application of road tolls (or other incentive mechanisms) for the purpose of

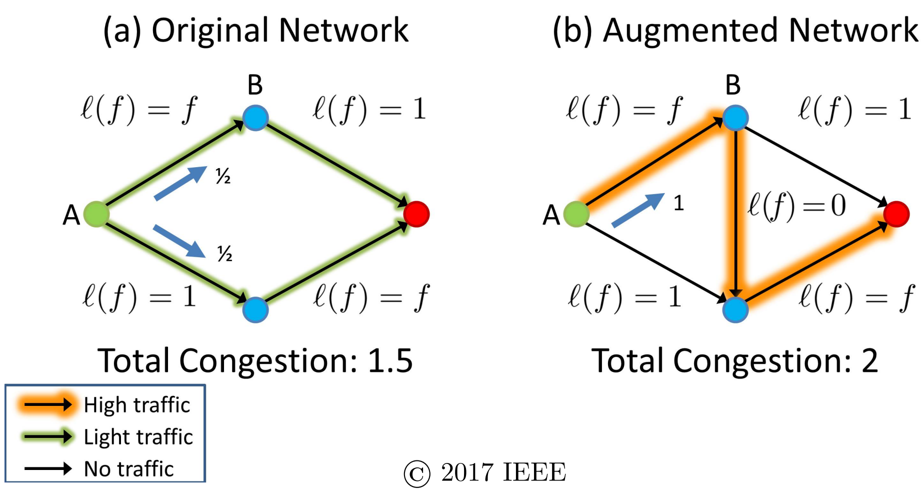

Figure 1.2: Braess's network, depicting an unintended consequence of attempting to influence social behavior. The image on the left in (a) depicts a transportation network and its associated Nash flow. A well-meaning traffic engineer, hoping to improve network congestion, adds a new link to the network connecting the two intermediate nodes. Despite the fact that this link's cost is zero, its addition to the network leads to the setting on the right in (b). Any user at node B can take the new link without increasing his cost, but in doing so, he increases the cost of the lower path, which in turn leads to more users at node A choosing the upper path. The ultimate effect of augmenting the network with a zero-cost link is that every driver's travel time increases by 33%.

influencing drivers to make routing choices that result in globally-optimal routing [29–38]. A simple pricing model assumes that a planner assigns each link in a network a pricing function $\tau(f)$, and then a driver on that link experiences a modified cost of $\ell(f) + \tau(f)$. How might prices be used to mitigate the problems seen in Braess's network?

One might postulate that the center zero-cost link is the cause of the problem, and so charge a single fixed price on that link, say a price of 1. This is essentially equivalent to removing the link from the network, since regardless of what others are doing, any agent at node (B) now weakly prefers the direct link to the destination (with cost 1) over the path containing the center link (whose cost is always at least 1).

Why might this be a poor choice? Suppose all of the parameters of this routing problem are fixed except the overall traffic rate. Rather than a total mass of 1 unit of traffic, suppose that there are $r > 0$ units of traffic. On this network, when $r < 1/2$, the optimal flow has all of the traffic using the zig-zag path, with a total latency of $2r^2$. However, the price on the center link effectively disincentivizes users from choosing this optimal flow, instead incentivizing an even split between the upper and lower links that has a total latency of $r^2 + r$, just as in Figure 1.2(a). For this low-r regime, the ratio of the total latency of the influenced flow and the total latency of the optimal flow is $\frac{1}{2} + \frac{1}{2r}$, which approaches ∞ as $r \to 0$.[3] In essence, this demonstrates that a fixed price of 1 on the center link in Braess's network is not *robust* to changes in traffic rate.

1.2.4 Robust Incentives for Selfish Routing

If road tolls are designed to incentivize good performance for one instance of a routing problem, and then some detail of the routing problem changes, it would be desirable for the original tolls to incentivize good performance on the changed routing problem as well. In Section 1.2.3 above, the total inflow of traffic changed, but we could consider robustness against a wide variety of other types of change: for example, a link is "removed" by a traffic

[3]This particular problem is investigated in considerably more detail in Chapter 3; in particular see Figure 3.2.

accident or natural disaster. That is, we would like to know if the performance guarantees provided by the original tolls are *robust* to changes or mischaracterizations in the underlying details of the system (e.g., network structure, traffic rate, user demands).

This goal can naturally be mapped into the robustness framework described by (1.2). Our typical approach is to select $\mathcal{G}$ to represent some class of routing problems (e.g., symmetric parallel-network routing problems with heterogeneous price-sensitive users), and then model the uncertainty of interest using the information mask $\mathcal{I}$. If we wish to understand robustness to variations in traffic rate, we select $\mathcal{I}$ to reveal all aspects of a routing problem instance *other than* the traffic rate.

1.3 Informal Overview of Findings

This dissertation is divided into two parts. Part I represents a logical grouping of work that deals specifically with robust pricing problems for purposes of mitigating congestion caused by selfish routing as described in Section 1.2. In Part II, we consider a somewhat more abstract "incentive" design problem motivated by the distributed control of a multiagent system, both from the perspective of a system designer (Chapter 8) and from the perspective of an adversary (Chapter 9). Here, we will survey the findings on a chapter-by-chapter basis.

1.3.1 Chapter 3: The Fragility of Fixed Tolls

Existing research has demonstrated that if a tax-designer has an accurate and detailed characterization of a routing problem, it is possible to design simple road tolls which incentivize optimal routing [39–41]. These are termed *fixed tolls*, since they are simple constant functions of traffic flow. One of the potential benefits of fixed tolls is that a single fixed price can easily and clearly be communicated to drivers, and this may be preferable to more complicated variable or contingent prices. However, as shown above in section 1.2.3, fixed tolls need not be robust to variations of network parameters. We show this more comprehensively and for more types of parameter variations in Chapter 3:

- Proposition 3.1 shows that if the latency/delay functions of a network are not characterized precisely, fixed tolls cannot guarantee optimal routing on that network.
- In an example depicted in Figure 3.2, we investigate the problem considered in Section 1.2.3 and show how fixed tolls can cause unbounded degradation in performance when the network's traffic rate changes.
- Finally, Theorem 3.2 shows that not only must fixed tolls depend on network structure to incentivize *optimal* flows, but that they must depend on network structure even to guarantee that they incentivize *improvements* in network flows.

It is simple to cast these results in the meta-framework of (1.2): each posits a particular class of routing problems (here, generally selfish routing problems with populations having homogeneous price-sensitivity), each posits a class of taxation mechanisms $\mathcal{T}$ subject to constraints (here, generally fixed tolls $\mathcal{T}_{\text{fixed}}$), and each posits an information mask $\mathcal{I}$ which specifies the information available to the taxation mechanism. Thus, since these are all negative results, they imply that even for simple $\mathcal{G}$, we have that for many different $\mathcal{I}$, the quality given by $Q_{\mathcal{G}}\left(\mathcal{T}_{\text{fixed}}, \mathcal{I}\right)$ is poor.

1.3.2 Chapter 4: The modest benefits of marginal-cost pricing

Another well-studied taxation mechanism is that of *marginal-cost tolls*, in which each network link is assigned a flow-varying tax that is specifically designed to penalize inefficient congestion. Marginal-cost tolls are known to incentivize optimal network routing in the special case that all network users are homogeneous in tax-sensitivity (i.e., they all value time equally) [42,43]. If users are heterogeneous in price-sensitivity (i.e., different users value their time differently), then marginal-cost pricing can still help, provided that each user is charged a personalized price that is "tuned" to her particular price-sensitivity [32]. If the pricing authority lacks access to a detailed characterization of the users' price-sensitivities or lacks the ability to charge each user an individualized price, what can be said about the effectiveness of marginal-cost pricing?

In this chapter, we investigate the robustness of marginal-cost tolls in settings with price-sensitive and heterogeneous user populations. In summary:

- In an example presented in Figure 4.1, it is demonstrated how marginal-cost tolls generally fail to incentivize optimal routing when the price-sensitivity of the user population is not precisely characterized.
- Proposition 4.1 shows, again via an example, that off-the-shelf marginal-cost tolls can actually *degrade* routing efficiency when the user population is heterogeneous.
- In contrast to these generally negative results, Theorem 4.2 considers a restricted problem (parallel networks, linear-affine cost functions, and high traffic) and derives the optimal scaled marginal-cost toll that minimizes worst-case congestion resulting from heterogeneous populations.

A message here is that while it is still possible to construct pathological examples on which marginal-cost tolls perform poorly, they are strictly more robust than the fixed tolls considered in Chapter 3. Furthermore, by appropriately restricting the class of problems under consideration, it can be shown (as in Theorem 4.2) that marginal-cost tolls can be quite robust.

With respect to the meta-problem (1.2), here we investigate the sensitivity of the quality metric to changes in $\mathcal{G}$, while holding the class of taxation mechanisms $\mathcal{T}$ and available information $\mathcal{I}$ constant. That is, the fundamental question of this chapter is on the role of design-time uncertainty.

1.3.3 Chapter 5: Advanced techniques for toll robustness

In this chapter, we ask if there exist any taxation mechanisms that are robust in the information-denied environments studied earlier. Fortunately, there are – but we show that these bring their own drawbacks:

- Theorem 5.1, the cornerstone of the chapter's results, shows that if tolls can be made large, there is a universal taxation mechanism that can incentive approximately-optimal flows without requiring information about the structure of the network or about the user population's price-sensitivity.

- Inspired by the fact that Theorem 5.1 prescribes arbitrarily-large tolls, we next consider the effect of placing an upper bound on tolls. Theorem 5.3 considers the same restricted setting seen in Theorem 4.2 and derives the optimal bounded taxation mechanism for that setting; interestingly, when latency functions are linear-affine, the optimal taxation mechanism also must be affine.

- Finally, Theorem 5.4 demonstrates a curious fact that the worst-case inefficiency due to *fixed* tolls can actually be expressed as the worst-case inefficiency due to the bounded tolls of Theorem 5.3 *for a very low toll bound*. Since efficiency is increasing in the toll upper bound, this serves as an additional indication that fixed tolls are considerably less robust than flow-varying tolls.

The work contained in this chapter should be viewed as an initial piece of a larger research agenda, to understand the relationship between constraints on information (i.e., uncertainty) and the constraints on taxation mechanisms, as discussed earlier with the meta-problem (1.2). In the meta-problem framework of (1.2), this chapter's primary question is on the sensitivity of the quality metric to changes in the allowable class of taxation mechanisms $\mathcal{T}$. Here we hold the class of routing problems $\mathcal{G}$ constant, and parameterize $\mathcal{T}$ by an upper bound on taxation functions to show how enlarging the set of available methodologies can improve performance.

1.3.4 Chapter 6: Avoiding perverse incentives

The foregoing was largely concerned with comparing the efficiency incentivized by a taxation mechanism to the efficiency of optimal routing profiles. Alternatively, a pricing designer may be interested in comparing the flows incentivized by a taxation mechanism with the uninfluenced flows. In this context, if for some routing problem a taxation mechanism incentivizes a Nash flow which is *worse* than the uninfluenced Nash flow, we call this mechanism *perverse.* This chapter formalizes this notion of *perverse incentives*, and asks which taxation mechanisms can be said to systematically avoid perverse incentives. One of the key insights here is that there is a fundamental tradeoff between minimizing congestion and avoiding perverse incentives; in essence, one goal cannot be accomplished without making some concession to suboptimality in the other goal.

- First, Theorem 6.1 shows this tradeoff qualitatively: if a taxation mechanism improves outcomes on any network, it must degrade them on some other network.

- Next, we show in Theorem 6.2 that perverse incentives are not totally ubiquitous. That is, there exist nontrivial classes of networks (e.g., parallel networks) on which perverse incentives can always be avoided with nontrivial taxation mechanisms.

- Theorem 6.5 and Proposition 6.6 are a step towards characterizing the perversity tradeoff quantitatively, on routing games with homogeneous populations. Here we provide tools that can help a tax designer minimize worst-case congestion while limiting the perversity of the applied mechanism.

- Lastly, we show in Corollary 6.7 that this work on perverse incentives has interesting implications for the theory of altruistic behavior in congestion games. In particular, our impossibility result from Theorem 6.1 implies that there exist "altruism paradoxes": routing problems for which making a fraction of the population *more* altruistic can actually increase aggregate traffic congestion.

1.3.5 Chapter 7: Price discrimination

In this final chapter on influencing selfish routing, we consider discriminatory pricing; i.e., pricing mechanisms that charge different prices to different users. One of the motivations here is to explore some of the consequences of modifying our design constraints.

- Theorem 7.1 is similar in spirit to our previous Theorem 5.1 for large tolls: if price-discrimination is fine enough (that is, enough distinct prices can be charged to different types of users), then arbitrarily-efficient Nash flows can be incentivized.

- Theorem 7.2 shows that there is a general equivalence between fine price discrimination and low variance in price sensitivity among the members of the user population. That is, fine price-discrimination on a poorly-characterized population gives the same worst-case performance as no price-discrimination on a well-characterized population.

- In Theorem 7.3, we apply the optimal bounded taxation mechanism from Chapter 5 (as considered in Theorem 5.3) to a discriminatory setting, and show how to compute the optimal discriminatory prices for any parallel-network routing problem.

1.3.6 Chapter 8: Applying the framework to a distributed control problem

Until here, the sole application of our work has been to mitigate traffic congestion in selfish routing problems. In Chapter 8, we turn to quite a different setting: distributed control of a multiagent system, posed as a problem of game design. Recent years have witnessed a surge of interest in using game theory to design distributed control laws for multiagent systems; overviews of some of this work can be found in [4]. In this paradigm, utility functions are assigned to decision-makers in a multiagent system; these utility functions induce a game between the agents. This paradigm has been proposed to address problems in many different domains, including sensor networks [44, 45], UAV control [46], and wireless spectrum sharing [47,48]. The agenda in this literature is to develop methodologies for designing utility functions that use information available from the nominal engineering problem.

One significant open question that remains in this area centers on the networked aspect of the designed game. The classical results here essentially assume the agents have a full specification of their utility functions *and* can accurately observe the action choices of all other agents at all times or can communicate among each other [49–52]. In other words, the graph that describes which agents can observe one another is a complete graph with edges connecting every pair of agents. If the observation graph is not complete, it is not known how to compute utility functions which induce good Nash equilibria.

Chapter 8 investigates a setting which may provide some insight into general answers to this open question. Here, we assume that the planner is given a game which is played on a complete observation graph, and that this game's Nash equilibria correspond to the optimizer of an objective function of interest. We then investigate the removal of a single directed edge from the observation graph; in essence, we are asking if a distributed control algorithm that is designed for a complete observation graph structure can be "projected" in a principled way onto a game with a sparser observation graph.

Crucially, when an edge is removed from the observation graph, this prevents some agent from evaluating its nominal utility function — so a new utility function must be assigned to that agent. Thus, we are seeking a robust means of assigning that agent a new utility function under the restriction that the new utility function has no information about the utility functions of other agents. That is, this problem maps to the robustness meta-problem of (1.2), with $\mathcal{T}$ representing sets of rules that govern the assignment of new utility functions, and $\mathcal{I}$ encoding the above informational restriction.

This chapter presents several negative results, showing that no robust utility reassignment is possible in many situations. Subsequently, we undertake an initial study on how it may be possible to circumvent these pathologies. In particular:

- Theorems 8.3 and 8.7 are the main negative results of this chapter. Informally, they show that nominally well-behaved games (potential games with unique Nash equilibria, or identical-interest games, respectively) can be made arbitrarily bad simply by the "removal" of a single directed edge in their respective interdependency graphs, even when that edge corresponds to a pair of agents that are very weakly interdependent.

- Theorem 8.9 explores a possible path to mitigating some of these pathologies via an informational paradox. Here, we show specifically for identical-interest games that if all interdependency connections to a single player are removed, this will essentially eliminate the harm of the pathologies described in Theorem 8.7.

1.3.7 Chapter 9: Robust influence by an adversary

No study on influencing social behavior would be complete without some mention of the influence that an adversary might have on behavior. In Chapter 9, we consider a simple model of coordination in a networked social or engineered system and ask how an adversary can influence the stable states of the system. In this setting, we assume that the adversary can introduce a fixed number of counterfeit nodes into the network which each impersonate a friendly node, effectively modifying the true nodes' utility functions in an attempt to indirectly influence their actions.

Our main question here is this: how does the capability of the adversary and the amount of information available to it affect its ability to influence behavior? Here, we model capability both by how precisely the adversary can target individual nodes, as well as how many nodes the adversary can target. In the language of the meta-problem (1.2), we play the role of the adversary and ask how the network can be influenced robustly. We posit three distinct possibilities for $\mathcal{T}$ (corresponding to static targeting, mobile targeting, and random targeting) and compute the adversary's ability to influence as a function of k, the number of friendly nodes it can target. The main results are

- Theorem 9.2 shows that if an adversary employs random targeting, its long-term influence on agent behavior is constant in k – suggesting that randomness is of high relative value when the adversary cannot influence many agents. This holds for any graph structure.

- Theorems 9.1 and 9.4 compare fixed and mobile deterministic adversaries for the special case of ring graphs, and show precise susceptibilities as a function of the adversaries' value of k and the size of the graph n. Particularly interesting is that for ring graphs of size $n > 3$, a mobile adversary is equally effective with $k = 3$ as it is with $k = n - 1$, starkly demonstrating the value of mobility.

1.3.8 Remarks

In the remainder of this dissertation, we develop the theory surrounding these contributions. Chapter 2 provides the complete technical details of the traffic routing model that is studied in Chapters 3–7. When an individual chapter or set of results requires a refinement or modification to the basic model, we make that refinement in the chapter in question. Chapter 8 then considers the design of a network game as a type of incentive design problem in which the "incentive mechanism" is a policy by which an autonomous agent modifies its utility function in response to the loss of some piece of information. Finally, Chapter 9 considers the impact of adversaries in distributed systems. Here, the adversary is solving an incentive design problem: by posing as a friendly agent, it can indirectly modify friendly agents' behavior. As their models differ significantly from that of the routing problems, Chapters 8 and 9 are nearly completely self-contained, each with their own complete model descriptions.

Part I

Robust Incentives for Selfish Routing

Chapter 2

Technical Preliminaries for Selfish Routing

In this chapter, we introduce the key technical details of the selfish routing model that is studied in Chapters 3–7. When a chapter specifically considers a variation of this model, we make this variation explicit in the chapter itself.

2.1 The Model

Consider a network routing problem for a network (V, E) comprised of vertex set V and edge set E. We call a source/destination vertex pair $(\sigma^c, t^c) \in (V \times V)$ a *commodity*, and the set of all such commodities $\mathcal{C}$. For each commodity $c \in \mathcal{C}$, there is a mass of traffic $r^c > 0$ that needs to be routed from σ^c to t^c. We write $\mathcal{P}^c \subset 2^E$ to denote the set of *paths* available to traffic in commodity c, where each path $p \in \mathcal{P}^c$ consists of a set of edges connecting σ^c to t^c. Let $\mathcal{P} = \cup\{\mathcal{P}^c\}$. A network is called *symmetric* if there is exactly one commodity: $\mathcal{C} = \{c\}$; i.e., all traffic routes from a common source σ to a common destination t using a common path set $\mathcal{P}$. A network is called a *parallel* network if all commodities share a single source-destination pair and all paths are disjoint; i.e., for all paths $p, p' \in \mathcal{P}$, $p \cap p' = \emptyset$. Note that a parallel network need not be symmetric; although all traffic must share a common source and destination, the various commodities' path sets $\mathcal{P}^c$ may still differ.

We write $f_p^c \geq 0$ to denote the mass of traffic from commodity c using path p, and $f_p \triangleq \sum_{c \in \mathcal{C}} f_p^c$. A *feasible flow* $f \in \mathbb{R}^{|\mathcal{P}|}$ is an assignment of traffic to various paths such that for each c, $\sum_{p \in \mathcal{P}^c} f_p^c = r^c$ and $\sum_{c \in \mathcal{C}} r^c = r$.

Given a flow f, the flow on edge e is given by $f_e = \sum_{p: e \in p} f_p$. To characterize transit delay as a function of traffic flow, each edge $e \in E$ is associated with a specific latency function $\ell_e : [0, r] \to [0, \infty)$; $\ell_e(f_e)$ denotes the delay experienced by users of edge e when the edge flow is f_e. We adopt the standard assumptions that each latency function is nondecreasing, convex, and continuously differentiable. We measure the cost of a flow f by the *total latency*, given by

$$\mathcal{L}(f) = \sum_{e \in E} f_e \cdot \ell_e(f_e) = \sum_{p \in \mathcal{P}} f_p \cdot \ell_p(f), \tag{2.1}$$

where $\ell_p(f) = \sum_{e \in p} \ell_e(f_e)$ denotes the latency on path p. We denote the flow that minimizes the total latency by

$$f^* \in \underset{f \text{ is feasible}}{\operatorname{argmin}} \ \mathcal{L}(f). \tag{2.2}$$

A *routing problem* is given by $G = (V, E, \mathcal{C}, \{\ell_e\})$; the shorthand $e \in G$ means $(e \in E : E \in G)$. We denote classes of routing problems with the calligraphic $\mathcal{G}$, and often write $e \in \mathcal{G}$ as a shorthand for $(e \in G : G \in \mathcal{G})$.

To study the effect of taxes on self-interested behavior, we model the above routing problem as a non-atomic congestion game. We assign each edge $e \in E$ a flow-dependent taxation function $\tau_e : \mathbb{R}^+ \to \mathbb{R}$. We characterize the taxation sensitivities of the users in commodity c with a monotone, nondecreasing function $s^c : [0, r^c] \to [S_{\rm L}, S_{\rm U}]$, where each user $x \in [0, r^c]$ has a taxation sensitivity $s_x^c \in [S_{\rm L}, S_{\rm U}] \subseteq \mathbb{R}^+$, where $S_{\rm L} \geq 0$ and $S_{\rm U} \leq +\infty$ are lower and upper sensitivity bounds, respectively. To avoid trivialities, we assume that in each c, zero measure of traffic has sensitivity exactly equal to 0, or that for all $\epsilon > 0$,

$s^c_\epsilon > 0$. If all users have the same sensitivity (i.e., $s^c_x = s^{c'}_y$ for all $c, c' \in \mathcal{C}$ and all $x \in [0, r^c]$ and $y \in [0, r^{c'}]$), the population is said to be *homogeneous*; otherwise it is *heterogeneous*.

Given a flow f, the cost that user $x \in [0, r^c]$ experiences for using path $\tilde{p} \in \mathcal{P}^c$ is of the form

$$J^c_x(f) = \sum_{e \in \tilde{p}} \left[\ell_e(f_e) + s^c_x \tau_e(f_e)\right]. \tag{2.3}$$

Thus, for each user $x \in [0, r^c]$, the sensitivity s^c_x can be viewed as a constant gain on the toll; a user's experienced cost is then the sum of the latency and sensitivity-weighted toll. Note that sensitivity can be interpreted as the reciprocal of an agent's value-of-time.[1] We assume that each user selects the lowest-cost path from the available source-destination paths. We call a flow f a *Nash flow* if all users are individually using minimum-cost paths given the choices of other users, or if for all commodities $c \in \mathcal{C}$ and all users $x \in [0, r^c]$ we have

$$J^c_x(f) = \min_{p \in \mathcal{P}^c} \left\{ \sum_{e \in p} \left[\ell_e(f_e) + s^c_x \tau_e(f_e)\right] \right\}. \tag{2.4}$$

It is well-known that a Nash flow exists for any non-atomic congestion game of the above form [54].

In our analysis, we typically assume that each sensitivity distribution function s^c is unknown to the pricing authority; for a given routing problem G and $S_\mathrm{U} \geq S_\mathrm{L} \geq 0$ we define the set of possible sensitivity distributions as the set of monotone, nondecreasing functions $\mathcal{S}_G = \{s^c : [0, r^c] \to [S_\mathrm{L}, S_\mathrm{U}]\}_{c \in \mathcal{C}}$. We write $s \in \mathcal{S}_G$ to denote a specific collection of sensitivity distributions, which we term a *population*.

2.2 The Robustness of a Taxation Mechanism

It was shown in Section 1.2 how self-interested behavior can lead to poor system performance; in both examples, the fundamental problem was that in an optimal flow, users

[1] We adopt this formulation from [39]. Note that constant sensitivity is a commonly-studied special case; alternative formulations are possible [53].

could decrease their own cost only by imposing a greater cost on those around them. If users were altruistic, willing to accept a personal degradation of service for the sake of the greater good, they might be expected to adopt this type of socially-optimal configuration. One way to influence users to choose this configuration is to charge tolls on over-congested links, hoping to increase their costs enough that a sufficient number of users will avoid them. The tolls are put in place essentially to induce an “artificial altruism” in the population; in fact, there are many parallels between the literature on altruism in congestion games and the literature on financial incentives in congestion games (some of which we explore in Chapter 6) [55]. Of course, the toll-designer must take care in choosing the tolls. It has already been seen in Braess’s Paradox that seemingly-innocuous approaches can have unexpected consequences, and the toll-designer must be certain not to fall into a similar trap. If tolls are too high on a particular edge, it may be that too many users will avoid that edge; if tolls are not properly balanced throughout the network, uneven and inefficient flow distributions could arise.

2.2.1 A Simple Robustness Taxonomy

It is the goal of robust social influence design to levy tolls that incentivize desirable behavior irrespective of changes or mischaracterizations of the underlying system. Figure 2.1 depicts several types of system changes which could potentially create problems for taxation methodologies. In these diagrams, the tolls were designed for the nominal system on the left, but after the respective change, these same tolls are effectively being applied to different networks than those for which they were designed. The hope is that the tolls designed for the original system provide comparable performance guarantees on the “new” systems; to this end, we will ask if each of several common taxation methodologies is robust to variation in parameters such as:

1. Network Changes: If tolls are designed to incentivize efficient flows for a particular network, and the network undergoes some change, do the original tolls still incentivize

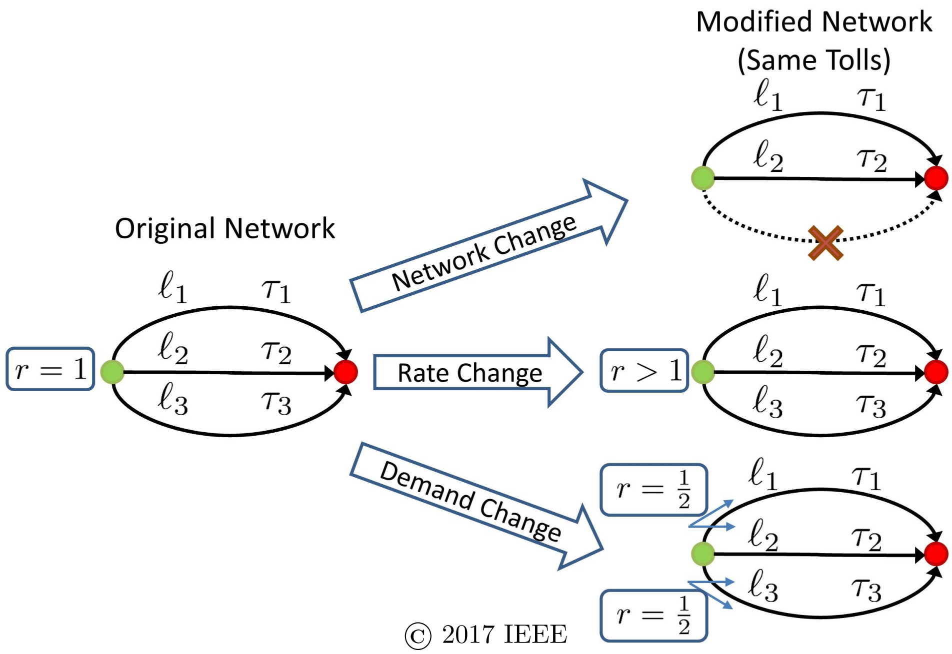

Figure 2.1: Diagram depicting possible network changes a designer may consider in robustness analysis. Suppose tolls τ_1, τ_2, andτ_3 are designed for the network on the left. To analyze the robustness of this toll design, the designer may subject the routing problem to various hypothetical changes: topology (deleting link 3), traffic rate (increasing r above its original value), and demand structure (restricting half the traffic to the upper two links, the other half of the traffic to the lower two links), while keeping the tolls designed for the original network. If the tolls are able to incentivize efficient behavior on the "new" networks despite having been designed for the original network, they are called robust.

efficient flows for the new, changed network?

2. Traffic Rate: Do tolls designed for one traffic rate (i.e., one value of r) still incentivize efficient flows if the rate changes?

3. Demand Structure: If some users have access to different paths than others, how does this impact the design of the correct tolls?

To investigate the robustness of a particular tolling strategy to variations of a parameter, a system-planner can design tolls for a specific system realization, hold the tolls constant, and study the effect on Nash total latency of varying the parameter in question. One way to model perturbations of games is to define a correspondence $\Gamma_G(\cdot)$ that returns games similar to some nominal game G that have been perturbed in the argument of Γ_G. For example, $\Gamma_G(\{\ell, r\})$ represents the set of routing games that differ from G only in their latency functions ℓ and total traffic rate r.

To formally discuss tolling strategies, let a *taxation mechanism* T be a mapping from games to edge taxation functions; thus, we write $T(G)$ to denote the edge taxes that T assigns to game G. We write $\mathcal{L}^{\text{nf}}(G^*, T(G))$ to mean the Nash flow total latency for some (possibly different) game G^* induced by tolls generated by T for the nominal game G. A taxation mechanism T is said to be *strongly robust on* G if T incentivizes optimal flows for all allowable perturbations of G. That is, the strong robustness of T to variations in game parameters $X \subseteq \{V, E, \mathcal{P}, \{\ell_e\}_{e\in E}, r, s\}$ implies that

$$\mathcal{L}^{\text{nf}}(G^*, T(G)) = \mathcal{L}^*(G^*) \text{ for all } G^* \in \Gamma_G(X). \tag{2.5}$$

We likewise say that T is strongly robust on a larger class of games $\mathcal{G}$ if (2.5) holds for all $G \in \mathcal{G}$.

This may be too strong a condition in some settings (though strongly-robust taxation mechanisms do exist in some contexts; see Section 2.3.2), so a taxation mechanism is said to be *weakly robust on* G if it never incentivizes Nash flows on perturbed networks that are *worse* than the un-tolled flows. That is, writing $\mathcal{L}^{\text{nf}}(G, \emptyset)$ to denote the Nash flow total latency on G without tolls, the weak robustness of T on G to parameters specified by X implies that

$$\mathcal{L}^{\text{nf}}(G^*, T(G)) \leq \mathcal{L}^{\text{nf}}(G^*, \emptyset) \text{ for all } G^* \in \Gamma_G(X). \tag{2.6}$$

Again, we say that T is weakly robust on a larger class of games $\mathcal{G}$ if (2.6) holds for all $G \in \mathcal{G}$. Put differently, if tolls are weakly robust, they will not create perverse incentives. Thus, by the definitions presented in this section, assigning a toll of 0 to every link is always a weakly robust taxation mechanism, as it certainly cannot make Nash flows worse.

2.2.2 Price of Anarchy

While the strong-weak taxonomy presented above may provide a simple binary check on the robustness of a mechanism, it is often desirable to measure robustness more quantitatively. One basic optimization objective which lends itself well to this goal is termed the *price of anarchy*. Informally, price of anarchy is a measure of how much worse Nash flows can be than optimal flows. For some class of routing problems $\mathcal{G}$, let $\mathcal{L}^{\rm nf}(G)$ and $\mathcal{L}^*(G)$ denote the total latencies of Nash flows and optimal flows on $G \in \mathcal{G}$, respectively. In its simplest form, the price of anarchy is defined over a class of games $\mathcal{G}$ as

$$\mathrm{PoA}(\mathcal{G}) = \sup_{G \in \mathcal{G}} \frac{\mathcal{L}^{\rm nf}(G)}{\mathcal{L}^*(G)}. \tag{2.7}$$

Price of anarchy was introduced in [56] and some of the first comprehensive analysis was done in the area of selfish routing games in [8,57,58]. Price of anarchy has found far-reaching applications in fields as diverse as supply-chain management [59], auction theory [60, 61], telecommunication systems [62], machine scheduling [63], distributed control [64] and a variety of networking problems [9, 65].

Note the similarity between (2.7) and the inner optimization of the overarching meta-problem (1.2); because of this similarity, we will use a price-of-anarchy-like object to assess the robustness of taxation mechanisms.[2]

For a given routing problem $G \in \mathcal{G}$, we gauge the efficacy of a collection of taxation functions $\tau = \{\tau_e\}_{e \in E}$ by comparing the total latency of the resulting Nash flow and the total latency associated with the optimal flow, and then performing a worst-case analysis over all possible user populations. For some routing problem G, let $\mathcal{L}^{\rm nf}(G, s, \tau)$ denote the total latency of the worst-performing Nash flow resulting from taxation functions τ and population $s \in \mathcal{S}_G$. To assess the impact of uncertainty on user sensitivities, we often use

[2] In so doing, we frequently overload the $\mathrm{PoA}(\cdot)$ notation; we shall endeavor to explicitly state the meaning of PoA when not abundantly clear from context.

the following formulation:

$$\mathrm{PoA}(G,\tau) = \sup_{s\in\mathcal{S}_G} \frac{\mathcal{L}^{\mathrm{nf}}(G,s,\tau)}{\mathcal{L}^*(G)} \geq 1. \tag{2.8}$$

Note that (2.10) considers the cost of routing induced by taxes τ on *a specific routing problem* (i.e., fixed network (V,E), commodities $\mathcal{C}$, and latency functions $\{\ell\}$) in the situation that the sensitivities s of the user population are unknown. Hence, this particular formulation is assessing the cost associated with not knowing the distribution of user sensitivities; other formulations are possible and we will introduce them as necessary.

2.2.3 Network-Agnosticity and Perverse Incentives

One type of taxation mechanism which we will consider repeatedly is a *network-agnostic* taxation mechanism. This concept was first formalized in [1]; here, each edge's taxation function is computed using only locally-available information. That is, $\tau_e(f_e)$ depends only on ℓ_e, not on edge e's location in the network, the network topology, the overall traffic rate, or the properties of any other edge. A network-agnostic taxation mechanism T is thus a mapping from latency functions to taxation functions, and the taxation function associated with latency function ℓ_e is given by

$$\tau_e(\cdot) = T(\ell_e). \tag{2.9}$$

Network-agnostic taxation mechanisms provide a way to formally study the value of information about network structure when designing behavior-influencing mechanisms. Here, if a network-agnostic taxation mechanism performs well, it can be inferred that information about network structure is not particularly valuable. On the other hand, if all network-agnostic taxation mechanisms perform poorly, this indicates that information about network structure is quite valuable indeed.

Here, we are often interested in formulating the price of anarchy in the following way:

the price of anarchy of a class of games $\mathcal{G}$ under the influence of taxes T is defined as

$$\mathrm{PoA}\,(\mathcal{G},T) \triangleq \sup_{G\in\mathcal{G}} \sup_{s\in\mathcal{S}_G} \frac{\mathcal{L}^{\mathrm{nf}}(G,s,T)}{\mathcal{L}^*(G)}. \tag{2.10}$$

However, the price of anarchy itself need not be the only metric of concern. In Chapter 6, we pose a somewhat different question: rather than measuring how far the influenced flows are from *optimal*, it may be desirable to measure how much the taxes help (or hurt) *with respect to the un-influenced flows*. Here, we may define a similar metric to that in (2.10), but with the un-influenced total latency $\mathcal{L}^{\mathrm{nf}}(G,\emptyset)$ in the denominator. We call such a metric the *index of perversity* of taxation mechanism T, defined as

$$\mathrm{PI}\,(\mathcal{G},T) \triangleq \sup_{G\in\mathcal{G}} \sup_{s\in\mathcal{S}_G} \frac{\mathcal{L}^{\mathrm{nf}}(G,s,T)}{\mathcal{L}^{\mathrm{nf}}(G,\emptyset)}. \tag{2.11}$$

Here, if T has a large index of perversity, this means that on some networks, it incentivizes flows that are considerably worse than the un-influenced flows; in other words, it can create perverse incentives. If a taxation mechanism has a perversity index of 1, we say that it is *non-perverse*; otherwise, it is *perverse*.

Note that it is always true that $\mathrm{PI}(\mathcal{G},T) \leq \mathrm{PoA}(\mathcal{G},T)$; this is because on any G, $\mathcal{L}^{\mathrm{nf}}(G,\emptyset) \geq \mathcal{L}^*(G)$. Finally, when these metrics are evaluated only over homogeneous populations (as opposed to heterogeneous), we write them as $\mathrm{PoA}^{\mathrm{hm}}(\mathcal{G},T)$ and $\mathrm{PI}^{\mathrm{hm}}(\mathcal{G},T)$, respectively.

2.3 Survey of Existing Taxation Methodologies

2.3.1 Fixed tolls for designers with detailed information

A simple way to apply tolls for social coordination would simply be to charge all users of each link a fixed price. Tolls of this form are known as "fixed" tolls, since the tolling function on each edge is a constant function of edge flow. To see an example of fixed tolls consider again Pigou's Example in Figure 1.1. If all users have a tax-sensitivity equal to 1, one set of edge tolls that enforces the optimal flow for Pigou's example is simply $\tau_1(f_1) = 0.5$, and $\tau_2(f_2) = 0$. Under these tolls, the optimal flow of $(1/2, 1/2)$ is a Nash flow, since in it all users experience a cost of 1. If users are heterogeneous, an optimal fixed toll can still easily be found by charging the price that would cause the most sensitive half of the users to deviate to the lower link. This fixed-tolling approach has been studied in general, and it is known that fixed tolls can be computed to enforce *any* feasible flow, provided that the system planner has a complete characterization of the system: network topology, user demand profile, latency functions, and user sensitivities [40, 41].

Note that in the language of the robustness meta-problem, these fixed-toll results essentially allow the class of problems $\mathcal{G}$ to be quite large, but then take the information mask $\mathcal{I}$ to be the identity: the taxation mechanism can specify tolls for each routing problem instance as a function of all relevant details.

2.3.2 Marginal-cost tolls: optimal and network-agnostic

In traffic routing, an agent's total cost can be viewed as being two-fold: the first component is the agent's own experienced delay, the second is the delay that the agent's presence imposes on others. A marginal-cost toll explicitly charges each agent for his imposition on other agents; in economic language, marginal-cost tolls internalize the agent's negative externalities [42, 43]. The marginal-cost taxation mechanism $T^{\rm mc}$ assigns tolls to each edge e given by

$$T_e^{\rm mc}(f_e) = f_e \cdot \frac{d}{df_e}\ell_e(f_e). \tag{2.12}$$

Note that each edge's toll depends only on that local edge's congestion properties and traffic flow; global information regarding traffic rate and network topology is not used.

It is well-known that for homogeneous user populations, charging marginal-cost tolls enforces exactly-optimal network flows [42, 43], or

$$\mathcal{L}^{\mathrm{nf}}(G, T^{\mathrm{mc}}(G)) = \mathcal{L}^*(G). \tag{2.13}$$

By construction, these tolls are strongly robust to variations of network topology and overall traffic rate.

This can be seen for Pigou's and Braess's networks in Figures 1.1 and 1.2, where marginal-cost tolls prescribe tolling functions of $\tau(f) = f$ for each of the linear-cost edges. For homogeneous users, the linear-cost edges have a resulting effective cost of $2f$, which incentivizes the desired optimal flows. Mapping these tolls to the robustness meta-problem (1.2), we see that the marginal-cost approach can handle fairly fairly restrictive choices of the uncertainty mask $\mathcal{I}$: if $\mathcal{I}$ hides the overall traffic rate or the network topology, marginal-cost prices still function optimally. One of the central themes of this dissertation is the question of whether there exist other taxation mechanisms which can perform well under these same informational restrictions.

2.3.3 Other Related Work

Recent years have seen a great deal of work on the subject of selfish routing, influencing user behavior in congestion games, and other types of issues surrounding these problems. The authors of [12, 66] study the "price of risk aversion," which measures how society's risk preferences affect aggregate congestion as compared to ordinary Nash flows. Similarly, [67] studies the "deviation ratio," which measures essentially the same quantity for arbitrary cost function biases. The term Stackelberg Routing describes a setting in which a planner controls some fraction of traffic and the remainder of the traffic routes selfishly [68]. In some cases the price of anarchy guaranteed by Stackelberg routing can be shown to be a lower bound for the price of anarchy of a taxation mechanism [36]. Work on the price of anarchy resulting from bounded tolls can be found in [69, 70]. In [71], the authors propose a centralized iterative algorithm whereby a system planner levies tolls on the network, records the resulting Nash flow, updates the tolls accordingly, and so on – and it is proved that Nash flows resulting from this sequence of tolls converge to the planner's target flow even if the initial latency functions are unknown. The authors of [72] study when networks admit demand-independent tolls. The impact of including autonomous vehicles in transportation networks is investigated in [73].

Chapter 3

The Fragility of Fixed Tolls

As shown in [40, 41], fixed tolls can enforce any feasible flow if chosen properly. How robust are they to variations in a system's underlying parameters? Here, we investigate the robustness of fixed tolls to variations of three parameters: latency functions, overall traffic rate, and general network changes. All of this analysis of fixed tolls will be in the simplified context of homogeneous unit-sensitivity populations; thus, any negative results here can only be worsened by extending to the broader heterogeneous case. This chapter deals simply with the questions of strong and weak robustness; for a further look at the price of anarchy associated with fixed tolls, see Theorem 5.4 in Chapter 5.

3.1 Fixed Tolls Cannot Be Strongly Robust

To begin, we ask if fixed tolls can ever be strongly robust to changes in latency functions. One could imagine sudden changes to latency functions arising as a result of traffic accidents, weather, or natural disasters; strong robustness to these would imply that optimal performance would be incentivized regardless of the severity of the disturbance. We have the following easy fact:

Proposition 3.1 *Fixed tolls are not strongly robust on $\mathcal{G}$ to changes in latency functions.*

Put differently, to guarantee optimal routing in general, fixed tolls require detailed characterizations of a network's latency functions.

Proof: Here, the network topology (V, E), path set $\mathcal{P}$, and total traffic rate r are held constant while latency functions $\{\ell\}$ are varied. Note from (2.5) that to disprove strong robustness on general networks $\mathcal{G}$, it suffices to exhibit two networks G and G^* that differ only in their latency functions for which $\mathcal{L}^{\rm nf}(G, T^{\rm ft}(G)) = \mathcal{L}^*(G)$ and $\mathcal{L}^{\rm nf}(G^*, T^{\rm ft}(G)) > \mathcal{L}^*(G^*)$. To achieve this, consider a perturbed version of Pigou's network. In Pigou's network, a fixed toll of $1/2$ on the upper congestible link incentivized optimal routing. This fixed toll was computed for a network whose lower link latency function was given precisely by $\ell_2(f_2) = 1$; what if, instead of 1, the lower latency function was some unknown constant b? To be precise, let G_b represent Pigou's network with $\ell_2(f_2) = b$ so that G_1 represents the nominal Pigou network. To ascertain whether fixed tolls can be strongly robust on Pigou's network, consider the quantity $\mathcal{L}^{\rm nf}(G_b, T^{\rm ft}(G_1))$ which represents the total latency on perturbed network G_b resulting from tolls computed for nominal G_1. If fixed tolls $T^{\rm ft}$ are strongly robust, then for each b, it will be true that $\mathcal{L}^{\rm nf}(G_b, T^{\rm ft}(G_1)) = \mathcal{L}^*(G_b)$. Thus, the robustness of $T^{\rm ft}$ can be checked by varying b in the following price-of-anarchy-like expression:

$$\mathrm{PoA}(b) \triangleq \frac{\mathcal{L}^{\rm nf}\left(G_b, T^{\rm ft}(G_1)\right)}{\mathcal{L}^*(G_b)}. \tag{3.1}$$

In Figure 3.1, $\mathrm{PoA}(b)$ is plotted as b varies between 0 and 1.5 for two sets of tolls: the solid curve corresponds to tolls computed for $b = 1$, and the dashed curve corresponds to tolls equal to 0. Note that the fixed toll only incentivizes perfectly efficient behavior exactly at $b = 1$ (that is, $\mathrm{PoA}(1) = 1$). For all other values, $\mathrm{PoA}(b) > 1$, which means that tolls lead to worse-than-optimal total latencies for $b \neq 1$, or $\mathcal{L}^{\rm nf}(G_b, T^{\rm ft}(G_1)) > \mathcal{L}^*(G_b)$, showing that these tolls are not strongly robust. Note that here we only checked the single fixed toll $\tau_1 = 1/2$, but it is easy to show that on this network, this toll is equivalent in all respects

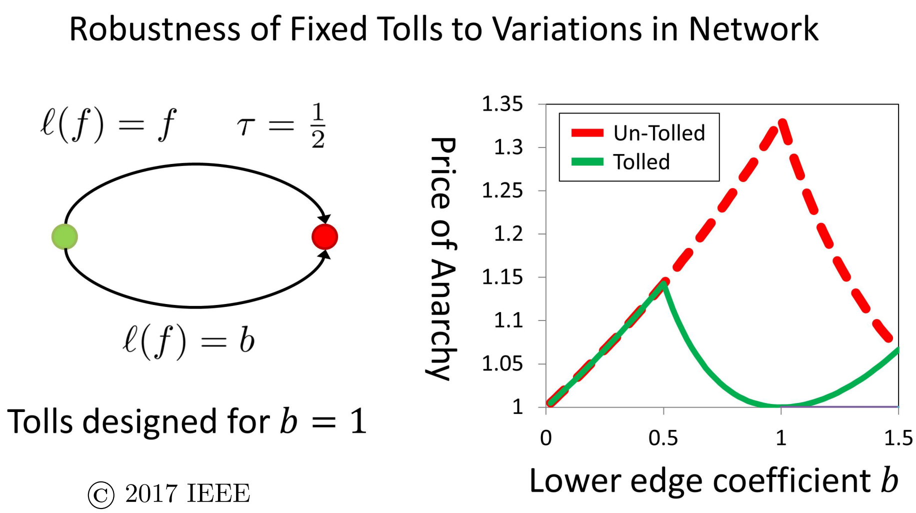

Figure 3.1: Pigou's Network: fixed tolls applied to the "wrong" network. Depicted here is an analysis of the robustness of fixed tolls to variations in latency functions. On the left, a Pigou-style network has a fixed toll of 1/2 charged to the upper link; this toll incentivizes optimal behavior when the lower latency function satisfies $\ell(f) = 1$. To investigate the robustness of fixed tolls to network variations, the toll is held constant while the lower-link latency function b is allowed to vary between 0 and 1.5. For each value of b, the total latency of tolled and un-tolled Nash flows as well as the optimal total latency for that particular value of b are recorded. Finally, the price of anarchy curves are generated by dividing the Nash latencies by the respective optimal latency for each b. Note that the tolled price of anarchy is only 1 when $b = 1$; that is, this fixed toll only incentivizes optimal behavior on the specific network for which it was designed. The fact that the toll does not incentivize optimal behavior for all networks proves Proposition 3.1, stating that fixed tolls are not strongly robust to latency function variations.

to any set of tolls for which $\tau_1 - \tau_2 = 1/2$, so assuming that $\tau_2 = 0$ is without loss of generality. Put differently, the optimal fixed tolls for G_1 are essentially unique, so we have shown that there can exist no fixed-toll taxation mechanism on G (and by extension on $\mathcal{G}$) that is robust to variations of latency functions. ■

Fixed Tolls and Rate-Dependence

We next investigate the robustness of fixed tolls along a different dimension: overall traffic rate. To this end, we return to the Braess's Paradox network of Figure 1.2. Now, suppose that r, the total amount of traffic on the network, is not fixed at 1, but can take any value between 0 and 1. Let G_r represent the Braess's Paradox network with r units of traffic, so the canonical version is simply given by G_1. Let $T^{\mathrm{Braess}}(G_1)$ represent a single fixed toll of 1 on the center zero-latency link.

As for Pigou, the robustness of T^{Braess} can be checked by varying r in the following price-of-anarchy-like expression:

$$\mathrm{PoA}(r) \triangleq \frac{\mathcal{L}^{\mathrm{nf}}\left(G_r, T^{\mathrm{Braess}}(G_1)\right)}{\mathcal{L}^*(G_r)}. \tag{3.2}$$

It has already been shown that for $r = 1$, the optimal flow has no traffic on the center zero-cost link, so the proposed fixed toll achieves the goal of enforcing optimal flows. On the other hand, if $r \leq 0.5$, the optimal flow is to send *all* the traffic on the center zero-cost link, so that no traffic uses the constant-latency links. Unfortunately, the fixed toll on the center link is still boldly incentivizing all users to avoid the now-optimal center link. In Figure 3.2, the price of anarchy from (3.2) is plotted as a function of r with and without the fixed toll on the center link. Note that the tolled curve in Figure 3.2 increases rapidly as r approaches 0, driving the price of anarchy above the un-tolled maximum of 4/3; by setting r low enough, the price of anarchy in this instance can be made arbitrarily high.

This example demonstrates that great care must be taken with fixed tolls when the total traffic rate is varying or unknown, because fixed tolls designed for one demand profile can cause arbitrarily poor performance under a different demand profile.

3.2 Weakly Robust Fixed Tolls Must Depend on Network Structure

A fundamental problem with the fixed tolls applied to an uncertain Pigou network (as in Figure 3.1) was that the correct toll on the upper edge depended on the latency function of the lower edge; if the lower latency function was unknown, there was no way to compute an optimal toll on the upper edge, so fixed tolls could not be strongly robust. This prompts the question: could weakly robust fixed tolls be designed by letting the tolling function

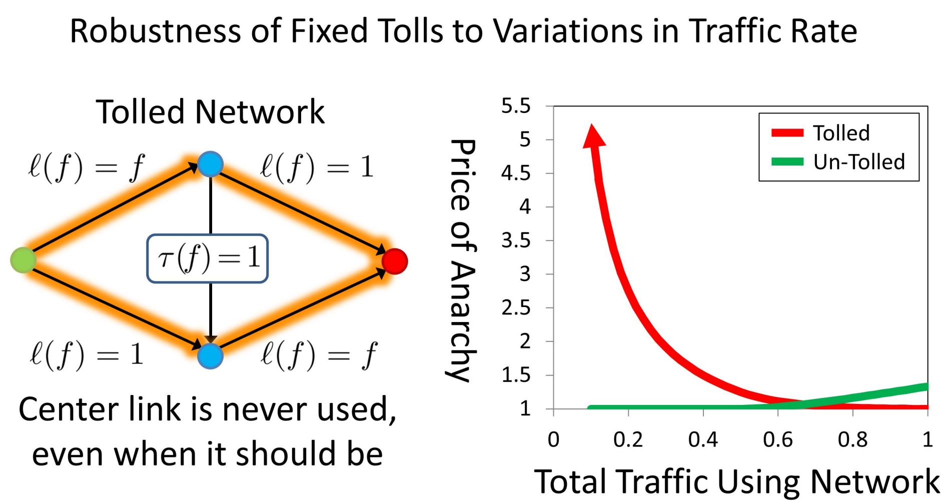

Figure 3.2: Braess' network: perverse incentives piled on top of unintended consequences. Following on the classical Braess's paradox (see Figure 1.2), this figure depicts an attempt to redeem the pathological network augmentation with a simple fixed toll. If a toll of 1 is levied on the link connecting the two intermediate nodes, it is simple to show that no traffic will ever use that link. Unfortunately, for low traffic rates (particularly when the total mass of traffic is less than 0.5), it is optimal for traffic to use the center link, but the fixed toll prevents this. On the right is plotted the price of anarchy as a function of the total traffic rate; note that under the influence of tolls, the price of anarchy can become arbitrarily large as traffic approaches 0.

for edge e depend only on the local latency function ℓ_e? Such a taxation mechanism is called *network-agnostic*, defined more formally in (2.9). If a taxation mechanism is network-agnostic, then each edge toll depends only on local information, so any efficiency guarantees are automatically robust to changes in network structure.

One way to view network-agnosticity is that the toll-designer pre-commits to a tolling function for each possible latency function without specific knowledge of which latency functions will appear in the final realization. Thus, the weak-robustness notation of (2.6) can be simplified by writing $\mathcal{L}^{\mathrm{nf}}(G, T)$ to mean the total latency of a Nash flow on network G resulting from the tolls generated by taxation mechanism T, and $\mathcal{L}^{\mathrm{nf}}(G, \emptyset)$ to mean the total latency of a Nash flow on G with no tolls. This leads to the main result about the lack of robustness of fixed tolls:

Theorem 3.2 *The only nonnegative network-agnostic fixed tolls that are weakly robust to network variations satisfy*

$$\tau_e = 0 \tag{3.3}$$

for all possible network edges.

Theorem 3.2 shows that positive fixed tolls must in general require some global information about network structure (e.g., network topology or latency functions) in order to ensure that they do not cause harm, even for the simple setting of homogeneous populations. Theorem 3.2 is proved with a series of simple example networks. This shows that even on simple networks, fixed tolls lack robustness, suggesting that complex networks could exhibit even more severe pathologies.

Proof: Let T^{naft} be a network-agnostic taxation mechanism such that for any network G, $\mathcal{L}^{\mathrm{nf}}(G, T^{\mathrm{naft}}) \leq \mathcal{L}^{\mathrm{nf}}(G, \emptyset)$. Write the toll assigned to an edge with latency function ℓ as $T^{\mathrm{naft}}(\ell)$.

First, consider the network shown in Figure 3.3(a) in which the latency functions satisfy $\ell_1 + \ell_2 = \ell_3$. The flow $(1/2, 1/2)$ is both a Nash flow and an optimal flow; the only tolls

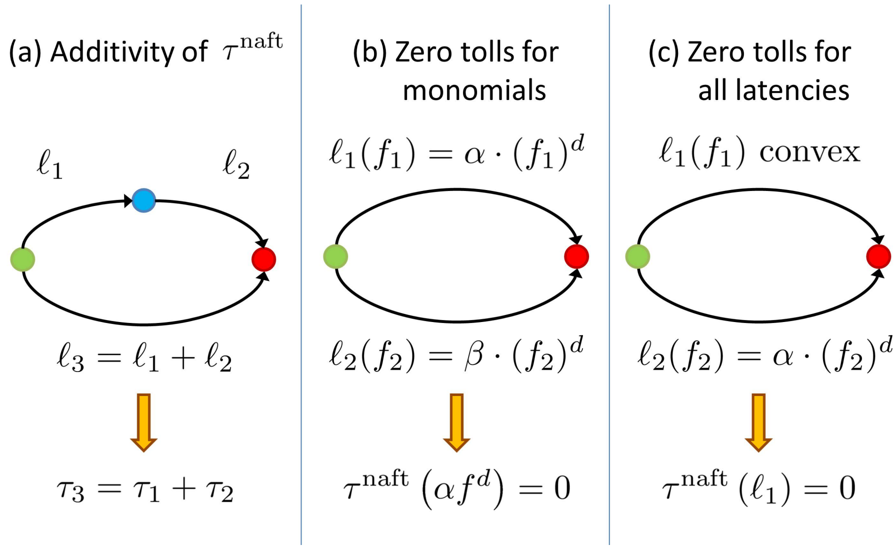

Figure 3.3: Networks used to prove Theorem 3.2, showing that any network-agnostic fixed-toll taxation mechanism must charge taxes of 0 on every link. The network in (a) is used to show that network-agnostic fixed-toll taxation mechanism T^{naft} is additive in latency functions. Next, (b) is used to show in conjunction with (a) that monomial cost functions of any degree must be assigned a zero toll. Finally, in (c), for any general latency function ℓ_1, it is shown how to create a polynomial latency function ℓ_2 which is more congestible at equilibrium than ℓ_1, so that a positive toll charged on ℓ_1 would increase the total latency of the network.

which will always support this must satisfy $T^{\text{naft}}(\ell_1) + T^{\text{naft}}(\ell_2) = T^{\text{naft}}(\ell_3)$. This is the first condition on T^{naft}:

$$\ell_1 + \ell_2 = \ell_3 \quad \Longrightarrow \quad T^{\text{naft}}(\ell_1) + T^{\text{naft}}(\ell_2) = T^{\text{naft}}(\ell_3). \tag{3.4}$$

Next, consider the network shown in Figure 3.3(b), a two-link parallel network with degree-d monomial cost functions $\ell_1(f_1) = \alpha(f_1)^d$ and $\ell_2(f_2) = \beta(f_2)^d$. It can be shown that the optimal flow on this network is equal to the untolled Nash flow for any $\alpha > 0, \beta > 0$, and $d \geq 1$. Thus, the tolls on each link must be equal; otherwise, the tolled flow will have a strictly higher total latency than the un-tolled flow. That is, all monomials of the same degree must be charged the same toll, regardless of the scale of the latency function: $T^{\text{naft}}\left(\alpha f^d\right) = T^{\text{naft}}\left(\beta f^d\right)$. In particular, letting $\beta = 2\alpha$ and appealing to (3.4), it holds that that $2T^{\text{naft}}\left(\alpha f^d\right) = T^{\text{naft}}\left(2\alpha f^d\right)$, which implies the second condition on T^{naft}:

$$\text{for all } \alpha > 0 \text{ and } d \geq 1, \quad T^{\text{naft}}\left(\alpha f^d\right) = 0. \tag{3.5}$$

Using this fact regarding polynomials, positive fixed tolls on any other latency function can now be ruled out. Refer to Figure 3.3(c), another two-link parallel network. Given an arbitrary convex latency function ℓ (with flow derivative ℓ') on the upper link, it is possible to design a polynomial latency function for the lower link $\ell_2(f_2) = \alpha(f_2)^d$ to show that $T^{\text{naft}}(\ell) = 0$. Let the polynomial degree satisfy $d > \frac{\ell(1/2)}{2\ell'(1/2)}$, and coefficient satisfy $\alpha = 2^d\ell(1/2)$. Then by design, the un-tolled Nash flow on the network is $(1/2, 1/2)$, and at this flow, shifting any positive mass of traffic from the upper link to the lower link strictly increases the total latency on the network. To avoid this, the toll on the upper link must be zero: $T^{\text{naft}}(\ell) = 0$. Since ℓ is an arbitrary convex latency function, the theorem is proved.

■

Paying for Optimality: A Note on Network-Agnostic Fixed Subsidies

Theorem 3.2 carefully specified that the fixed tolls in question be nonnegative – and this prompts the question of negative tolls, i.e., subsidies. It turns out that for the special case of linear-latency networks, strongly robust network-agnostic fixed *subsidies* do exist. For a linear latency function of the form $\ell_e(f_e) = a_e f_e + b_e$, the corresponding network-agnostic fixed subsidy (represented as a negative toll) is given by

$$\tau_e^{\text{subsidy}} = -\frac{b_e}{2}. \tag{3.6}$$

By simply paying users half the constant-term cost on each link, optimal flows can be incentivized as Nash flows. As a side note, these subsidies are a special case of the "variable price schemes" of [43] with $\bar{\eta} = 1$.

The cost functions resulting from these subsidies can be related to the cost functions resulting from standard marginal-cost tolls τ^{mc} (see (2.12)), given in this case by $\tau_e^{\mathrm{mc}} = a_e f_e$. Under the homogeneous-user model, the cost functions resulting from τ^{subsidy} are

$$J_e^{\mathrm{subsidy}}(f_e) = \underbrace{a_e f_e + b_e}_{\ell_e} \underbrace{- \frac{b_e}{2}}_{\tau_e} = a_e f_e + \frac{b_e}{2}, \tag{3.7}$$

while the cost functions resulting from marginal-cost tolls are

$$J_e^{\mathrm{mc}}(f_e) = 2a_e f_e + b_e. \tag{3.8}$$

Since these two cost functions are related by a constant multiplicative factor for all agents (i.e., $J^{\mathrm{mc}} = 2J^{\mathrm{subsidy}}$), they induce the same optimal Nash flows, and these subsidies inherit the strong robustness of marginal-cost tolls. While these strongly robust network-agnostic fixed subsidies are appealing, it is not clear that the concept generalizes beyond linear cost functions or homogeneous users.

Chapter 4

Can Marginal-Cost Tolls Help?

This chapter represents a first step towards characterizing the robustness of the popular taxation mechanism known as "marginal-cost" tolls, as defined in (2.12) [42, 43]. For convenience, we repeat the definition here. The marginal-cost taxation mechanism T^{mc} assigns tolls to each edge e (with associated latency function $\ell_e(f_e)$) of

$$T_e^{\mathrm{mc}}(f_e) = f_e \cdot \frac{d}{df_e}\ell_e(f_e). \tag{4.1}$$

It has long been known that marginal-cost tolls incentivize congestion-minimal flows on all networks for homogeneous users [42, 43], but heretofore the robustness of these prices to mischaracterization of underlying parameters has been left unstudied.

4.1 Marginal-cost tolls are not strongly robust for price sensitive users

The strong robustness guarantees of marginal-cost tolls have been proved by [42, 43] in the setting of homogeneous known-sensitivity users; do these guarantees carry over to

---

Section 4.1 contains material that is adapted, with permission, from [1] and was previously published by IEEE Control Systems. Section 4.2 is adapted, with permission, from [74] and was previously published by IEEE Transactions on Automatic Control.

the more detailed heterogeneous model? As a first step towards studying robust taxation mechanisms for heterogeneous users, the performance of "off-the-shelf" marginal-cost tolls is investigated on simple price-sensitive settings.

In Pigou's network (see Figures 1.1 and 4.1), the marginal-cost taxation mechanism T^{mc} assigns a flow-varying toll of $\tau_1^{\mathrm{mc}}(f_1) = f_1$ to the upper link; for homogeneous users, this incentivizes optimal routing. What if the users' sensitivities remained homogeneous, but took on some unknown value other than 1? We can employ a parallel argument to that seen for fixed tolls in Proposition 3.1: Write G_s to denote Pigou's network in which all users have sensitivity s. To ascertain whether marginal-cost tolls can be strongly robust to sensitivity variations, consider the quantity $\mathcal{L}^{\mathrm{nf}}(G_s, T^{\mathrm{mc}}(G_1))$, which represents the total latency on the perturbed population in G_s resulting from marginal-cost tolls computed for unit-sensitivity G_1. If marginal-cost tolls $\tau^{\mathrm{mc}}(G_1)$ are strongly robust, then for each s, it will be true that $\mathcal{L}^{\mathrm{nf}}(G_s, T^{\mathrm{mc}}(G_1)) = \mathcal{L}^*(G_s)$. Since the optimal flow on a network does not depend on the user sensitivities, this is simply equivalent to writing $\mathcal{L}^{\mathrm{nf}}(G_s, T^{\mathrm{mc}}(G_1)) = 0.75$.

In Figure 4.1, $\mathcal{L}^{\mathrm{nf}}(G_s, T^{\mathrm{mc}}(G_1))$ is plotted as s varies between 0 and 2. Note that the marginal-cost toll only incentivizes perfectly efficient behavior exactly at $s = 1$. For all other values, the total latency is strictly greater than the optimal 0.75, showing that these tolls are not strongly robust. This implies that when considering user price-sensitivity, marginal-cost tolls are not strongly robust to sensitivity variations, even in the simplified setting of homogeneous users on Pigou's network.

Perverse Marginal-Cost Tolls For Heterogeneous Users

Here, we ask a similar question of marginal-cost tolls in the heterogeneous model to that asked for fixed tolls in the homogeneous model: Despite lacking strong robustness to user sensitivities, is it at least possible to show that marginal-cost tolls are weakly robust? Unfortunately, in general settings, it is possible to show that even marginal-cost tolls can incentivize flows that are strictly worse than their un-tolled counterparts, thus lacking even

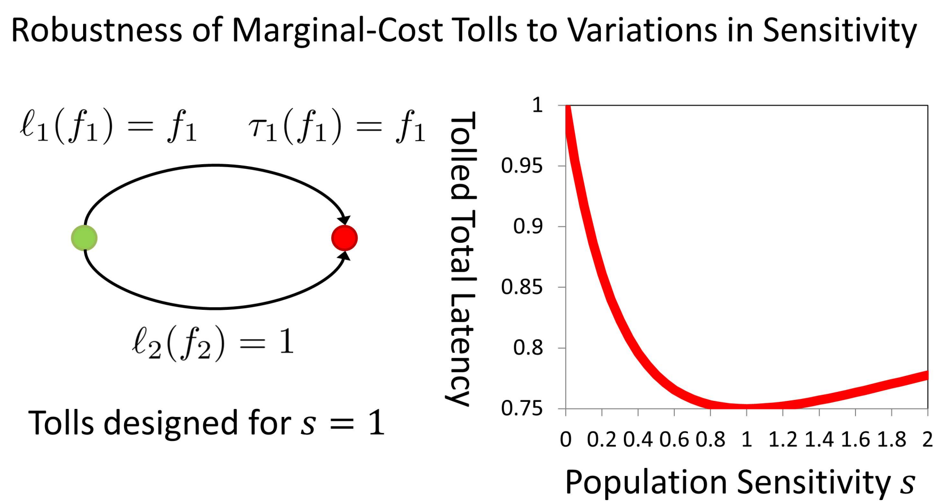

Figure 4.1: A setting in which marginal-cost tolls are shown not to be strongly robust to variations of user sensitivity. Depicted is the canonical Pigou Network with marginal-cost toll $\tau^{\mathrm{mc}}(f) = f$ assigned to the top link. If marginal-cost tolls were strongly robust to variations of user sensitivities, they would incentivize optimal flows for every user sensitivity profile. To check this, tolls are chosen without *a priori* knowledge of the user sensitivities, and then the population's homogeneous sensitivity is swept from 0 to 2 and the resulting total latency is plotted. The optimal total latency is only obtained when the user sensitivities are exactly 1; all other sensitivity values incentivize some suboptimal total latency $\mathcal{L}(f) > 0.75$.

weak robustness to heterogeneity. Consider the network and demand profile in Figure 4.2. This is essentially a three-link network; a population of mass 0.5 has access only to the upper two links, and a population of mass 1 has access only to the lower two links. The optimal flow has all users from the upper population using the center (congestible) link and all users from the lower population using the lower link. In the unique un-tolled Nash flow, half the traffic from the lower link has shifted to the center link so as to equalize the latencies of those two links.

Note that the un-tolled version bears a strong resemblance to Pigou's example if the upper link is ignored: self-interested users from the lower source have over-congested the center link and degraded its performance. To see how tolls can make things even worse, suppose that the upper population is toll-sensitive (with sensitivities equal to 1), but the lower population is not (i.e., they have sensitivities close to 0). The insensitivity of the lower population effectively fixes the flow on the center link at 1, since this is the flow that equalizes the latencies of the lower two links. That is, regardless of what the upper population does, the center link flow will always be 1. Thus, marginal-cost tolls on this network serve only to force the upper population to choose the upper link, resulting in the flow depicted on the right in Figure 4.2. This pathological flow has a total latency of 1.75, which corresponds to a price of anarchy of 1.4. This is greater than the 4/3 guaranteed as a worst-case on linear-latency networks, demonstrating that marginal-cost tolls lack even weak robustness for heterogeneous populations. In this case, instead of incentivizing altruism, marginal-cost tolls simply amplified the selfishness-induced inefficiency that was already present. This allows us to state the following fact:

Proposition 4.1 *Let $\mathcal{G}$ denote the class of all routing problems. Marginal-cost tolls* (2.12) *are not weakly-robust on $\mathcal{G}$ to variations of user sensitivities.*

Proof: Let G denote the routing game in Figure 4.2. Per the above analysis, $T^{\rm mc}$ induces a Nash total latency of $\mathcal{L}^{\rm nf}(G, T^{\rm mc}(G)) = 1.75$, but the un-tolled total latency is $\mathcal{L}^{\rm nf}(G, \emptyset) = 1.5$; since the tolled total latency is strictly worse, $T^{\rm mc}$ cannot be weakly robust

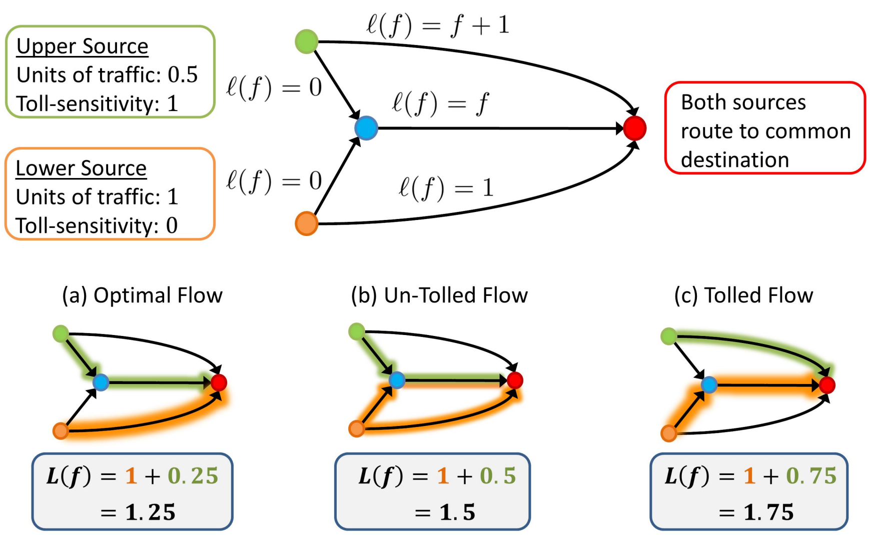

Figure 4.2: A network demonstrating that marginal-cost tolls are not weakly robust to user heterogeneity. This figure depicts a simple two-source network in which 0.5 units of traffic route from the upper (green) source, and 1 unit of traffic routes from the lower (orange) source. If traffic from the upper source trades off time and money equally (i.e., $s \equiv 1$), but traffic from the lower source cares only about time (i.e., $s \equiv 0$), marginal-cost tolls result in a price of anarchy of 1.4 on this network. The optimal flow here requires all of the traffic from the lower source to use the lower, constant-latency link. However, only the traffic from the upper source responds to tolls; when marginal-cost tolls are levied, all of the upper-source (green traffic) moves to the inefficient upper path, and the lower-source (orange) traffic moves to replace it on the middle path, as depicted on the right.

to heterogeneous user sensitivities. ■

As a side note, the most extreme pathology in this example arises when the lower population has a sensitivity of 0, but perversities still arise for any low, positive sensitivity. That is, the poor performance in this example can still occur when every agent has a nonzero price-sensitivity.

4.2 Scaled marginal-cost tolls offer some robustness

Though Proposition 4.1 showed that off-the-shelf marginal-cost tolls are not weakly-robust on the class of all networks, this left open the possibility that something *like* marginal-cost tolls may yet exhibit some robustness, perhaps on a reduced set of networks.

In this section, we study the efficacy of the *scaled marginal-cost* taxation mechanism for parallel networks with linear-affine cost functions and under a particular utilization condition in situations in which both the number of links and the users' price-sensitivities are unknown or time-varying.

In the following, we write $\mathcal{G}$ to denote the class of symmetric parallel networks with linear-affine latency functions of the form $\ell_e(f_e) = a_e f_e + b_e$. Furthermore, we assume that the traffic rate on each network in $\mathcal{G}$ is such that all edges have positive flow in an un-tolled Nash flow.[1]

We study tolls of the following form: for any scalar coefficient $\kappa \geq 0$, the scaled marginal-cost taxation mechanism, denoted by $T^{\mathrm{smc}}(\kappa)$, assigns taxation functions

$$\tau_e^{\mathrm{smc}}(f_e;\kappa) = \kappa \cdot f_e \cdot \frac{d}{df_e}\ell_e(f_e) = \kappa a_e f_e,\ \forall f_e \geq 0. \tag{4.2}$$

[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.

To formalize a notion of worst-case efficiency guarantees, we define the set of possible sensitivity distributions for the users as $\mathcal{S} = \{s : [0,1] \to [S_{\rm L}, S_{\rm U}]\}$. Let $\mathcal{L}^*(G)$ denote the total latency associated with the optimal flow, and $\mathcal{L}^{\rm nf}(G, s, \tau)$ denote the total latency associated with the Nash flow resulting from taxation functions τ and sensitivity distribution $s \in \mathcal{S}$. In this section, we formulate the price of anarchy of the scaled marginal-cost taxation mechanism with respect to both uncertainity in the underlying network and the users' price-sensitivity, i.e.,

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}, T^{\rm smc}(\kappa)) = \sup_{s \in \mathcal{S}, G \in \mathcal{G}} \left\{ \frac{\mathcal{L}^{\rm nf}(G, s, T^{\rm smc}(\kappa))}{\mathcal{L}^*(G)} \right\} \geq 1. \tag{4.3}$$

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 4.2 *For any affine-cost parallel network $G \in \mathcal{G}$ with flow on all edges in an un-tolled Nash flow, 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 $\kappa > 0$*

$$\mathcal{L}^{\rm nf}\left(G, s, T^{\rm smc}(\kappa)\right) < \mathcal{L}^{\rm nf}\left(G, s, \emptyset\right). \tag{4.4}$$ [2]

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

$$\kappa^* = \frac{1}{\sqrt{S_{\rm L} S_{\rm U}}} = \arg\min_{\kappa \geq 0} \left\{ \mathrm{PoA}(\mathcal{G}, \mathcal{S}, T^{\rm smc}(\kappa)) \right\}. \tag{4.5}$$

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

$$\mathrm{PoA}\left(\mathcal{G}, \mathcal{S}, T^{\rm smc}(\kappa^*)\right) = \frac{4}{3}\left(1 - \frac{\sqrt{S_L/S_U}}{\left(1 + \sqrt{S_L/S_U}\right)^2}\right) \leq \frac{4}{3}. \tag{4.6}$$

[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 section assume that $\mathcal{L}^{\rm nf}(G, s, \emptyset) > \mathcal{L}^*(G)$.

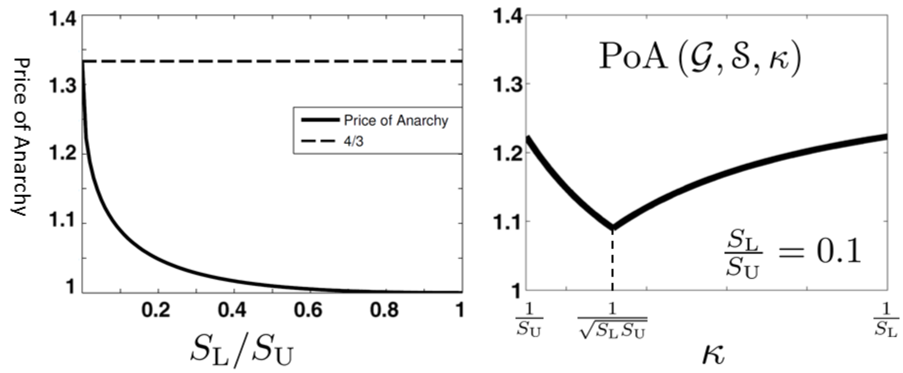

Figure 4.3: Left: An illustration of the price of anarchy bound from Theorem 4.2, with optimal toll scalar $\kappa = (S_{\rm L}S_{\rm U})^{-1/2}$. Since the bound depends only on $S_{\rm L}/S_{\rm U}$, this plot neatly expresses the effect of model uncertainty on toll effectiveness. As expected, we inherit the canonical price of anarchy of $4/3$ when $S_{\rm L}/S_{\rm U} = 0$ (i.e., we may be entirely unable to influence behavior). 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 $S_{\rm U}$.

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. [39], that can guarantee higher efficiencies only at the expense of strict informational requirements. See Figure 4.3 for plots of the price of anarchy with respect to various parameters.

Theorem 4.2 Proof

We begin with some notation before delving into the proof of Theorem 4.2. 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., $\left|\cup_{x\in[0,1]} s_x\right| \leq m$.

Let $\mathcal{F}(G, \mathcal{S}, T) \subset \mathbb{R}^n$ denote the set of Nash flows associated with all routing games

[3]This price of anarchy bound is also unchanged by increases in the total mass of traffic flowing through the network; see Claim 4.2.1.1.

(G, s, T) where $s \in \mathcal{S}$. Note that we are representing Nash flows anonymously: a particular $f^{\mathrm{nf}} \in \mathcal{F}(G, \mathcal{S}, T)$ describes merely *how many* agents are on each edge, not *which* agents are on each edge. For brevity, we often express $T^{\mathrm{smc}}(\kappa)$ as merely κ.

The proof of Theorem 4.2 involves proving that the scaling coefficient $\kappa \geq 0$ 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 Lemmas 4.2.1 and 4.2.3 can be found in Section 4.3.

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 T as $\mathcal{F}^{\mathrm{mi}}(\mathcal{G}, \mathcal{S}_m, T)$. Note that on a network with n links, there are at most $(n-1)$ sensitivity types in a minimally-indifferent Nash flow.

First, Lemma 4.2.1 proves that a Nash flow on an n-link network for any heterogeneous population can be represented as a minimally-indifferent Nash flow for a population with only $(n-1)$ sensitivities. Thus, we can assume without loss of generality that any Nash flow is minimally-indifferent.

Lemma 4.2.1 *For any network $G \in \mathcal{G}$ consisting of n links, with $n \geq 2$, and $\kappa \geq 0$,*

$$\mathcal{F}(G, \mathcal{S}, \kappa) = \mathcal{F}^{\mathrm{mi}}\left(G, \mathcal{S}_{n-1}, \kappa\right). \tag{4.7}$$

Second, Lemma 4.2.2 shows that we may further refine our search to the set of homogeneous sensitivity distributions. In particular, when $\kappa \leq \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}}$, the worst-case total latency is realized by Nash flows for a homogeneous population with sensitivity S_{L}.

Lemma 4.2.2 *Let* $\kappa \leq \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}}$. *Then for any* $G \in \mathcal{G}$,

$$\max_{s \in \mathcal{S}} \mathcal{L}^{\mathrm{nf}}\left(G, s, \kappa\right) = \mathcal{L}^{\mathrm{nf}}\left(G, S_{\mathrm{L}}, \kappa\right). \tag{4.8}$$

Proof: 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 $G \in \mathcal{G}$, any minimally-indifferent Nash flow $f^{\mathrm{nf}} \in \mathcal{F}^{\mathrm{mi}}\left(G, \mathcal{S}_{n-1}, \kappa\right)$ with sensitivity values $\{s_i\}_{i=1}^{n-1}$ satisfies

$$a_i f_i^{\mathrm{nf}} - a_{i+1} f_{i+1}^{\mathrm{nf}} = \frac{b_{i+1} - b_i}{1 + \kappa s_i} \tag{4.9}$$

for each pair of adjacent edges (for details, see (4.32) in the proof of Lemma 4.2.1 in Section 4.3). Note that the expression in (4.9) 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 (4.9) as a linear expression in both f^{nf} and z.

The $(n-1)$ equations obtained from (4.9) combined with the flow-conservation constraint $\sum_{i=1}^{n} f_i^{\mathrm{nf}} = 1$, yield the n-dimensional linear system

$$P f^{\mathrm{nf}} = r + Qz \tag{4.10}$$

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

$$f^{\mathrm{nf}}(z) = R + Mz, \tag{4.11}$$

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

$$R = P^{-1}r, \qquad\qquad M = P^{-1}Q. \tag{4.12}$$

The following observations will be helpful to our proof:

Observation 4.2.2.1 *The matrices M and R possess the following properties for any $G \in \mathcal{G}$:*

$$\mathbf{1}^{\mathrm{T}} M = \mathbf{0}^{\mathrm{T}}, \tag{4.13}$$

$$\mathbf{1}^{\mathrm{T}} R = 1, \tag{4.14}$$

$$AR \in \mathrm{sp}\left\{\mathbf{1}\right\}, \tag{4.15}$$

$$M^T A M \mathbf{1} = -M^T b. \tag{4.16}$$

Observation 4.2.2.2 *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:*

$$\mathcal{L}^{\mathrm{nf}}(z) = z^T M^T A M z + z^T M^T b + L_R, \tag{4.17}$$

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:

$$\mathcal{L}^{\mathrm{nf}}(G, s, 0) = L_R. \tag{4.18}$$

Proof: [Proof of Observation 4.2.2.1]These facts follow algebraically from the fact that by definition, for any $z \in \mathbb{R}^{n-1}$, $f^{\mathrm{nf}}(z)$ satisfies (4.10). ■

Proof: [Proof of Observation 4.2.2.2]We simply substitute $f^{\mathrm{nf}}(z)$ (that is, equation (4.11)) into (4.27) to obtain

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

Consider the last term, $2R^T AMz$. By (4.15) in Observation 4.2.2.1, $\exists \alpha \in \mathbb{R}$ such that $R^T A = \alpha \mathbf{1}^{\mathrm{T}}$, and by (4.13), $\mathbf{1}^{\mathrm{T}} M = \mathbf{0}^T$, so $2R^T AMz = 0$. Simplifying, we obtain

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

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 (4.16) in Observation 4.2.2.1, 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 (4.11) to parameterize the set of all Nash flows for any network.

Characterizing the set of Nash flows

To formalize our definition of $f^{\mathrm{nf}}(z)$ (given in (4.11)), 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:

$$\frac{1}{1+\kappa S_{\mathrm{L}}} \geq z_1 \geq \cdots \geq z_i \geq z_{i+1} \geq \cdots \geq z_{n-1} \geq \frac{1}{1+\kappa S_{\mathrm{U}}}. \tag{4.19}$$

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 (4.10) 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 4.2.2, 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}.

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}), $Mz_v = z_{\mathrm{L}}\mu_{\mathrm{L}} + z_{\mathrm{U}}\mu_{\mathrm{U}}$. By substituting the expression for a Nash flow (4.11) into the incentive constraints (4.9), it can be shown via Observation 4.2.2.1 that the first $(\ell-1)$ elements of μ_{U} are nonnegative, but elements ℓ through $(n-1)$ of μ_{U} are nonpositive. This corresponds to the fact that increases in κ always shift traffic to higher-index links. Furthermore, this operation implies that the vector $(A\mu_{\mathrm{U}} + b)$ is nonnegative and ordered nondecreasing. Equation (4.13) implies that $\mu_{\mathrm{U}}^T \mathbf{1} = 0$, so it follows that

$$\mu_{\mathrm{U}}^T (A\mu_{\mathrm{U}} + b) \leq 0 \tag{4.20}$$

because $(A\mu_{\mathrm{U}} + b)$ places more weight on the negative elements of μ_{U}.

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

$$(z_{\mathrm{L}} - z_{\mathrm{U}})\, \mu_{\mathrm{U}}^T \left[(z_{\mathrm{L}} + z_{\mathrm{U}} - 1)\, A\mu_{\mathrm{U}} + (1 - 2z_{\mathrm{L}})\, (A\mu_{\mathrm{U}} + b) \right]. \tag{4.21}$$

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 (4.20) shows that the expression in (4.21) 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. ■

Finally, Lemma 4.2.3 gives the unique optimal value of κ for homogeneous populations; heterogeneous populations ultimately inherit this optimal result.

Lemma 4.2.3 *For all $G \in \mathcal{G}$, and for all $\kappa \neq \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}} = \kappa^*$,*

$$\max_{s \in \mathcal{S}_1} \mathcal{L}^{\mathrm{nf}}\left(G, s, \kappa^*\right) < \max_{s \in \mathcal{S}_1} \mathcal{L}^{\mathrm{nf}}\left(G, s, \kappa\right). \tag{4.22}$$

Finally, the price of anarchy of $T^{\mathrm{smc}}(\kappa^*)$ for homogeneous populations is given by (4.6).

Proof: [Proof of Theorem 4.2] We combine the inequalities on the price of anarchy proved in each lemma. Lemma 4.2.1 implies that

$$\mathrm{PoA}\left(\mathcal{G}, \mathcal{S}, \kappa^*\right) = \mathrm{PoA}(\mathcal{G}, \mathcal{S}_{n-1}, \kappa^*). \tag{4.23}$$

Lemma 4.2.2 implies that

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}_{n-1}, \kappa^*) = \mathrm{PoA}(\mathcal{G}, \mathcal{S}_1, \kappa^*) \tag{4.24}$$

and the worst-case total latency with $\kappa = \kappa^*$ is better than the un-tolled total latency. By Lemma 4.2.3, we have that for any $\kappa \neq \kappa^*$,

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}_1, \kappa^*) < \mathrm{PoA}(\mathcal{G}, \mathcal{S}_1, \kappa). \tag{4.25}$$

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

$$\mathrm{PoA}(\mathcal{G}, \mathcal{S}_1, \kappa) \leq \mathrm{PoA}(\mathcal{G}, \mathcal{S}, \kappa). \tag{4.26}$$

Combining inequalities (4.23), (4.24), (4.25), and (4.26), we have that for any $\kappa \neq \kappa^*$,

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

Thus, (4.6) is valid for heterogeneous populations as well. ■

4.3 Chapter Proofs

4.3.1 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

$$\mathcal{L}(f) = f^T A f + f^T b. \tag{4.27}$$

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 $\ell_i(f_i) = a_i f_i + b_i$.

4.3.2 Proof of Lemma 4.2.1 and Associated Claims

We first prove two intermediate claims. In Claim 4.2.1.1 we show that if every link has positive flow in an un-tolled Nash flow, then under $T^{\mathrm{smc}}(\kappa)$, every link in that network will have positive flow in a Nash flow induced by any finite $\kappa > 0$.

Claims 4.2.1.1 and 4.2.1.2 use the following definition: for Nash flow $f^{\mathrm{nf}} \in \mathcal{F}(G, \mathcal{S}, \kappa)$, for each edge $e_i \in E$, define s_i^- and s_i^+ by the following:

$$s_i^- = \inf_{x \in [0,1]} \left\{ s_x : \text{agent } x \text{ uses edge } e_i \text{ in flow } f^{\mathrm{nf}} \right\}, \tag{4.28}$$

$$s_i^+ = \sup_{x \in [0,1]} \left\{ s_x : \text{agent } x \text{ uses edge } e_i \text{ in flow } f^{\mathrm{nf}} \right\}. \tag{4.29}$$

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 4.2.1.1 *For any network $G \in \mathcal{G}$, let $f^{\mathrm{nf}} \in \mathcal{F}(G, \mathcal{S}, \kappa)$ for any $\kappa \geq 0$. Then f^{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 $f^{\mathrm{t}} \in \mathcal{F}(G, \mathcal{S}, \kappa)$ for $\kappa > 0$ in which some edge e_k has

$f_k^{\rm t} = 0$. Thus, for every edge e_i,

$$(1 + s_i^+ \kappa) a_i f_i^{\rm t} + b_i \leq b_k \leq a_i f_i + b_i. \tag{4.30}$$

Simplifying (4.30) 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 4.2.1.2 we show that under scaled marginal-cost tolls, heterogeneous users sort themselves onto the links in a predictable order.

Claim 4.2.1.2 *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 $f^{\rm nf}$: (i) $a_i f_i^{\rm nf} \geq a_j f_j^{\rm nf}$, and (ii) $s_i^+ \leq s_j^-$.*

Proof: Consider edges e_i and e_{i+1} in network G. By hypothesis, $b_i \leq b_{i+1}$. Consider a Nash flow $f^{\rm nf} \in \mathcal{F}(G, s, \kappa)$ with $\kappa \geq 0$ and $s \in \mathcal{S}$. By Claim 4.2.1.1, $f_{i+1}^{\rm nf} > 0$. 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 $\tau_e(f_e) = a_e f_e$,

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

Thus, $a_i f_i^{\rm nf} \geq a_{i+1} f_{i+1}^{\rm nf} \geq 0$, 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:

$$(1 + \kappa s_{i+1}^-) a_{i+1} f_{i+1}^{\rm nf} + b_{i+1} \leq (1 + \kappa s_{i+1}^-) a_i f_i^{\rm nf} + b_i. \tag{4.31}$$

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

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

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 $f^{\text{nf}} \in \mathcal{F}(G, s, \kappa)$; by Claim 4.2.1.2, $s_i^+ \leq s_{i+1}^-$. We may rearrange (4.31) (and the opposite inequality for s_i^+) to obtain

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

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

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

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 $\forall i \in \{2, \ldots, n-2\}$, $(f_i^{\text{nf}} + f_{i+1}^{\text{nf}})/2$ agents have a sensitivity of s_i; $\left(f_1^{\text{nf}} + f_2^{\text{nf}}/2\right)$ agents have sensitivity s_1, and $\left(f_{n-1}^{\text{nf}}/2 + f_n^{\text{nf}}\right)$ agents have sensitivity s_{n-1}. Then $f^{\text{nf}} \in \mathcal{F}^{\text{mi}}(G, \mathcal{S}_{n-1}, \kappa)$; 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:

$$(1 + \kappa s_i) a_i f_i + b_i = (1 + \kappa s_i) a_{i+1} f_{i+1} + b_{i+1}. \tag{4.32}$$

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

[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.

and the desired result is immediate. ■

4.3.3 Proof of Lemma 4.2.3

The proof of Lemma 4.2.3 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 (4.17), if we write $\Theta = -\mathbf{1}^\mathrm{T} b^T M = \mathbf{1}^\mathrm{T} M^T A M \mathbf{1} \geq 0$ (see Observation 4.2.2.1), we may explicitly write the total latency of a homogeneous Nash flow as

$$\begin{aligned} \mathcal{L}^\mathrm{nf}(G, s, \kappa) &= L_R + \frac{\mathbf{1}^\mathrm{T} M^T A M \mathbf{1}}{(1+s\kappa)^2} + \frac{b^T M \mathbf{1}}{1+s\kappa} \\ &= L_R - \frac{s\kappa}{(1+s\kappa)^2}\Theta. \end{aligned} \tag{4.33}$$

It is easy to verify that the minimum of (4.33) occurs whenever $\kappa = 1/s$, and is equal to $L_R - \Theta/4$. Furthermore, partial derivatives of (4.33) 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 (4.33) that the solution to this equation is

$$\kappa^* = \frac{1}{\sqrt{S_\mathrm{L} S_\mathrm{U}}}. \tag{4.34}$$

The partial derivatives of (4.33) 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 (4.34). Since we know that an un-tolled latency can never be more than 4/3 times an optimal latency, from (4.33) we can write

$$\frac{\mathcal{L}^{\mathrm{nf}}(G,s,0)}{\mathcal{L}^*(G)} = \frac{L_R}{L_R - \frac{1}{4}\Theta} \leq \frac{4}{3}. \tag{4.35}$$

This implies that $\Theta \leq L_R$, and it follows algebraically that for κ^* as defined in (4.34), $s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]$, and $\mathcal{G}$, the expression for the price of anarchy is given by (4.6). ■

Remarks

While this essentially concludes our study on marginal-cost tolls *per sé*, it should be noted that the principles applied in this chapter will resurface again several times in coming chapters. As our study proceeds, we will show that several taxation mechanisms of interest can be analyzed as simple variations of scaled marginal-cost tolls. As such, several of the analytical tools in this chapter will find uses for other purposes.

Chapter 5

The Robustness and Universality of Large Tolls

In this chapter, we ask if it is possible to compute optimal taxes with minimal information about the system, and present several results showcasing the relationship between available tolling methodologies, uncertainty, and achievable performance. We term this goal "robust coordination," as we desire to incentivize agents to behave as though they are coordinating with one another, but we require that our behavior-influencing mechanisms are robust to mischaracterizations of the system. As discussed in Chapter 2, since price of anarchy is simply a cost metric in worst-case over some set of unknown information, it lends itself naturally to quantifying the robustness of taxation mechanisms to unknown information. Thus, our analysis represents a departure both from the typical descriptive price of anarchy research as well as from the complete-information assumptions of the taxation literature.

The main contribution here is to derive a universal taxation mechanism that guarantees arbitrarily-good performance for any routing game while requiring no prior knowledge of the specific network, user demand profile, or distribution of user sensitivities. That is, the derived taxes are robust to gross mischaracterizations of the above quantities. This

This chapter is adapted, with permission, from [75], scheduled to be published by IEEE Transactions on Automatic Control. © 2018 IEEE.

result holds for networks with general latency functions and any topology, suggesting that surprisingly-little information is required in principle.

Our next result explores the effect of reducing the designer's capabilities while maintaining a high level of uncertainty. To this end, our second contribution is to explore the effect of placing an upper bound on the allowable tolling functions. This may have practical value in settings where very large tolls may be impossible (or politically unpalatable) to implement. For parallel networks with linear-affine latency functions, we derive the optimal tolling functions that minimize worst-case performance degradation for any unknown distribution of user sensitivities and toll upper bound, requiring no prior knowledge of the number of network links. These optimal tolls are simple affine functions of flow. We show that for parallel networks with linear-affine cost functions and simple user demands, the worst-case performance degradation strictly decreases with the toll upper bound. Our results suggest that large tolls can compensate for a poor characterization of user sensitivities. Unfortunately, by imposing an upper bound on allowable taxation functions, optimal behavior can no longer be guaranteed. Thus, this result additionally implies that unbounded tolls are *necessary* to enforce optimal flows if both the network topology and user sensitivities are unknown.

Our results in Section 5.3 explore a further restriction on the designer's capabilities, requiring that tolls do not depend on flow (i.e., requiring fixed tolls rather than tolling functions). This section complements the work on fixed tolls from Chapter 3; here, rather than studying strong/weak robustness, we compute a lower bound on the price of anarchy of fixed tolls for heterogeneous price-sensitive users. In this setting, we show that even if fixed tolls are allowed to depend on the network topology and user demands, they provide relatively poor performance guarantees when the user sensitivities are unknown. Here, by reducing the designer's capability (by disallowing access to flow-varying taxation functions), we dramatically reduce the achievable performance guarantees in the presence of uncertainty. That is, in support of the results of Chapter 3, we show here that fixed tolls are significantly less robust than flow-varying tolls.

5.1 A Universal Taxation Mechanism

In this section, we prove that network- and sensitivity-agnostic tolls exist which can drive the price of anarchy to 1 for general networks and latency functions. We term these "universal" because they take the same form and provide the same performance guarantee regardless of which particular routing scenario they are applied to. Using this taxation mechanism, we show in Theorem 5.1 that for any network, regardless of network topology, traffic rates $\{r^c\}$, or price-sensitivity functions $\{s^c\}$, the price of anarchy can be made arbitrarily close to 1 with sufficiently-large edge tolls, indicating that tolls exist which are robust to mischaracterizations of all the aforementioned system parameters.

Theorem 5.1 *Let $\mathcal{G}$ be the set of multi-commodity routing games where $S_{\mathrm{U}} \geq S_{\mathrm{L}} > 0$. For any network edge $e \in \mathcal{G}$ with convex, nondecreasing, continuously differentiable latency function ℓ_e, define the universal taxation function on edge e with gain parameter $\kappa \geq 0$ as*

$$\tau_e^{\mathrm{u}}(f_e;\kappa) = \kappa\left(\ell_e(f_e) + f_e \cdot \frac{d}{df_e}\ell_e(f_e)\right). \tag{5.1}$$

Then for any routing problem $G \in \mathcal{G}$,

$$\lim_{\kappa\to\infty} \mathrm{PoA}\left(G, T^{\mathrm{gpt}}(\kappa)\right) = 1. \tag{5.2}$$

That is, on *any* network being used by *any* population of users, the total latency can be made arbitrarily close to the optimal latency, and each individual link toll is a simple continuous function of that link's flow. The reason for this is that as κ increases, the original latency function has a smaller and smaller relative effect on the users' cost functions; in the large-toll limit, the only cost experienced by the users is the tolling function itself which is specifically designed to induce optimal Nash flows.

Proof: Using a sequence of tolls, we construct a sequence of Nash flows that converges to an optimal flow. Let κ_n be an unbounded, increasing sequence of tolling coefficients.

For any routing problem $G \in \mathcal{G}$ and price-sensitivities $s \in \mathcal{S}_G$, let $f^n = \left(f_p^n\right)_{p\in\mathcal{P}}$ denote the Nash flow resulting from the tolling coefficient κ_n. For each commodity c, let $\mathcal{P}_n^c \subseteq \mathcal{P}^c$ denote the set of paths that have positive flow in f^n. For any $p \in \mathcal{P}_n^c$, there must be some user $x \in [0, r^c]$ using p with sensitivity s_x^c; the cost experienced by this user is given by

$$J_x^c(f^n) = \sum_{e\in p} \left[\ell_e(f_e) + \kappa_n s_x^c \left(\ell_e(f_e) + f_e \cdot \frac{d}{df_e}\ell_e(f_e)\right)\right].$$

Define $\gamma_{n,x} \triangleq \frac{\kappa_n s_x^c}{1+\kappa_n s_x^c}$. Let $\ell_e^*(f_e) = f_e \cdot \frac{d}{df_e}\ell_e(f_e)$; then for any other path $p' \in \mathcal{P}^c \setminus p$, user x must experience a lower cost on p than on p', or

$$\sum_{e\in p}\ell_e(f_e) - \sum_{e\in p'}\ell_e(f_e) \leq \gamma_{n,x}\left[\sum_{e\in p'}\ell_e^*(f_e) - \sum_{e\in p}\ell_e^*(f_e)\right]. \tag{5.3}$$

Therefore, for any $n \geq 1$, f^n must satisfy some set of inequalities defined by (5.3). Note that for all $c \in \mathcal{C}$ and any $x \in [0, r^c]$, $\lim_{\kappa_n\to\infty} \gamma_{n,x} = 1$, so because all the functions in (5.3) are continuous, f^n converges to a set F^* of feasible flows that satisfy

$$\sum_{e\in p}\ell_e(f_e) - \sum_{e\in p'}\ell_e(f_e) \leq \left[\sum_{e\in p'}\ell_e^*(f_e) - \sum_{e\in p}\ell_e^*(f_e)\right] \tag{5.4}$$

for all c, all $p \in \mathcal{P}^{c*}$, and $p' \in \mathcal{P}^c$, where $\mathcal{P}^{c*} \subseteq \mathcal{P}^c$ is some subset of paths. But inequalities (5.4) (combined with the feasibility constraints on f) also specify a Nash flow for G for a unit-sensitivity population with marginal-cost taxes as specified by (2.12). Any such Nash flow must be optimal [42]; that is, any $f \in F^*$ is a minimum-latency flow for G. Thus, since $\mathcal{L}(f)$ is a continuous function of f,

$$\lim_{n\to\infty} \mathcal{L}\left(f^n\right) = \mathcal{L}^*\left(G\right), \tag{5.5}$$

obtaining the proof of the theorem. ■

5.1.1 Price of Anarchy Bounds for Homogeneous Populations

The result in Theorem 5.1 is encouraging since it ensures that no routing game or user population is so pathological that we cannot enforce optimal routing with sufficiently-high tolls, but it gives no indication of *how high* these tolls must be. In our next result in Proposition 5.2 (which follows from a result in [55]), we state that for homogeneous price-sensitive populations (i.e., all users have the same non-zero price sensitivity), the performance degradation is uniformly bounded in all games by a simple expression.

Proposition 5.2 *If all users have (unknown) homogeneous price-sensitivity $s \geq S_{\rm L} > 0$, the price of anarchy induced by $T^{\rm gpt}(\kappa)$ is given by*

$$\sup_{G \in \mathcal{G}} \mathrm{PoA}\left(G, T^{\rm gpt}(\kappa)\right) \leq \frac{1 + \kappa S_{\rm L}}{\kappa S_{\rm L}}. \tag{5.6}$$

Proof: Immediate from Proposition 6.4 of [55]. ■

5.2 Optimal bounded tolls for parallel affine networks

Of course, it may be impractical or politically infeasible to charge extremely high tolls. For example, if network demand is elastic, very large tolls could induce some users to avoid travel altogether. Therefore, in Theorem 5.3, we analyze the effect of an upper bound on the allowable tolling functions. For simplicity, we focus on the same class of parallel networks that we studied in Section 4.2; additionally, we assume without loss of much (if any) generality that for each network, the total traffic rate is $r = 1$. For parallel networks with affine cost functions in which every edge has positive flow in an un-tolled Nash flow, we explicitly derive the optimal bounded taxation mechanism, and then provide an expression for the price of anarchy. These optimal tolls are simple affine functions of flow, and the price of anarchy is strictly decreasing in the upper bound. Formally, we say a taxation mechanism is *bounded* if all its taxation functions respect some upper bound:

Definition 5.1 *Taxation mechanism T is* bounded *by T on a class of routing problems $\bar{\mathcal{G}}$ if for every edge $e \in \bar{\mathcal{G}}$, T assigns a (possibly flow-varying) tolling function that satisfies*

$$\tau_e : [0,1] \to [0,T]. \tag{5.7}$$

$\mathcal{T}\left(T, \bar{\mathcal{G}}\right)$ denotes the set of mechanisms bounded by T on $\bar{\mathcal{G}}$.

For the following results, let $\mathcal{G}^p \subseteq \mathcal{G}$ represent the class of all single-commodity, parallel-link routing problems with affine latency functions. That is, for all $e \in \mathcal{G}^p$, the latency function satisfies $\ell_e(f_e) = a_e f_e + b_e$ where $a_e \geq 0$ and $b_e \geq 0$ are edge-specific constants. "Single-commodity" implies that all traffic has access to all network edges. Furthermore, we assume as in Section 4.2 that every edge has positive flow in an un-tolled Nash flow. In order to meaningfully discuss uniform toll bounds on a broad class of networks, it is necessary to describe classes of networks with bounded latency functions. To this end, we define $\mathcal{G}\left(\bar{a}, \bar{b}\right) \subset \mathcal{G}^p$ as the set of parallel, affine-cost networks such that for every $e \in \mathcal{G}\left(\bar{a}, \bar{b}\right)$, the latency function coefficients satisfy $a_e \leq \bar{a}$ and $b_e \leq \bar{b}$.

In this chapter we use the following slight refinement of our previous definition of network agnosticity:

Definition 5.2 *For every edge $e \in \mathcal{G}$ with latency function ℓ_e a* network-agnostic *taxation mechanism is a mapping $T^{\mathrm{na}} : [0,1] \times \{\ell_e\}_{e \in \mathcal{G}} \to \{\tau_e\}$ that assigns the following flow-dependent taxation function to edge e:*

$$\tau_e(f_e) = \tau^{\mathrm{na}}\left(f_e; \ell_e\right) \tag{5.8}$$

where $\tau^{\mathrm{na}}(f,\ell)$ *satisfies the following additivity condition:*[1] *for all* $e, e' \in \mathcal{G}$ *and* $f \in [0,1]$,

$$\tau^{\mathrm{na}}(f;\ell_e + \ell_{e'}) = \tau^{\mathrm{na}}(f;\ell_e) + \tau^{\mathrm{na}}(f;\ell_{e'})\,. \tag{5.9}$$

Thus, both marginal-cost tolls (2.12) and universal tolls (5.1) are network-agnostic according to Definition 5.2.

Our goal is to derive the bounded network-agnostic taxation mechanism that minimizes the worst-case selfish routing on $\mathcal{G}^p$. We define the price of anarchy with respect to class of problems $\mathcal{G}$ and bound T as the best price of anarchy we can achieve on $\mathcal{G}$ with a taxation mechanism bounded by T:

$$\mathrm{PoA}_T(\mathcal{G}) \triangleq \inf_{T\in\mathfrak{T}(T,\mathcal{G})} \left\{ \sup_{G\in\mathcal{G}} \mathrm{PoA}(G,T) \right\}. \tag{5.10}$$

Theorem 5.3 *Let* $\mathcal{G}(\bar{a},\bar{b}) \subset \mathcal{G}^p$ *be some subset of parallel, affine-cost networks with finite* $\bar{a}$ *and* $\bar{b}$. *For any toll bound* T *and* $S_{\mathrm{U}} \geq S_{\mathrm{L}} > 0$, *define the set of universal parameters by the tuple* $U_T = \left(S_{\mathrm{L}}, S_{\mathrm{U}}, \bar{a}, \bar{b}\right)$. *Then there exist functions* $\kappa_1(U_T)$ *and* $\kappa_2(U_T)$ *such that the optimal network-agnostic taxation mechanism bounded by* T *on* $\mathcal{G}(\bar{a},\bar{b})$ *assigns tolling functions*

$$\tau_e(f_e) = \kappa_1(U_T)a_e f_e + \kappa_2(U_T)b_e. \tag{5.11}$$

Furthermore, the price of anarchy $\mathrm{PoA}_T\left(\mathcal{G}\left(\bar{a},\bar{b}\right)\right)$ *is given by the following:*

$$\begin{cases} \frac{4}{3}\left(1-\frac{\kappa_1(U_T)S_{\mathrm{L}}}{(1+\kappa_1(U_T)S_{\mathrm{L}})^2}\right) & \text{if } \kappa_1(U_T) < \frac{1}{\sqrt{S_{\mathrm{L}}S_{\mathrm{U}}}} \\ \frac{4}{3}\left(1-\frac{(1+\kappa_1(U_T)S_{\mathrm{L}})\left(\frac{S_{\mathrm{L}}}{S_{\mathrm{U}}}+\kappa_1(U_T)S_{\mathrm{L}}\right)}{\left(1+2\kappa_1(U_T)S_{\mathrm{L}}+\frac{S_{\mathrm{L}}}{S_{\mathrm{U}}}\right)^2}\right) & \text{if } \kappa_1(U_T) \geq \frac{1}{\sqrt{S_{\mathrm{L}}S_{\mathrm{U}}}}. \end{cases} \tag{5.12}$$

See Figure 5.1 for a comparison of the price of anarchy afforded by Theorems 5.1 and 5.3.

[1] The additivity condition in Definition 5.2 requires that two edges connected in series will be assigned the same taxation function as if they were replaced by a single edge whose latency function is the sum of the underlying latency functions. It ensures that the incentive design process be independent of network specifications, isolating the role of network information in the design process.

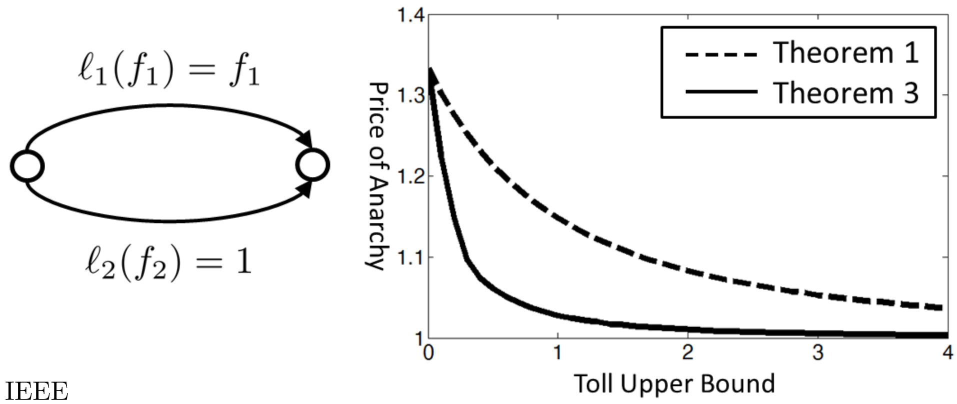

Figure 5.1: Price of Anarchy plot contrasting the Universal toll result from Theorem 5.1 (dashed line) with the optimal toll result from Theorem 5.3 (solid line) on the special case of the two-link network depicted on the left. For both price of anarchy curves, the user sensitivities satisfy $S_{\mathrm{L}} = 1$ and $S_{\mathrm{U}} = 10$. The price of anarchy of either taxation mechanism converges to 1 as the toll upper bound increases, but the solid line converges much more quickly. This is because Theorem 5.3 gives the optimal tolls for a specific class of networks (parallel networks), but the universal tolls from Theorem 5.1 are designed to work on *all* classes of networks.

Note that the tolls of Theorem 5.3 incentivize considerably lower system costs than those of Theorem 5.1; this is due to the fact that Theorem 5.3 is optimized for a smaller class of networks.

A closed-form expression for $\kappa_2(\cdot)$ can be found in the proof of Theorem 5.3 as (5.20); no convenient closed form for $\kappa_1(\cdot)$ appears to exist. Nonetheless, it holds that $\kappa_1(\cdot)$ and $\kappa_2(\cdot)$ are both nondecreasing and unbounded in T; among other things, this implies that $\lim_{T\to\infty} \mathrm{PoA}_T\left(\mathcal{G}\left(\bar{a},\bar{b}\right)\right) = 1$.

We now proceed with the proof of Theorem 5.3, which relies on two supporting lemmas. For our first milestone, we restrict attention to simple affine taxation functions:

Lemma 5.3.1 *Let $T^A(\kappa_1,\kappa_2)$ denote an affine taxation mechanism that assigns tolling functions $\tau_e(f_e) = \kappa_1 a_e f_e + \kappa_2 b_e$. For any $\kappa_{\max} \geq 0$, the optimal coefficients κ_1^* and κ_2^* satisfying*

$$(\kappa_1^*,\kappa_2^*) \in \underset{\kappa_1,\kappa_2\leq\kappa_{\max}}{\arg\min} \left\{ \sup_{G\in\mathcal{G}^p} \mathrm{PoA}\left(G, T^A(\kappa_1,\kappa_2)\right)\right\} \tag{5.13}$$

are given by

$$\kappa_1^* = \kappa_{\max}, \tag{5.14}$$

$$\kappa_2^* = \max\left\{0,\ \frac{\kappa_{\max}^2 S_{\mathrm{L}} S_{\mathrm{U}} - 1}{S_{\mathrm{L}} + S_{\mathrm{U}} + 2\kappa_{\max} S_{\mathrm{L}} S_{\mathrm{U}}}\right\}. \tag{5.15}$$

Furthermore, for any $G \in \mathcal{G}^p$, $\mathrm{PoA}\left(G, T^A(\kappa_1^*, \kappa_2^*)\right)$ is upper-bounded by the following expression:

$$\begin{cases} \frac{4}{3}\left(1 - \frac{\kappa_{\max} S_{\mathrm{L}}}{(1+\kappa_{\max} S_{\mathrm{L}})^2}\right) & \text{if} \quad \kappa_{\max} < \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}} \\ \frac{4}{3}\left(1 - \frac{(1+\kappa_{\max} S_{\mathrm{L}})\left(\frac{S_{\mathrm{L}}}{S_{\mathrm{U}}} + \kappa_{\max} S_{\mathrm{L}}\right)}{\left(1+2\kappa_{\max} S_{\mathrm{L}} + \frac{S_{\mathrm{L}}}{S_{\mathrm{U}}}\right)^2}\right) & \text{if} \quad \kappa_{\max} \geq \frac{1}{\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}}. \end{cases} \tag{5.16}$$

See Section 5.4 for the proof of Lemma 5.3.1.

Next, in Lemma 5.3.2, we investigate the possibility that some other taxation mechanism could perform better than the affine $T^A(\kappa_1^*, \kappa_2^*)$ while still respecting the bound T. To that end, we assume that some arbitrary taxation mechanism outperforms affine tolls, and deduce various properties of these hypothetical tolls. We show that this hypothetical "better" taxation mechanism must universally charge higher tolls than our optimal affine tolls.

Lemma 5.3.2 *Let T^* be any network-agnostic taxation mechanism such that for $\kappa_{\max} \geq 0$*

$$\sup_{G \in \mathcal{G}^p} \mathrm{PoA}\left(\mathcal{G}^p, T^*\right) < \sup_{G \in \mathcal{G}^p} \mathrm{PoA}\left(\mathcal{G}^p, T^A(\kappa_1^*, \kappa_2^*)\right). \tag{5.17}$$

Then T^* must charge strictly higher tolls than $T^A(\kappa_1^*, \kappa_2^*)$ on every edge in every network:

$$\forall\ e \in \mathcal{G}^p, \quad \forall\ f_e \in (0,1], \quad \tau_e^*(f_e) > \tau_e^A(f_e). \tag{5.18}$$

The proof of Lemma 5.3.2 appears in Section 5.4.

Proof: [Proof of Theorem 5.3] For any non-negative κ_1 and κ_2, $T^A(\kappa_1, \kappa_2)$ is tightly bounded by $\left(\kappa_1 \bar{a} + \kappa_2 \bar{b}\right)$ on $\mathcal{G}\left(\bar{a}, \bar{b}\right)$. Note that for κ_1^* and κ_2^* as defined in Lemma 5.3.1, $\left(\kappa_1^* \bar{a} + \kappa_2^* \bar{b}\right)$ is a strictly increasing, continuous function of $\kappa_{\max}$. Thus, for any $T \geq 0$, there is a unique $\kappa_{\max}^* \geq 0$ for which $T^A(\kappa_1^*, \kappa_2^*)$ is tightly bounded by T on $\mathcal{G}\left(\bar{a}, \bar{b}\right)$. We define the function $\kappa_1(U_T)$ as the maximal $\kappa^*_{\max}$ for any $T \geq 0$, given $S_{\rm L}, S_{\rm U}, \bar{a}$, and $\bar{b}$. That is, $\kappa_1(U_T)$ is defined implicitly as the unique function satisfying

$$\kappa_1(U_T)\bar{a} + \max\left\{0, \frac{\left(\kappa_1^2(U_T)S_{\rm L}S_{\rm U} - 1\right)\bar{b}}{S_{\rm L} + S_{\rm U} + 2\kappa_1(U_T)S_{\rm L}S_{\rm U}}\right\} = T. \tag{5.19}$$

We define $\kappa_2(U_T)$ as

$$\kappa_2(U_T) = \max\left\{0, \ \frac{\kappa_1^2(U_T)S_{\rm L}S_{\rm U} - 1}{S_{\rm L} + S_{\rm U} + 2\kappa_1(U_T)S_{\rm L}S_{\rm U}}\right\}. \tag{5.20}$$

Let $e' \in \bar{\mathcal{G}}$ be an edge with latency function $\ell_{e'}(f_{e'}) = \bar{a}f_{e'} + \bar{b}$. By construction, the tolling function assigned by $T^A(\kappa_1(U_T), \kappa_2(U_T))$ to e' satisfies bound T with equality: $\tau^A_{e'}(1) = T$.

Now let T^* be any taxation mechanism with a strictly lower price of anarchy than $T^A(\kappa_1(U_T), \kappa_2(U_T))$. By Lemma 5.3.2, T^* assigns higher tolling functions than $T^A(\kappa_1(U_T), \kappa_2(U_T))$ on *every* edge for every flow rate. In particular, on edge e', $\tau^*_{e'}(1) > \tau^A_{e'}(1) = T$, violating bound T and proving the optimality of $T^A\left(\kappa_1(U_T), \kappa_2(U_T)\right)$ over the space of all network-agnostic taxation mechanisms bounded by T. By substituting $\kappa_1(U_T)$ for $\kappa_{\max}$ in expression (5.16), we obtain the complete price of anarchy expression (5.12). ■

5.3 Comparing bounded tolls with fixed tolls

Recall that Chapter 3 contained a series of negative results regarding fixed tolls, chief of which was Theorem 3.2 which showed that no network-agnostic fixed toll mechanism exists with desirable properties. In light of this, we now ask in Theorem 5.4 what price of anarchy guarantees are possible with fixed tolls if the tolls are allowed to depend on network structure, but user sensitivities are unknown. Since we are allowing these fixed tolls to depend on network structure (e.g., the number of edges in the network), we denote such taxation functions by $T^{\rm ft}(G) = \{T^{\rm ft}_e(G)\}_{e \in G}$. The following theorem demonstrates that any network-dependent fixed-toll taxation mechanism generally provides poor performance guarantees when compared with the optimal bounded taxation mechanism from Theorem 5.3.

Theorem 5.4 *Consider any network-dependent fixed-toll taxation mechanism T^{ft}. For any network $G \in \mathcal{G}^p$,*

$$\sup_{s \in \mathcal{S}} \mathcal{L}^{\mathrm{nf}}\left(G, s, T^{\mathrm{ft}}(G)\right) \geq \sup_{s \in \mathcal{S}} \mathcal{L}^{\mathrm{nf}}\left(G, s, T^{A}(1/S_{\mathrm{U}}, 0)\right), \tag{5.21}$$

with affine tolls $T^A(\cdot)$ as defined in Lemma 5.3.1. Thus,

$$\begin{aligned} \sup_{G \in \mathcal{G}} \mathrm{PoA}\left(G, \tau^{\mathrm{ft}}\right) &\geq \sup_{G \in \mathcal{G}^p} \mathrm{PoA}\left(G, \tau^{A}(1/S_{\mathrm{U}}, 0)\right) \\ &= \frac{4}{3}\left(1 - \frac{S_{\mathrm{L}}/S_{\mathrm{U}}}{\left(1 + S_{\mathrm{L}}/S_{\mathrm{U}}\right)^2}\right). \end{aligned} \tag{5.22}$$

We point out that the right-hand side of (5.22) represents the price of anarchy due to network-agnostic affine tolls for a *very low toll upper bound.* For example, in the canonical Pigou network depicted in Figure 5.1, if $S_{\mathrm{U}} = 10$, affine tolls prescribed by $\tau^A(1/S_{\mathrm{U}}, 0)$ imply a toll upper-bound of just 0.1. As shown in Figure 5.1, the price of anarchy for optimal affine tolls is steeply decreasing in the toll upper-bound, so a designer wishing to exploit the simplicity of fixed tolls may need to accept lower performance guarantees as a result.

Furthermore, it is important to note that Theorem 5.4 shows that T^A, a *network-agnostic* tolling mechanism, provides better performance guarantees (even for moderately low tolls) than T^{ft}, a *network-dependent* tolling mechanism. This shows the power of Theorem 5.3's taxation mechanism: given *less* information, it performs *better* than any fixed-toll taxation mechanism.

See Figure 5.2 for a comparison of the price of anarchy afforded by Theorems 5.3 and 5.4, and note that fixed tolls only outperform flow-varying affine tolls when both uncertainty and the toll upper bound are low. In all other situations, optimal affine tolls provide better

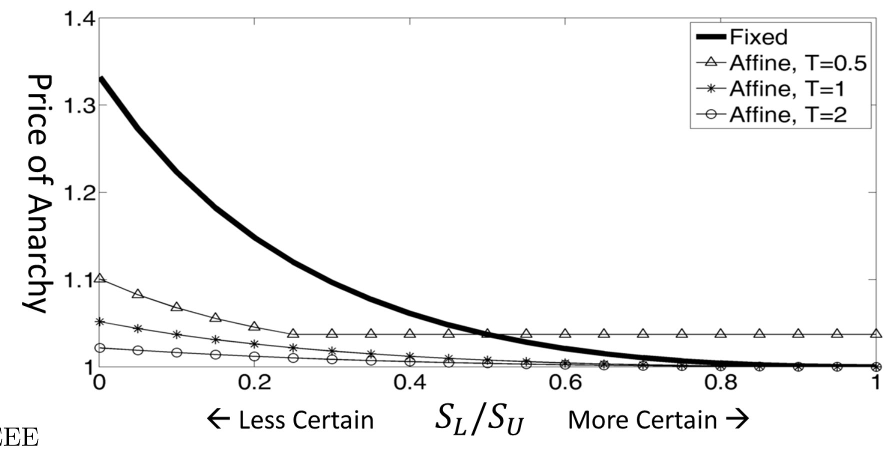

Figure 5.2: Comparison of Price of Anarchy guaranteed by Theorems 5.3 and 5.4. All plots are for $S_{\mathrm{L}} = 1$ and $\bar{a} = \bar{b} = 1$. The horizontal axis represents the level of certainty in price-sensitivity; note that most taxation mechanisms guarantee a price of anarchy of 1 for complete certainty unless they are restricted by a very low upper-bound. The solid line represents the price of anarchy resulting from fixed tolls (according to (5.22)), and the marked lines represent the price of anarchy resulting from optimal flow-varying affine tolls for a given toll bound (according to (5.12)). Note that for a very low toll bound, fixed tolls slightly outperform affine tolls for well-characterized populations; this is due to the fact that the fixed tolls are not restricted by the toll upper bound.

performance guarantees.

The proof of Theorem 5.4 first considers homogeneous sensitivity distributions and then extends to heterogeneous. We write $f^{\mathrm{ft}}(G, s, \tau)$ and $\mathcal{L}^{\mathrm{nf}}(G, s, \tau)$ to denote a Nash flow and its associated total latency induced by fixed tolls $\tau \in \mathbb{R}^n$ on network G, with homogeneous sensitivity $s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]$. Similarly, we write the total latency of a Nash flow resulting from affine tolls $\tau^A(\kappa_1, \kappa_2)$ as $\mathcal{L}^{\mathrm{nf}}\left(G, s, \tau^A(\kappa_1, \kappa_2)\right)$.

Define the optimal fixed tolls τ^* as

$$\tau^* \in \operatorname*{arg\,min}_{\tau \in \mathbb{R}^n} \max_{s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]} \mathcal{L}^{\mathrm{nf}}(G, s, \tau). \tag{5.23}$$

That is, τ^* is in the set of edge tolls that minimize the total latency for the worst possible user sensitivity.

In Lemma 5.4.1, we see that there is a curious relationship between the total latencies of Nash flows resulting from fixed tolls and those resulting from affine tolls $\tau^A(1/S_{\mathrm{U}}, 0)$.

That is, the optimal fixed tolls guarantee the same worst-case performance as affine tolls with extremely low coefficients.

Lemma 5.4.1 *For any $G \in \mathcal{G}^p$, for a homogeneous population, the worst-case total latency resulting from the optimal fixed tolls τ^* is equal to the worst-case total latency resulting from $\tau^A(1/S_{\rm L}, 0)$:*

$$\max_{s\in[S_{\rm L},S_{\rm U}]} \mathcal{L}^{\rm nf}\left(G, s, \tau^*\right) = \max_{s\in[S_{\rm L},S_{\rm U}]} \mathcal{L}^{\rm nf}\left(G, s, \tau^A(1/S_{\rm U}, 0)\right). \tag{5.24}$$

The proof of Lemma 5.4.1 appears in the appendix.

Proof of Theorem 5.4: Since the set of homogeneous populations is a strict subset of the set of heterogeneous ones, we can only make things worse by extending from homogeneous to heterogeneous populations, so the bound in (5.22) must hold. The expression in (5.22) is obtained by substituting $1/S_{\rm U}$ in for $\kappa_{\max}$ in the first part of expression (5.16). ∎

5.4 Chapter Proofs

To prove Lemma 5.3.1, we analytically relate the Nash flows induced by affine tolls with coefficients κ_1 and κ_2 to the Nash flows induced by marginal-cost tolls scaled by κ_1 for some other sensitivity distribution s'. We can then use techniques from Section 4.2 to derive the optimal κ_1 and κ_2.

Proof of Lemma 5.3.1

Let $G \in \mathcal{G}^p$ and $\kappa_1 \geq \kappa_2 \geq 0$.[2] For user $x \in [0, 1]$ with sensitivity $s_x \in [S_{\rm L}, S_{\rm U}]$, the cost of edge $e \in G$ given flow f under affine tolls is given by

$$J_x^e(f) = (1 + \kappa_1 s_x)a_e f_e + (1 + \kappa_2 s_x)b_e.$$

[2]Here, the requirement that $\kappa_1 \geq \kappa_2$ is without loss of generality; later analysis shows that $\kappa_2 > \kappa_1$ would always result in a Nash flow with higher congestion than the un-tolled case.

Note that we may scale $J_x^e(f)$ by any edge-independent factor without changing the underlying preferences of agent x. Thus, without loss of generality, we may write

$$J_x^e(f) = \frac{1+\kappa_1 s_x}{1+\kappa_2 s_x} a_e f_e + b_e. \tag{5.25}$$

Now, define sensitivity distribution s' by the following: for any $x \in [0,1]$, s'_x satisfies

$$s'_x = \frac{s_x(\kappa_1 - \kappa_2)}{\kappa_1(1+\kappa_2 s_x)}. \tag{5.26}$$

By a series of algebraic manipulations, we may combine (5.25) and (5.26) to obtain

$$J_x^e(f) = \left(1+\kappa_1 s'_x\right) a_e f_e + b_e, \tag{5.27}$$

which is simply the cost resulting from scaled marginal-cost tolls (4.2). Thus, for any sensitivity distribution s, we may model a Nash flow resulting from affine tolls with coefficients κ_1 and κ_2 as a Nash flow for sensitivity distribution s' resulting from scaled marginal-cost tolls with $\kappa = \kappa_1$.

Thus, by Theorem 4.2, assuming first that $\kappa_{\max}$ is sufficiently high, our optimal choice of κ_1 is that which satisfies

$$\kappa_1 = \frac{1}{\sqrt{S'_{\mathrm{L}} S'_{\mathrm{U}}}}, \tag{5.28}$$

where S'_{L} and S'_{U} are computed according to (5.26).

Combining (5.26) and (5.28) yields the following characterization of the optimal κ_2 with respect to κ_1, for $\kappa_{\max} \geq (S_{\mathrm{L}} S_{\mathrm{U}})^{-1/2}$:

$$\kappa_2 = \frac{\kappa_1^2 S_{\mathrm{L}} S_{\mathrm{U}} - 1}{S_{\mathrm{L}} + S_{\mathrm{U}} + 2\kappa_1 S_{\mathrm{L}} S_{\mathrm{U}}}. \tag{5.29}$$

Evaluating (4.6) at $q = S'_{\mathrm{L}}/S'_{\mathrm{U}}$ verifies the second part of (5.16) as the correct expression for $\mathrm{PoA}\left(G, T^A(\kappa_1^*, \kappa_2^*)\right)$ when $\kappa_{\max}$ is large.

Consider the case when $\kappa_{\max} < (S_{\mathrm{L}} S_{\mathrm{U}})^{-1/2}$. Now, (5.29) would prescribe a negative value for κ_2, so the optimal choice is to let κ_2 saturate at 0. Now, we are precisely applying scaled marginal-cost tolls with $\kappa = \kappa_1$, so we apply the fact shown in Lemma 4.2.2 that on this class of networks, if $\kappa \leq (S_{\mathrm{L}} S_{\mathrm{U}})^{-1/2}$, the worst-case total latency of a Nash flow always occurs for the extreme low-sensitivity homogeneous sensitivity distribution given by $s_x \equiv S_{\mathrm{L}}$ for all $x \in [0,1]$.

The total latency of a Nash flow for a homogeneous population with sensitivity S_{L} is given by (4.33) as

$$\mathcal{L}^{\mathrm{nf}}(G, S_{\mathrm{L}}, \kappa) = L_R - \frac{\kappa S_{\mathrm{L}}}{(1+\kappa S_{\mathrm{L}})^2}\Theta, \tag{5.30}$$

where L_R and Θ are positive constants depending only on G, satisfying $\Theta \leq L_R$. It is easy to verify that the above expression is minimized on a subset of $\left[0, (S_{\mathrm{L}} S_{\mathrm{U}})^{-1/2}\right]$ by maximizing κ, and using the fact that $\Theta \leq L_R$, we may verify that the price of anarchy for $\kappa_{\max} < (S_{\mathrm{L}} S_{\mathrm{U}})^{-1/2}$ is given by the first part of (5.16), completing the proof of Lemma 5.3.1.

■

Proof of Lemma 5.3.2

Here, we derive properties of any taxation mechanism that outperforms $T^A(\kappa_1^*, \kappa_2^*)$. We define the set of routing problems $\mathcal{G}^0$ as follows: $G \in \mathcal{G}^0$ is a parallel network consisting of two edges, with $\ell_1(f_1) = cf_1$ and $\ell_2(f_2) = c$.

Let $G \in \mathcal{G}^0$. For any c, the optimal flow on G is $(f_1^{\mathrm{opt}}, f_2^{\mathrm{opt}}) = (1/2, 1/2)$ and the optimal total latency is $\mathcal{L}^*(G) = 3c/4$, but the un-tolled Nash flow has a total latency of $\mathcal{L}^{\mathrm{nf}}(G, s, \emptyset) = c$, so the un-tolled price of anarchy is $4/3$. It is straightforward to show furthermore that if $S_{\mathrm{U}} > S_{\mathrm{L}} \geq 0$, for any $\kappa_{\max} > 0$, this network constitutes a worst-case example and the price of anarchy bound of this particular network is tight; i.e., it equals the expression given in (5.16): $\mathrm{PoA}\left(G, T^A(\kappa_1^*, \kappa_2^*)\right) = \sup_{G \in \mathcal{G}^p} \mathrm{PoA}\left(G, T^A(\kappa_1^*, \kappa_2^*)\right)$. Thus, if our hypothetical T^* outperforms T^A in general, it must specifically outperform T^A on any network $G \in \mathcal{G}^0$, or

$$\mathrm{PoA}\,(G, T^*) < \mathrm{PoA}\,\left(G, T^A(\kappa_1^*, \kappa_2^*)\right). \tag{5.31}$$

Now, we investigate the performance of the hypothetical tolling mechanism T^* on networks in $\mathcal{G}^0$. Given a network $G \in \mathcal{G}^0$, T^* assigns edge tolling functions $\tau_1^*(f_1)$ and $\tau_2^*(f_2)$. Recall that since T^* is network-agnostic, there is some function $\tau^*\,(f; a, b)$ such that an edge $e \in E$ with latency function $\ell_e(f_e) = a_e f_e + b_e$ is assigned tolling function $\tau^*(f_e; a_e, b_e)$. By analyzing networks in $\mathcal{G}^0$, we can deduce properties of the function with the 2nd and 3rd arguments set to 0, since $\tau_1^*(f_1) = \tau^*(f_1; c, 0)$ and $\tau_2^*(f_2) = \tau^*(f_2; 0, c)$.

Now we show that T^* must assign higher tolls than $T^A(\kappa_1^*, \kappa_2^*)$. Let $S_{\mathrm{U}} > S_{\mathrm{L}}$. By design, the worst-case Nash flows resulting from $T^A(\kappa_1^*, \kappa_2^*)$ occur for homogeneous populations with $s = S_{\mathrm{L}}$ and $s = S_{\mathrm{U}}$. Since any network $G \in \mathcal{G}^0$ has only 2 links, we can characterize a Nash flow simply by the flow on edge 1; accordingly, let $f_{\mathrm{L}}(c)$ denote the flow as a function of c on edge 1 in the Nash flow resulting from sensitivity distribution $s = S_{\mathrm{L}}$, and $f_{\mathrm{H}}(c)$ the corresponding edge 1 flow for $s = S_{\mathrm{U}}$. These flows are the solutions to the following equations:

$$c f_{\mathrm{L}}(c)\,(1 + \kappa_1^* S_{\mathrm{L}}) = c\,(1 + \kappa_2^* S_{\mathrm{L}})\,, \tag{5.32}$$

$$c f_{\mathrm{H}}(c)\,(1 + \kappa_1^* S_{\mathrm{U}}) = c\,(1 + \kappa_2^* S_{\mathrm{U}})\,. \tag{5.33}$$

Summing (5.32) and (5.33) yields

$$\kappa_1^*\,(f_{\mathrm{L}}(c) - f_{\mathrm{H}}(c)) = \frac{f_{\mathrm{H}}(c)}{S_{\mathrm{U}}} - \frac{f_{\mathrm{L}}(c)}{S_{\mathrm{L}}} + \frac{1}{S_{\mathrm{L}}} - \frac{1}{S_{\mathrm{U}}}. \tag{5.34}$$

It is always true that $f_{\mathrm{H}}(c) < f_{\mathrm{L}}(c)$. By design, $\mathcal{L}(f_{\mathrm{L}}(c)) = \mathcal{L}(f_{\mathrm{H}}(c))$. Note that $\mathcal{L}$ is simply a concave-up parabola in the flow on edge 1.

Now, let $f_{\mathrm{L}}^*(c)$ and $f_{\mathrm{H}}^*(c)$ be defined as the Nash flows resulting from T^* for a given value of c; i.e., the solutions to

$$cf_{\mathrm{L}}^*(c) + \tau_1^*(f_{\mathrm{L}}^*(c))S_{\mathrm{L}} = c + \tau_2^*(1 - f_{\mathrm{L}}^*(c))S_{\mathrm{L}}, \tag{5.35}$$

$$cf_{\mathrm{H}}^*(c) + \tau_1^*(f_{\mathrm{H}}^*(c))S_{\mathrm{U}} = c + \tau_2^*(1 - f_{\mathrm{H}}^*(c))S_{\mathrm{U}}. \tag{5.36}$$

Since T^* guarantees better performance than $T^A(\kappa_1^*, \kappa_2^*)$, it must do so in particular for these homogeneous sensitivity distributions $s = S_{\mathrm{L}}$ and $s = S_{\mathrm{U}}$. Since $\mathcal{L}$ is a parabola, this means that for any c, $f_{\mathrm{H}}(c) < f_{\mathrm{H}}^*(c) < f_{\mathrm{L}}^*(c) < f_{\mathrm{L}}(c)$.

Define the nondecreasing function $\Delta^*(f) = \tau_2^*(f) - \tau_1^*(1-f)$ (which is implicitly also a function of c), so equations (5.35) and (5.36) can be combined and rearranged to show

$$\begin{aligned}\Delta^*(f_{\mathrm{L}}^*(c)) - \Delta^*(f_{\mathrm{H}}^*(c)) &> c\left[\frac{f_{\mathrm{H}}(c)}{S_{\mathrm{U}}} - \frac{f_{\mathrm{L}}(c)}{S_{\mathrm{L}}} + \frac{1}{S_{\mathrm{L}}} - \frac{1}{S_{\mathrm{U}}}\right] \\ &= \kappa_1^* c \left(f_{\mathrm{L}}(c) - f_{\mathrm{H}}(c)\right) \end{aligned} \tag{5.37}$$

The above inequality can be further loosened by replacing $f_{\mathrm{L}}^*(c)$ with $f_{\mathrm{L}}(c)$ and $f_{\mathrm{H}}^*(c)$ with $f_{\mathrm{H}}(c)$, and substituting from (5.34) and rearranging, we finally obtain

$$\frac{\Delta^*(f_{\mathrm{L}}(c)) - \Delta^*(f_{\mathrm{H}}(c))}{f_{\mathrm{L}}(c) - f_{\mathrm{H}}(c)} > \kappa_1^* c. \tag{5.38}$$

Since this must be true for any $c > 0$, the average slope of $\Delta^*(f)$ must be greater than $\kappa_1^* c$ for all $f > 0$. Since $\tau_2^*(f) \geq 0$ this implies that $\tau_1^*(f) > \kappa_1^* cf$ for all $f > 0$, or that

$$\tau^*(f; a, 0) > \tau^A(f; a, 0) \tag{5.39}$$

for all positive f and a.

Now consider the following rearrangement of (5.36):

$$\begin{aligned}\tau_2^*(1-f_{\mathrm{H}}^*(c)) &= [cf_{\mathrm{H}}^*(c)+\tau_1^*(f_{\mathrm{H}}^*(c))-cS_{\mathrm{U}}]\cdot\frac{1}{S_{\mathrm{U}}}\\ &> c\left[(1+\kappa_1^*S_{\mathrm{U}})\,f_{\mathrm{H}}(c)-1\right]\cdot\frac{1}{S_{\mathrm{U}}}\\ &= \kappa_2^*cS_{\mathrm{U}}=\tau_2^A(f).\end{aligned}\tag{5.40}$$

This implies that $\tau_2^*(f)>\kappa_2^*c$ for all $f>0$, or that

$$\tau^*(f;0,b)>\tau^A(f;0,b)\tag{5.41}$$

for all positive f and b.

Finally, note that the additivity assumption of Definition 5.2 implies that $\tau^*(f;a,b)$ is additive in its second and third arguments. That is, we may add inequalities (5.39) and (5.41) to conclude that for all nonnegative f, a, and b, it is true that

$$\tau^*(f;a,b)>\kappa_1^*af+\kappa_2^*b,\tag{5.42}$$

or that a necessary condition for $\sup_{G\in\mathcal{G}^p}\mathrm{PoA}(G,T^*)<\sup_{G\in\mathcal{G}^p}\mathrm{PoA}(G,T^A)$ is that T^* must charge higher tolls on *every* edge in *every* network. ■

Proof of Lemma 5.4.1

We first derive a simple expression for a Nash flow for a homogeneous population as a linear function of the tolls τ. Note that in the context of fixed tolls, Nash flows are unique in cost: for a given routing game, every Nash flow exhibits the same cost on all edges [39,76].

Claim 5.4.1.1 *A Nash flow on $G\in\mathcal{G}$ for sensitivity $s\in\mathcal{S}_1$ and fixed tolls $\tau\in\mathbb{R}^n$ that has positive traffic on all links can be described by the following linear function:*

$$f^{\mathrm{ft}}(G,s,\tau)=R+H(b+s\tau),\tag{5.43}$$

where $R \in \mathbb{R}^n$ and $H \in \mathbb{R}^{n\times n}$ are constant matrices depending only on G. The total latency of this flow is given by the following convex quadratic in τ:

$$\mathcal{L}^{\mathrm{ft}}(G,s,\tau) = L_R + s\tau^T H^T(2AH+I)b + s^2\tau^T H^T AH\tau. \tag{5.44}$$

Proof: Since all users share the same sensitivity, all links have equal cost to all agents in a Nash flow, so when all network edges have positive flow, for any $e_i, e_j \in E$,

$$a_i f_i + b_i + s\tau_i = a_j f_j + b_j + s\tau_j.$$

Similar to the approach in the proof of Lemma 4.2.2 in Section 4.2 (see, for example, (4.10)), a Nash flow $f^{\mathrm{ft}}(G,s,\tau)$ is a solution f to the linear system

$$\underbrace{\begin{bmatrix} a_1 & -a_2 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{P} f = \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{r} + \underbrace{\begin{bmatrix} -1 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}}_{X} (b+s\tau). \tag{5.45}$$

P is invertible, so letting $H = P^{-1}X$ and $R = P^{-1}r$, a Nash flow is given by the linear equation (5.43).

The following observations will be helpful to our proof:

Observation 5.4.1.1 *The matrices H and R satisfy the following properties for any $G \in \mathcal{G}$:*

$$\mathbf{1}^{\mathrm{T}} Hb = \mathbf{0}^{\mathrm{T}}, \tag{5.46}$$

$$\mathbf{1}^{\mathrm{T}} R = 1, \tag{5.47}$$

$$AR \in \mathrm{sp}\left\{\mathbf{1}\right\}, \tag{5.48}$$

$$b^T H^T AHb = -M^T b. \tag{5.49}$$

Finally, every column of $(AH + I)$ *is in* $\mathrm{sp}\{\mathbf{1}\}$.

These facts follow algebraically from the fact that by definition, $f^{\mathrm{ft}}(G, s, \tau)$ satisfies (5.45). Subsutiting (5.43) into (2.1) and simplifying using the facts in Observation 5.4.1.1 yields (5.44).

■

Next, we establish a necessary condition for a set of fixed tolls to be optimal in the sense of (5.23).

Claim 5.4.1.2 *Fixed tolls* τ^* *satisfying* (5.23) *must also satisfy*

$$H\left(\tau^* + \frac{b}{S_{\mathrm{L}} + S_{\mathrm{U}}}\right) = \mathbf{0}. \tag{5.50}$$

Proof: By (5.44) the total latency due to fixed tolls is a convex parabola in s, so for any τ, the maximum total latency on $[S_{\mathrm{L}}, S_{\mathrm{U}}]$ occurs at either S_{L} or S_{U}. Since $\mathcal{L}^{\mathrm{ft}}(G, s, \tau)$ is continuous and convex in τ, this means that τ^* must satisfy

$$\mathcal{L}^{\mathrm{ft}}(G, S_{\mathrm{L}}, \tau^*) = \mathcal{L}^{\mathrm{ft}}(G, S_{\mathrm{U}}, \tau^*). \tag{5.51}$$

Thus, for any optimal fixed tolls τ^*, $\mathcal{L}^{\mathrm{ft}}(G, s, \tau^*)$ is a parabola centered at $s = \frac{S_{\mathrm{L}}+S_{\mathrm{U}}}{2}$:

$$\underset{s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]}{\operatorname{argmin}} \ \mathcal{L}^{\mathrm{ft}}(G, s, \tau^*) = (S_{\mathrm{L}} + S_{\mathrm{U}})/2. \tag{5.52}$$

Our goal is to find the parabola with minimum as in (5.52) which minimizes the values in (5.51).

Equation (5.44) implies that for all $\tau, \tau' \in \mathbb{R}^n$, $\mathcal{L}^{\mathrm{ft}}(G, 0, \tau) = \mathcal{L}^{\mathrm{ft}}(G, 0, \tau')$; that is, the $s = 0$ endpoint of the parabola has the same value for all tolls. Thus, for τ satisfying (5.52), $\mathcal{L}^{\mathrm{ft}}(G, S_{\mathrm{L}}, \tau^*) < \mathcal{L}^{\mathrm{ft}}(G, S_{\mathrm{L}}, \tau)$ if and only if $\mathcal{L}^{\mathrm{ft}}\left(G, \frac{S_{\mathrm{L}}+S_{\mathrm{U}}}{2}, \tau^*\right) < \mathcal{L}^{\mathrm{ft}}\left(G, \frac{S_{\mathrm{L}}+S_{\mathrm{U}}}{2}, \tau\right)$.

By convexity, any tolls which result in globally optimal routing for $s = \frac{S_{\mathrm{L}}+S_{\mathrm{U}}}{2}$ will also be optimal in the sense of (5.23). It is easily verified that for a known homogeneous sensitivity s, any tolls τ which satisfy

$$H\left(\tau + b/(2s)\right) = \mathbf{0} \tag{5.53}$$

result in globally optimal routing. The proof of this is obtained by substituting (5.53) into the gradient (with respect to τ) of $\mathcal{L}^{\mathrm{ft}}(G, s, \tau)$ and applying the facts from Observation 5.4.1.1.

Therefore, any τ which satisfies (5.53) with $s = \frac{S_{\mathrm{L}}+S_{\mathrm{U}}}{2}$ will be uncertainty-optimal. That is, τ^* satisfies (5.50). ■

Evaluating (5.43) with tolls satisfying (5.50) yields an expression for a Nash flow induced by τ^* as a function of s:

$$f^{\mathrm{ft}}(G, s, \tau^*) = R + Hb\frac{S_{\mathrm{L}} + S_{\mathrm{U}} - s}{S_{\mathrm{L}} + S_{\mathrm{U}}}, \tag{5.54}$$

implying that $(R + Hb)$ specifies an un-tolled Nash flow. For parallel networks, it is easy to show that every element of R is non-negative; thus, since $\alpha \triangleq \left(\frac{S_{\mathrm{L}}+S_{\mathrm{U}}-s}{S_{\mathrm{L}}+S_{\mathrm{U}}}\right) \in [0, 1]$, it must be that $(R + Hb\alpha)$ represents a feasible flow.

There are two possible worst-case flows using fixed toll τ^*: one when the sensitivity is S_{U}, the other when the sensitivity is S_{L}. In terms of (5.54), we write these flows as:

$$f_-^{\mathrm{ft}} = f^{\mathrm{ft}}(G, S_{\mathrm{L}}, \tau^*) = R + Hb\left(S_{\mathrm{U}}/(S_{\mathrm{L}} + S_{\mathrm{U}})\right). \tag{5.55}$$

$$f_+^{\mathrm{ft}} = f^{\mathrm{ft}}(G, S_{\mathrm{U}}, \tau^*) = R + Hb\left(S_{\mathrm{L}}/(S_{\mathrm{L}} + S_{\mathrm{U}})\right). \tag{5.56}$$

Next we show that f_-^{ft} and f_+^{ft}, the worst-case flows for optimal fixed tolls, are actually exactly equal to worst-case flows achievable with *scaled marginal-cost* tolls (4.2) with a particular scalar. The machinery of Claim 5.4.1.1 describes the Nash flows $f^{\mathrm{smc}}(G, s, \kappa)$ resulting from homogeneous sensitivity s and marginal-cost tolls scaled by $\kappa > 0$:

$$f^{\mathrm{smc}}(G, s, \kappa) = R + Hb/\left(1 + s\kappa\right). \tag{5.57}$$

The derivation of this is straightforward; it is detailed in Chapter 4.

The *worst* worst-case flows occur when the sensitivity of the population has been grossly over- or under-estimated; for example, if a population with sensitivity S_{U} is using a network with $\kappa = 1/S_{\mathrm{L}}$ (and vice-versa). There are two such flows:

$$f_{-}^{\mathrm{smc}} = R + \frac{Hb}{1 + S_{\mathrm{L}}/S_{\mathrm{U}}} \quad \text{and} \quad f_{+}^{\mathrm{smc}} = R + \frac{Hb}{1 + S_{\mathrm{U}}/S_{\mathrm{L}}}.$$

Comparing the above to (5.55) and (5.56), we see that $f_{-}^{\mathrm{smc}} = f_{-}^{\mathrm{ft}}$ and $f_{+}^{\mathrm{smc}} = f_{+}^{\mathrm{ft}}$. Thus, since

$$\begin{aligned} f^{\mathrm{ft}}(G, S_{\mathrm{L}}, \tau^*) &= f^{\mathrm{smc}}(G, S_{\mathrm{L}}, 1/S_{\mathrm{U}}), \\ f^{\mathrm{ft}}(G, S_{\mathrm{U}}, \tau^*) &= f^{\mathrm{smc}}(G, S_{\mathrm{U}}, 1/S_{\mathrm{L}}), \end{aligned}$$

it must be true that (re-writing now in terms of affine tolls)

$$\mathcal{L}^{\mathrm{nf}}(G, S_{\mathrm{L}}, \tau^*) = \mathcal{L}^{\mathrm{nf}}(G, S_{\mathrm{L}}, \tau^A(1/S_{\mathrm{U}}, 0)), \tag{5.58}$$

$$\mathcal{L}^{\mathrm{nf}}(G, S_{\mathrm{U}}, \tau^*) = \mathcal{L}^{\mathrm{nf}}(G, S_{\mathrm{U}}, \tau^A(1/S_{\mathrm{L}}, 0)). \tag{5.59}$$

By design, (5.58) equals (5.59), so we have that

$$\max_{s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]} \mathcal{L}^{\mathrm{nf}}\left(G, s, \tau^A(1/S_{\mathrm{U}}, 0)\right) = \max_{s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]} \mathcal{L}^{\mathrm{nf}}\left(G, s, \tau^*\right).$$

■

Chapter 6

Avoiding Perverse Incentives

In Chapters 4 and 5, our main goal was to derive taxation mechanisms that optimize the price of anarchy; i.e., reduce the congestion associated with worst-case problem instances. Marginal-cost tolls and their network-agnosticity appeared to offer some promising approaches to accomplishing this goal. However, Proposition 4.1 in Chapter 4 showed that there exist pathological networks and user populations on which marginal-cost tolls actually *increase* the congestion of Nash flows as compared to the un-tolled case. The purpose of this chapter is to investigate this phenomenon more fully, specifically to answer the following question:

Do there exist network-agnostic taxation mechanisms that improve the price of anarchy without degrading congestion on any network?

To examine this rigorously, we study a new performance metric that we term the *perversity index*, formally defined in Section 2.2.3. The perversity index of a taxation mechanism is defined as the ratio between the performance it incentivizes and the un-influenced performance, taken in worst case over user populations and networks. That is, if a taxation mechanism has a perversity index strictly greater than 1, this indicates that networks exist

This chapter contains material that is adapted, with permission, from [77–79], previously published in the proceedings of the 55th and 56th IEEE Conferences on Decision and Control and the 2017 American Control Conference. © 2016, 2017 IEEE on which this mechanism degrades the quality of Nash flows, rather than improves them. If this is the case, we we say that a taxation mechanism is *perverse*.

The main result in this chapter is Theorem 6.1, stating that

If networks are sufficiently complex and the user population is sufficiently diverse, every network-agnostic taxation mechanism is either trivial or perverse.

That is, a network agnostic taxation mechanism improves routing efficiency on some networks only if it degrades efficiency on other networks, and that such perverse incentives can even arise on simple parallel networks.

Nonetheless, we show that there do exist non-trivial classes of networks on which network-agnostic taxation mechanisms can be guaranteed to improve outcomes; in particular, parallel networks in which all traffic can use all paths constitute such a well-behaved network class. In Theorem 6.2 we define the *generalized marginal-cost* taxation mechanism; for a network edge e with delay function $\ell_e(f_e)$, this mechanism assigns taxation functions of

$$\tau_e^{\mathrm{gmc}}(f_e) = \kappa_1 \ell_e(f_e) + \kappa_2 f_e \ell_e'(f_e). \tag{6.1}$$

We show that this is the only non-trivial network-agnostic taxation mechanism that weakly improves routing efficiency on the class of parallel networks (i.e., that has a perversity index of 1). Thus, a system planner can apply generalized marginal-cost tolls on any parallel network without fear of causing perverse incentives.

Following this is a series of characterization results which seeks to understand how the price of anarchy and perversity index of generalized marginal-cost tolls are related. The chapter closes with a note on the connections between marginal-cost pricing and altruistic behavior; in particular, our main impossibility result in Theorem 6.1 implies the existence of "altruism paradoxes" in congestion games in which increasing the altruism of some members of society can actually degrade performance.

Unfortunately, it has recently been shown that if the user population is diverse in price-sensitivity, the optimality guarantees of marginal-cost tolls vanish [1].

6.1 An Impossibility Theorem

Our first question is this: do there exist network-agnostic taxation mechanisms which have a perversity index of 1? Example 6.1 shows that at least the marginal-cost taxation mechanism (2.12) has a perversity index strictly greater than 1; subsequently, Theorem 6.1 shows that this is true for any network-agnostic taxation mechanism.

Example 6.1 *Consider the network depicted in Figure 6.1, consisting of the well-known Braess's Paradox network [28] in parallel with a single constant-latency edge. Let marginal-cost tolls be charged on the network according to* (2.12)*; that is, edges e_1 and e_4 are each charged a flow-varying toll of $\tau_e(f_e) = f_e$. If the user population has 2 units of traffic and a homogeneous toll sensitivity of $s \in [0, 1]$, the unique Nash flow on this network is the one labeled "Efficient Nash Flow" in Figure 6.1, since all agents are experiencing a cost of $2+s$; deviating to the zig-zag path or to e_6 would yield a larger cost of $2 + 2s$ or 3, respectively. Since there are 2 units of traffic experiencing a delay of 2 each, the total latency is $2 \cdot 2 = 4$.*

Now consider a heterogeneous population in which 1 unit of traffic has a sensitivity of $s_1 = 0$ (the orange traffic in Figure 6.1), and 1 unit of traffic has a sensitivity of $s_2 = 1$. In this case, a new Nash flow emerges: one in which all the insensitive traffic uses the zig-zag path, and all the sensitive traffic uses the constant-latency link, labeled "Inefficient Nash Flow" in Figure 6.1. In this flow, any agent on the zig-zag path has a delay of 2, but any agent on the constant-latency path has a delay of 3, for a total latency of $2 + 3 = 5$, which is considerably greater than the un-tolled total latency of 4.

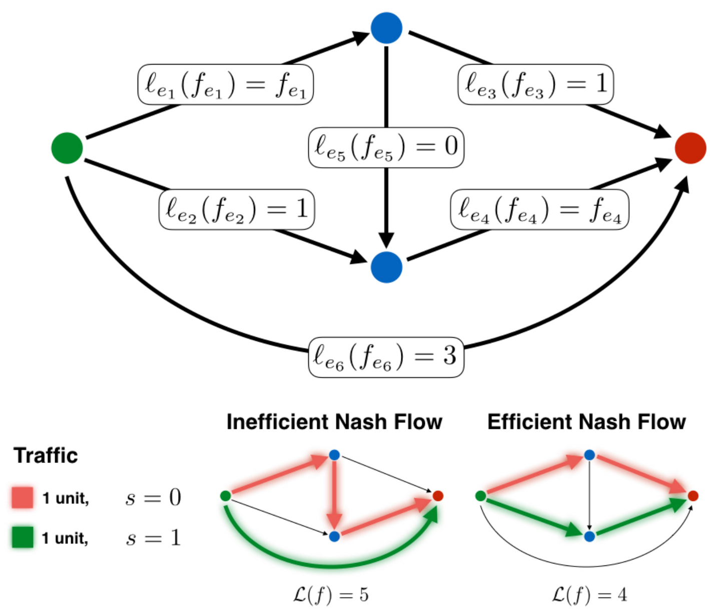

Figure 6.1: Example 6.1: A network demonstrating that marginal-cost tolls are perverse on symmetric networks. The user population has mass $r = 2$, divided equally between commodities with $s = 0$ and $s = 1$. Here, marginal-cost tolls induce more than one Nash flow; two such flows are exhibited in the figure. On the left, all of the insensitive traffic (orange) is using the zig-zag path, experiencing a latency of 2; all the sensitive traffic (green) is using the constant-latency edge, experiencing a latency of 3 – for a total latency of 5. On the right flow, all traffic is experiencing a latency of 2, for a total latency of 4. Here, *any* homogeneous population using this network has the right-hand flow as a unique Nash flow.

6.1.1 Perverse incentives are unavoidable if networks are sufficiently complex

Our first theorem shows that the pathology shown for marginal-cost taxes in Example 6.1 is generic both for symmetric and for parallel networks; that is, all network-agnostic taxation mechanisms can create perverse incentives if the class of networks is rich enough.

Theorem 6.1 *Let $\mathcal{G}^{\mathrm{p}}$ denote the class of all routing problems with parallel networks (not necessarily symmetric), and let $\mathcal{G}^{\mathrm{s}}$ denote the class of all routing problems with symmetric networks (not necessarily parallel). If $S_{\mathrm{U}} > S_{\mathrm{L}} = 0$[1], every non-trivial network-agnostic taxation mechanism has a strictly positive perversity index on $\mathcal{G} \in \{\mathcal{G}^{\mathrm{p}}, \mathcal{G}^{\mathrm{s}}\}$:*

$$\mathrm{PI}\left(\mathcal{G}, T\right) > 1. \tag{6.2}$$

A cornerstone of the proof of Theorem 6.1 is the following lemma, which gives a set of necessary conditions for a network-agnostic taxation mechanism to be non-perverse. The key insight from Lemma 6.1.1 is that all non-perverse taxation mechanisms are essentially a generalized form of marginal-cost tolls.

Lemma 6.1.1 *Let $\mathcal{G}^{\mathrm{sp}}$ denote the class of all routing problems with symmetric parallel networks. If network-agnostic taxation mechanism T has $\mathrm{PI}\left(\mathcal{G}^{\mathrm{sp}}, T\right) = 1$, then for every edge e, it assigns taxation functions satisfying*

$$\tau_e(f_e) = \kappa_1 \ell_e(f_e) + \kappa_2 f_e \ell_e'(f_e), \tag{6.3}$$

where $\kappa_1 > -1/S_{\mathrm{U}}$, $\kappa_2 \geq 0$, and $\kappa_2 \leq \kappa_1 + 1/S_{\mathrm{U}}$.

The proof of Lemma 6.1.1 appears in Section 6.4.

[1]Note that $S_{\mathrm{L}} = 0$ can be satisfied even when 0 mass of traffic has 0 sensitivity; $S_{\mathrm{L}} = 0$ simply ensures that the population's sensitivity is not bounded away from 0.

The space of network-agnostic taxation mechanisms is quite large, but Lemma 6.1.1 reduces the search, allowing us to search over just two parameters, κ_1 and κ_2.

Proof of Theorem 6.1: Here we prove Theorem 6.1 for the case of symmetric networks $\mathcal{G}^{\mathrm{s}}$; the proof for parallel networks $\mathcal{G}^{\mathrm{p}}$ is similar and utilizes the network shown in Figure 4.2.

Lemma 6.1.1 rules out all taxation mechanisms other than those satisfying (6.3), so suppose we are given a taxation mechanism assigning taxes of $\tau_e(f_e) = \kappa_1 \ell_e(f_e) + \kappa_2 f_e \ell_e'(f_e)$, where $\kappa_1 > -1/S_{\mathrm{U}}$, and $\kappa_2 \leq \kappa_1 + 1/S_{\mathrm{U}}$. If $\kappa_2 = 0$, this taxation mechanism is trivial, so let $\kappa_2 > 0$. Our task is to create a user population s (that is, a distribution of tax-sensitivities) and a network G such that $\mathcal{L}^{\mathrm{nf}}(G, s, T(\kappa_1, \kappa_2)) > \mathcal{L}^{\mathrm{nf}}(G, \emptyset)$. We will do this with a population having two sensitivity values $s_2 > s_1 > 0$ and a network resembling that in Figure 6.1. Construct the population as follows: let a unit mass of users have sensitivity s_1 (which we will specify momentarily) and a unit mass have $s_2 = S_{\mathrm{U}}$, for a total of 2 units of traffic. Define $\gamma_2 \triangleq \frac{s_2 \kappa_2}{1 + s_2 \kappa_1} \in (0, 1]$, and choose s_1 so that $\gamma_1 \triangleq \frac{s_1 \kappa_2}{1 + s_1 \kappa_1} = \gamma_2/8$. Then an agent with sensitivity $s_i \in \{s_1, s_2\}$ experiences an effective cost function on edge e of

$$J_e(f_e) = \ell_e(f_e) + \gamma_i f_e \ell_e'(f_e). \tag{6.4}$$

Now, let G be the network depicted in Figure 6.1; let the latency functions on edges e_2 and e_3 be $\ell_{e_2}(f_{e_2}) = \ell_{e_3}(f_{e_3}) = 1 + \gamma_2/8$ and let the latency function on edge e_6 be $\ell_{e_6}(f_{e_6}) = 2 + \gamma_2$. Enumerate the paths as follows: denote the "zig-zag" path $p_1 = \{e_1, e_5, e_4\}$, the remaining two paths in the upper sub-network $p_2 = \{e_1, e_3\}$ and $p_3 = \{e_2, e_4\}$, and the isolated constant-latency path $p_4 = \{e_6\}$; and denote the path flow of p_i by f_i. We will refer to paths p_1, p_2, and p_3 in the upper subnetwork as the "Braess subnetwork."

On this network, the flow (depicted on the left in Figure 6.1) $f^{\mathrm{perverse}} \triangleq (1, 0, 0, 1)$ is a Nash flow for this population, with total latency $\mathcal{L}^{\mathrm{nf}}(G, s, T) = 4 + \gamma_2$. However, it can be verified that if tolls are removed, the unique Nash flow is $f^{\mathrm{nf}} \triangleq (\gamma_2/4, 1 - \gamma_2/8, 1 - \gamma_2/8, 0)$, which has a total latency of $\mathcal{L}^{\mathrm{nf}}(G,\emptyset) = 4 + \gamma_2/2$, or

$$\mathcal{L}^{\mathrm{nf}}(G,s,T) > \mathcal{L}^{\mathrm{nf}}(G,\emptyset) \tag{6.5}$$

and the considered tolls are perverse. ■

6.1.2 Symmetric parallel networks prevent perverse incentives

Theorem 6.1 shows that it does not take much complexity to render a network-agnostic taxation mechanism perverse. Does this mean that it is never possible to achieve a perversity index of 1? Fortunately, the answer is no – and our Theorem 6.2 shows that on symmetric parallel networks (precisely the intersection of $\mathcal{G}^{\mathrm{s}}$ and $\mathcal{G}^{\mathrm{p}}$), the necessary condition from Lemma 6.1.1 is also sufficient to achieve a perversity index of 1. Thus, Theorem 6.2 gives a full characterization of non-perverse taxation mechanisms for symmetric parallel networks.

Theorem 6.2 *Let $\mathcal{G}^{\mathrm{sp}}$ denote the class of routing problems with symmetric parallel networks. For any $S_{\mathrm{U}} \geq S_{\mathrm{L}} \geq 0$, a network-agnostic taxation mechanism T has unity perversity index on $\mathcal{G}^{\mathrm{sp}}$*

$$\mathrm{PI}\left(\mathcal{G}^{\mathrm{sp}},T\right) = 1 \tag{6.6}$$

if and only if $T = T(\kappa_1,\kappa_2)$, assigning the tolling functions

$$\tau_e(f_e) = \kappa_1 \ell_e(f_e) + \kappa_2 f_e \ell_e'(f_e), \tag{6.7}$$

with $\kappa_1 > -1/S_{\mathrm{U}}$, $\kappa_2 \geq 0$, and $\kappa_2 \leq \kappa_1 + 1/S_{\mathrm{U}}$.

Note that if $\kappa_1 = 0$, then for any $\kappa_2 \geq 0$, the above corresponds to simple scaled marginal-cost tolls. In this case, the coefficient constraints reduce to $\kappa_2 \in [0, 1/S_{\mathrm{U}}]$; that is, Theorem 6.2 says that scaled marginal-cost taxes have a perversity of 1 if and only if they are scaled *conservatively,* i.e., they are no larger than would be required to induce optimal flows for a homogeneous population of high sensitivity S_{U}.

6.1.3 The price of anarchy of non-perverse tolls

Having shown that symmetric parallel networks do admit non-perverse taxation mechanisms, we now ask how effective those mechanisms are in reducing worst-case congestion. Simply because taxes have a perversity index of 1 does not immediately imply that their associated PoA is small; nonetheless, we show that generalized marginal-cost tolls can provide modest reductions of worst-case congestion.

Theorem 6.3 *Let $\mathcal{G}_d^{\rm sp}$ denote the class of all symmetric parallel networks with polynomial latency functions of degree at most $d \geq 1$. For any $S_{\rm U} \geq S_{\rm L} \geq 0$, levy the generalized marginal-cost taxation mechanism $T(\kappa_1, \kappa_2)$ as defined in (6.7) with coefficients $\kappa_1 \geq -1/S_{\rm U}$, $\kappa_2 \geq 0$, and $\kappa_2 \leq \kappa_1 + 1/S_{\rm U}$. Let $\beta_{\kappa_1,\kappa_2} \triangleq \frac{\kappa_2 S_{\rm L}}{1+\kappa_1 S_{\rm L}} \in [0,1]$. Then the price of anarchy associated with these tolls is*

$$\mathrm{PoA}\left(\mathcal{G}_d^{\rm sp}, T(\kappa_1,\kappa_2)\right) = \frac{1}{1 + d\beta_{\kappa_1,\kappa_2} - d\left(\frac{1+d\beta_{\kappa_1,\kappa_2}}{1+d}\right)^{\frac{d+1}{d}}}. \tag{6.8}$$

We provide a new proof of Theorem 6.3 in Section 6.4 that relies on our arguments for Theorem 6.2, but note that it is also a consequence of [55, Theorem 7.1], which gives the price of anarchy associated with heterogeneous, partially-altruistic populations. Though the two proofs are substantially different, they share the high-level idea that on symmetric parallel networks, increasing the fraction of players that are merely delay-averse always leads to worse congestion. This implies (in our model) that the price of anarchy is realized by a homogeneous population with sensitivity equal to $S_{\rm L}$, and the expression in (6.8) due to [80] applies immediately.

Furthermore, this yields the following simple characterization of the PoA-minimizing coefficients κ_1 and κ_2, where here we minimize the price of anarchy subject to a constraint that the tolls are non-perverse. Before stating the result, we point out that worst-case performance guarantees provided by a taxation mechanism can often be improved by increasing all edge tolls appropriately (see, e.g., discussion in Chapter 5). In order to make meaningful statements about congestion-minimizing tolls, it is useful to parameterize tolls by a stylized upper-bound; the parameter $\kappa_{\max} > 0$ plays this role in the following result. Thus, Corollary 6.4 solves the following optimization problem:

$$(\kappa_1^*, \kappa_2^*) \in \underset{\kappa_1, \kappa_2 \leq \kappa_{\max}}{\arg\inf} \ \mathrm{PoA}\left(\mathcal{G}_d^{\mathrm{sp}}, T(\kappa_1, \kappa_2)\right). \tag{6.9}$$

Corollary 6.4 *For any $d \geq 1$ and taxation coefficient upper bound $\kappa_{\max}$, the price of anarchy in (6.9) due to bounded generalized marginal-cost taxes is minimized by setting $\kappa_2^* = \kappa_{\max}$ and $\kappa_1^* = \kappa_{\max} - 1/S_{\mathrm{U}}$. If $S_{\mathrm{U}} = +\infty$, then this simplifies to $\kappa_1^* = \kappa_2^* = \kappa_{\max}$.*

Proof: The PoA expression in (6.8) is decreasing in $\beta_{\kappa_1,\kappa_2}$, which for any $S_{\mathrm{L}}, S_{\mathrm{U}}$ and fixed $\kappa_{\max}$ is maximized by saturating the bounds $\kappa_1 \geq \kappa_2 - 1/S_{\mathrm{U}}$ and $\kappa_2 \leq \kappa_{\max}$. ∎

The price of anarchy due to tolls as in Corollary 6.4 is plotted for several values of d in Figure 6.2. Note that even when d is unbounded (the dotted red curve in Figure 6.2), the price of anarchy is bounded whenever $S_{\mathrm{L}}/S_{\mathrm{U}} > 0$.

In the special case of $S_{\mathrm{U}} = +\infty$ (that is, no upper bound on sensitivity is known), the PoA-minimizing coefficients in Corollary 6.4 reduce to $\kappa_1 = \kappa_2 = \kappa_{\max}$. Note that this is identically the universal taxation mechanism from [75], which was developed to serve an entirely different purpose of optimizing the price of anarchy in the large-$\kappa_{\max}$ limit.

6.2 Quantifying the Tradeoff Between Optimality and Perversity

In this section we initiate a study on the tradeoff between optimizing for the perversity index and optimizing for the PoA. As an initial step, we characterize the price of anarchy for homogeneous populations associated with the taxation mechanism $T(\kappa_1, \kappa_2)$. Intuitively, Theorem 6.5 shows that the PoA-minimizing taxation mechanism (for homogeneous pop-

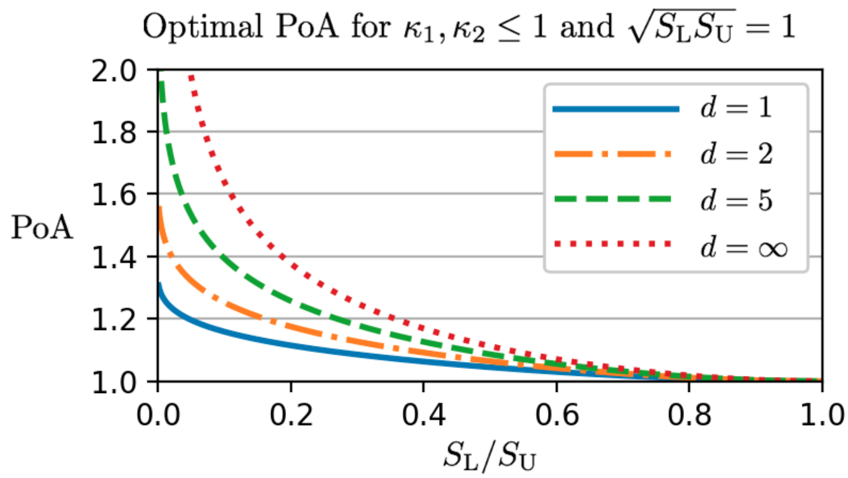

Figure 6.2: The optimal price of anarchy achievable using non-perverse network-agnostic tolls, where d indicates the largest degree of polynomial allowed in the considered latency functions. These values are plotted using the machinery of Theorem 6.3. The PoA is plotted with respect to S_L/S_U, which can serve as a proxy for the variance of the price sensitivities in the user population. On the far left, the price of anarchy resolves to the un-tolled value; on the right, the price of anarchy is 1. Our result continuously bridges the space in between.

ulations) is the taxation mechanism that perfectly balances the harm that can be caused by a homogeneous S_L population with the harm that can be caused by a homogeneous S_U population. Crucially, this taxation mechanism has a perversity index equal to the price of anarchy and strictly greater than 1. That is, the price of anarchy cannot be minimized without risking perverse incentives. In all the following results, we frequently refer to the quantity

$$\beta(s, (\kappa_1, \kappa_2)) := \frac{\kappa_2 s}{1 + \kappa_1 s}, \tag{6.10}$$

since under the influence of $T(\kappa_1, \kappa_2)$, a user with sensitivity s experiences a cost on edge e equivalent to

$$\ell_e(f_e) + \beta(s(\kappa_1, \kappa_2)) f_e \ell_e'. \tag{6.11}$$

That is, $\beta(s, (\kappa_1, \kappa_2))$ indicates users' induced sensitivity to their marginal effect on others: when $\beta = 1$, users interpret their marginal effect on others correctly, as though they were being charged perfect marginal-cost tolls. When $\beta < 1$, users are overly delay-sensitive; when $\beta > 1$, users are not delay-sensitive enough. Note that the non-perverse tolls defined in Theorem 6.2 are crafted to guarantee that every user has $\beta \leq 1$.

First, we present the following lemma, which is adapted from [80, Propositions 7.6, 7.7]. Here, we give the price of anarchy for known-sensitivity homogeneous population for fixed κ_1, κ_2:

Lemma 6.4.1 (Meir and Parkes, 2015 [80]) *Let $\mathcal{G}_d$ be the class of routing problems with polynomial latency functions of degree no more than $d \geq 1$, and let $\mathcal{G}$ denote the class of all routing problems with polynomial latency functions. In the following, let $\kappa := (\kappa_1, \kappa_2)$ and let $\beta(s,\kappa) := \frac{\kappa_2 s}{1+\kappa_1 s} \geq 0$. Then the price of anarchy resulting from $T(\kappa_1, \kappa_2)$ for a homogeneous population with sensitivity s is*

$$\overline{\mathrm{PoA}}(\mathcal{G}_d, s, \kappa) = \begin{cases} \dfrac{1}{1+d\beta(s,\kappa)-d\left(\frac{1+d\beta(s,\kappa)}{1+d}\right)^{\frac{d+1}{d}}} & \text{if } \beta(s,\kappa) \leq 1, \\ \beta(s,\kappa)^{-d}\left(\frac{1+d\beta(s,\kappa)}{1+d}\right)^{d+1} & \text{if } \beta(s,\kappa) > 1. \end{cases} \tag{6.12}$$

If d is not known, the PoA is given by

$$\overline{\mathrm{PoA}}(\mathcal{G}, s, \kappa) = \begin{cases} \left(\beta(s,\kappa)(1-\log\beta(s,\kappa))\right)^{-1} & \text{if } \beta(s,\kappa) \leq 1, \\ \beta(s,\kappa)\exp\left(\frac{1}{\beta(s,\kappa)}-1\right) & \text{if } \beta(s,\kappa) > 1. \end{cases} \tag{6.13}$$

Lemma 6.4.1 is a consequence of [80, Propositions 7.6, 7.7], since for a homogeneous population with sensitivity s, setting taxation functions equal to $\tau_e(f_e) = \kappa_1 \ell_e(f_e) + \kappa_2 f_e \ell_e'(f_e)$ induces the same subjective costs as charging marginal-cost taxes (2.12) to a homogeneous population with sensitivity $\beta(s,\kappa) := \frac{\kappa_2 s}{1+\kappa_1 s}$. The limit-invariant expressions (6.13) are found by taking the limit as $d \to \infty$. Given this expression, we are ready to present the theorem:

Theorem 6.5 *Let $\mathcal{G}_d$ denote the class of all networks with polynomial latency functions of degree at most $d \geq 1$. For homogeneous populations, for any $0 \leq S_{\mathrm{L}} < S_{\mathrm{U}}$ and taxation* *coefficient upper bound* $\kappa_{\max}$*, the PoA-minimizing toll scalars*

$$(\kappa_1^*, \kappa_2^*) \triangleq \inf_{\kappa_1, \kappa_2 \leq \kappa_{\max}} \mathrm{PoA}^{\mathrm{hm}}\left(\mathcal{G}_d, T(\kappa_1, \kappa_2)\right) \tag{6.14}$$

have $\kappa_2^* = \kappa_{\max}$*, and* κ_1^* *is the unique solution on the interval* $(-1/S_{\mathrm{U}}, \kappa_{\max} - 1/S_{\mathrm{U}})$ *to*

$$\overline{\mathrm{PoA}}(\mathcal{G}_d, S_{\mathrm{L}}, (\kappa_1^*, M)) = \overline{\mathrm{PoA}}(\mathcal{G}_d, S_{\mathrm{U}}, (\kappa_1^*, M)), \tag{6.15}$$

where $\overline{\mathrm{PoA}}$ *is defined in* (6.12)*. Its resulting price of anarchy and perversity index for homogeneous populations are equal and greater than one:*

$$\mathrm{PoA}^{\mathrm{hm}}\left(\mathcal{G}_d, T(\kappa_1^*, \kappa_2^*)\right) = \mathrm{PI}^{\mathrm{hm}}\left(\mathcal{G}_d, T(\kappa_1^*, \kappa_2^*)\right) > 1. \tag{6.16}$$

Before presenting the proof, we wish to compare the optimal price of anarchy from Theorem 6.5 to the price of anarchy of the non-perverse (conservative) taxation mechanism from Theorem 6.2; we do this by plotting the ratio of the two in Figure 6.3. Note that even when d is unbounded, for the plotted parameter values, the conservative PoA is no more than about 12% worse than the optimal PoA; this suggests that the benefits of perversity may be rather limited. The following lemma will be instrumental in proving the last part of Theorem 6.5; its proof appears in Section 6.4.

Lemma 6.5.1 *Let* $\mathcal{G}_d^{\mathrm{sp}}$ *denote the class of routing problems with symmetric parallel networks and polynomial latency functions of degree no more than* $d \geq 1$*. Given* $\kappa_2 > 0$*, when* $\kappa_1 \in (-1/S_{\mathrm{U}}, \kappa_2 - 1/S_{\mathrm{U}})$*, the homogeneous perversity index of* $T(\kappa_1, \kappa_2)$ *is greater than* 1 *and equal to the price of anarchy experienced by a population with sensitivity* S_{U}*:*

$$\mathrm{PI}^{\mathrm{hm}}\left(\mathcal{G}_d^{\mathrm{sp}}, T(\kappa_1, \kappa_2)\right) = \overline{\mathrm{PoA}}(\mathcal{G}_d, S_{\mathrm{U}}, (\kappa_1, \kappa_2)). \tag{6.17}$$

Proof of Theorem 6.5: The restriction to homogeneous populations here allows us to apply

Lemma 6.4.1 directly, and leverage its monotonicity properties to obtain the result. Consider the expressions given by (6.12) as a function of $\beta(s,\kappa)$. The price of anarchy as a function of β is bowl-shaped: when $\beta(s,\kappa) < 1$, the price of anarchy is strictly decreasing in $\beta(s,\kappa)$, when $\beta(s,\kappa) > 1$, the price of anarchy is strictly increasing in $\beta(s,\kappa)$, and when $\beta(s,\kappa) = 1$, the price of anarchy is equal to 1. Thus, minimizing the price of anarchy reduces to choosing κ_1 and κ_2 such that for all $s \in [S_{\rm L}, S_{\rm U}]$, $\beta(s,\kappa)$ takes values as "close" to 1 as possible, where this closeness is measured by the expressions in (6.12). Furthermore, when $\kappa_1 > -1/S_{\rm U}$ and $\kappa_2 \geq 0$, the monotonicity of $\beta(s,\kappa)$ ensures that the price of anarchy is achieved by an extreme sensitivity population with $s \in \{S_{\rm L}, S_{\rm U}\}$.

For any fixed $\kappa_2 > 0$ and fixed $s \in [S_{\rm L}, S_{\rm U}]$, $\beta(s,\kappa)$ is decreasing in κ_1 whenever $\kappa_1 > -1/S_{\rm U}$; also, when $\kappa_1 = \kappa_2 - 1/s$, we have $\beta(s,\kappa) = 1$ (thus, the PoA for that s is 1). Combined with the above, this means the price of anarchy is minimized in κ_1 on the interval $I := (\max\{-1/S_{\rm U}, \kappa_2 - 1/S_{\rm L}\}, \kappa_2 - 1/S_{\rm U})$.

Accordingly, for any κ_2, let $\delta := \kappa_{\max} - \kappa_2$. For any $\kappa_2 > 0$ and $\kappa_1 \in I$, it holds that

$$\beta(S_{\rm L}, (\kappa_1, \kappa_2) < \beta(S_{\rm L}, (\kappa_1 + \delta, \kappa_2 + \delta)),$$

and

$$\beta(S_{\rm U}, (\kappa_1, \kappa_2) > \beta(S_{\rm U}, (\kappa_1 + \delta, \kappa_2 + \delta)).$$

That is, when $\kappa_1 \in I$, the price of anarchy can be decreased by adding δ to both κ_1 and κ_2, showing that $\kappa_2^* = \kappa_{\max}$.

Finally, given that $\kappa_2^* = \kappa_{\max}$ and $\kappa_1 \in I$, note that $\overline{\rm PoA}(\mathcal{G}_d, S_{\rm L}, (\kappa_1, \kappa_2^*))$ is strictly increasing in κ_1, that $\overline{\rm PoA}(\mathcal{G}_d, S_{\rm U}, (\kappa_1, \kappa_2^*))$ is strictly decreasing in κ_1, and both are continuous for all κ_1. This guarantees that the PoA-minimizing κ_1^* must be the unique solution to (6.15) on the interval[2] $(-1/S_{\rm U}, \kappa_{\max} - 1/S_{\rm U})$, and (6.16) is an immediate consequence of Lemma 6.5.1. ■

[2]Though our argument considers only the interval I, our uniqueness claim holds on $(-1/S_{\rm U}, \kappa_{\max} - 1/S_{\rm U})$ because for $\kappa_1 \in (-1/S_{\rm U}, \kappa_{\max} - 1/S_{\rm L})$, it holds that $\overline{\rm PoA}(\mathcal{G}_d, S_{\rm U}, (\kappa_1, \kappa_{\max})) \geq \overline{\rm PoA}(\mathcal{G}_d, S_{\rm L}, (\kappa_1, \kappa_{\max}))$.

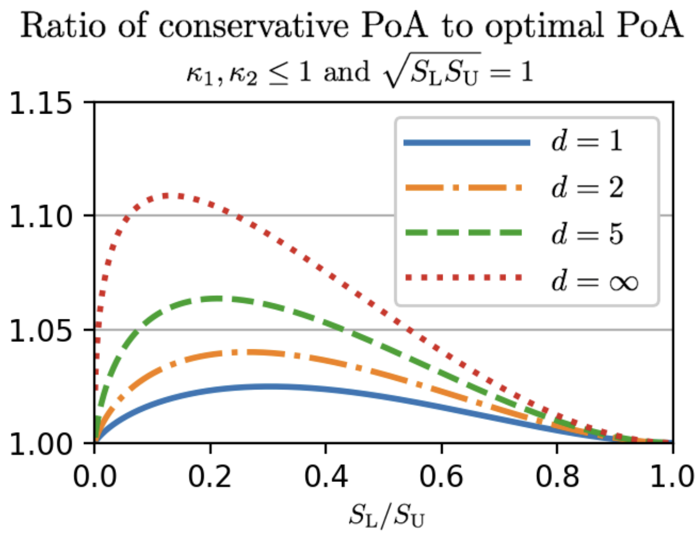

Figure 6.3: The conservative (non-perverse) PoA divided by the minimum-achievable PoA as a function of the sensitivity ratio $S_{\mathrm{L}}/S_{\mathrm{U}}$. This indicates the cost of avoiding perverse incentives, in the sense that if perverse incentives are not allowed, the price of anarchy can be worse by a factor indicated in this plot.

Section 6.2 assumes that minimizing the price of anarchy is the system planner's sole design goal, and Theorem 6.5 has shown that this cannot be done without creating perverse incentives on some networks. Nonetheless, there may be situations in which the planner wishes to reduce the price of anarchy, but desires to limit perversity as well. To explore the effect of this additional constraint, here we characterize the space of achievable PoA and PI. That is, if planner can tolerate perversity no more than some number α, this imposes a lower bound on the achievable price of anarchy. Proposition 6.6 provides tools to compute a full characterization of this tradeoff for homogeneous populations on symmetric parallel networks with polynomial latency functions. Formally, we wish to find

$$(\kappa_1^\alpha, \kappa_2^\alpha) \triangleq \inf_{\kappa_1, \kappa_2 \leq \kappa_{\max}} \mathrm{PoA}^{\mathrm{hm}}\left(\mathcal{G}_d, T(\kappa_1, \kappa_2)\right), \tag{6.18}$$

subject to $\mathrm{PI}^{\mathrm{hm}}(\mathcal{G}_d^{\mathrm{sp}}, T(\kappa_1^\alpha, \kappa_2^\alpha)) \leq \alpha$.

Proposition 6.6 *Let $\mathcal{G}_d^{\mathrm{sp}}$ be the class of symmetric parallel networks with polynomial latency functions of degree no more than $d \geq 1$. First, $\kappa_2^\alpha = \kappa_{\max}$. Let $\kappa_1^\dagger$ be the unique* *solution to*

$$\overline{\mathrm{PoA}}\Big(\mathcal{G}_d, S_{\mathrm{U}}, (\kappa_1^{\dagger}, M)\Big) = \alpha, \tag{6.19}$$

and let κ_1^* *be as defined in Theorem 6.5. Then* $\kappa_1^{\alpha} = \max\left\{\kappa_1^*, \kappa_1^{\dagger}\right\}$. *The resulting price of anarchy is achieved by a low-sensitivity population:*

$$\mathrm{PoA}(\mathcal{G}_d^{\mathrm{sp}}, T(\kappa_1^{\alpha}, \kappa_2^{\alpha})) = \overline{\mathrm{PoA}}(\mathcal{G}_d, S_{\mathrm{L}}, T(\kappa_1^{\alpha}, \kappa_2^{\alpha}))\,, \tag{6.20}$$

where $\overline{\mathrm{PoA}}$ *is defined in Lemma 6.4.1.*

Proof: It is shown in the proof of Theorem 6.5 that the optimal price of anarchy is achieved with $\kappa_2^{\alpha} = \kappa_{\max}$; that applies here as well. Lemma 6.5.1 shows that whenever $\kappa_1 < \kappa_2 - 1/S_{\mathrm{U}}$, the perversity index of $T(\kappa_1, \kappa_2)$ is equal to the price of anarchy for S_{U}; thus, choosing $\kappa_1^{\dagger}$ as the solution to (6.19) ensures a perversity index of exactly α. If α is less than the minimum-achievable price of anarchy for these parameters, then $\kappa_1^{\alpha} = \kappa_1^{\dagger}$. On the other hand, if the minimum-achievable PoA is lower than α, then by definition it can be achieved by setting $\kappa_1^{\alpha} = \kappa_1^*$. In either case, since the tolls are being applied (weakly) conservatively, the price of anarchy is realized by a population with sensitivity S_{L} (this can be verified by the expressions (6.12)). ■

6.3 Implications for Altruistic Behavior

The foregoing has assumed that users are selfish and act with the sole objective of minimizing personal cost. However, real users may act altruistically, with the public good in mind. This is investigated in the α-altruism model, which has no tolls but assigns each user x an *altruism level* $\alpha_x \in [0, 1]$; a user with $\alpha = 0$ is totally selfish, whereas a user with $\alpha = 1$ is totally altruistic [55]. This is modeled by assuming that on edge e, user x with altruism level α_x experiences cost

$$J_e^x(f_e) = \ell_e(f_e) + \alpha_x f_e \ell_e'(f_e). \tag{6.21}$$

In other words, a totally-altruistic user fully accounts for the marginal effects that his actions have on those around him.

By comparing the cost functions induced by marginal-cost tolls (2.12) with the cost functions experienced by altruistic players (6.21), it is clear that there is a deep connection between this model of α-altrusim and the theory of marginal-cost taxation. In essence, marginal-cost taxes are designed to induce artificial altruism in the user population.

The authors of [55] exhibit two contexts in non-atomic congestion games in which worst-case performance improves with increasing levels of altruism: the first is in general networks with homogeneous altruism, and the second is in parallel networks with heterogeneous altruism. In both cases, if the average level of altruism in the population increases, worst-case performance improves.

Given the equivalence of marginal-cost taxation and altruism, our Corollary 6.7 strengthens the parallel-network result of [55], showing that *on any network*, the worst-case flows are realized by a low-altruism homogeneous population. On the other hand, given our impossibility result in Theorem 6.1, Corollary 6.7 shows that increased altruism does not, in general, improve performance. That is, on the network in Figure 6.1, a totally-selfish population is associated with the efficient Nash flow, but a partially-altruistic population is associated with the inefficient Nash flow.

In the following, $\mathcal{L}^{\mathrm{nf}}_{\mathrm{alt}}(G,\alpha)$ denotes the worst-case Nash flow total latency on G for a given altruism distribution α, where users in α take altruism levels in the interval $[A_{\mathrm{L}}, A_{\mathrm{U}}] \subseteq [0,1]$. A homogeneous altruism distribution in which all users have value A_{L} is denoted α^{L}.

Corollary 6.7 *For any $G \in \mathcal{G}^{\mathrm{sp}}$,*

$$\mathcal{L}^{\mathrm{nf}}_{\mathrm{alt}}(G,\alpha) \leq \mathcal{L}^{\mathrm{nf}}_{\mathrm{alt}}\left(G,\alpha^{\mathrm{L}}\right). \tag{6.22}$$

However, there exist $G \in \mathcal{G}^{\mathrm{s}} \cup \mathcal{G}^{\mathrm{p}}$ such that there exists an altruism distribution α satisfying

$$\mathcal{L}^{\mathrm{nf}}_{\mathrm{alt}}(G, \alpha) > \mathcal{L}^{\mathrm{nf}}_{\mathrm{alt}}\left(G, \alpha^{\mathrm{L}}\right). \tag{6.23}$$

Proof: Any Nash flow induced by the tolls of Theorem 6.2 is a Nash flow for some altruism distribution (see, e.g., the argument in the proof of Lemma 6.7.1 in Section 6.4). Thus, Corollary 6.7 is implied by Theorems 6.1 and 6.2. ■

6.4 Chapter Proofs

Proof of Lemma 6.1.1: We shall exhibit example networks on which various tolls are perverse, thus eliminating all but tolls of the form in (6.7). First, consider the network in Figure 6.4(a).

This network has two paths in parallel; the first path is a pair of edges in series with arbitrary latency functions ℓ_1 and ℓ_2, the second path consists of a single edge with latency function ℓ_3 satisfying $\ell_1 + \ell_2 = \ell_3$. For any such network, any nominal Nash flow f^{nf} is optimal; thus, a non-perverse taxation mechanism T would need to incentivize a flow f^T that satisfies $f^T = f^{\mathrm{nf}} = f^{\mathrm{opt}} = (r/2, r/2)$ by charging tolls satisfying $\tau_1(r/2) + \tau_2(r/2) = \tau_3(r/2)$. That is, T is additive: if $\ell_1 + \ell_2 = \ell_3$, it is true that $T(\ell_1) + T(\ell_2) = T(\ell_1 + \ell_2)$. Note that this also implies that $T(0) = 0$, since any latency function ℓ_1 can be written as $\ell_1 + 0$.

Next we show that $T(\ell)$ is constant when ℓ is constant. Consider again the network in Figure 6.4(a) when $\ell_1(f_1) = 0$, $\ell_2(f_2) = b_2$, and $\ell_3(f_3) = b_3$. It is clear that if $b_2 < b_3$, the unique Nash flow routes all traffic on the upper path, and this flow is also optimal. Writing $T(b_2)(\cdot)$ as the tolling function assigned to $\ell_2(f_2) = b_2$ by T, it follows that for all f, $T(b_2)(f) < T(b_3)(0)$. If this were not the case, there would exist perverse Nash flows for large r which route a positive mass of traffic on the lower path. Since this must hold for all b_2 and b_3, it implies that for all b, $T(b)(\cdot)$ is a nonincreasing function of flow.

By an opposite argument, it must be that for all f, $T(b_2)(0) < T(b_3)(f)$, implying that $T(b)(\cdot)$ must be a nondecreasing (and thus constant) function of flow. Because T is simply a mapping from $\mathbb{R}$ to $\mathbb{R}$ for constant functions, its additivity implies linearity:[3] $T(b) = \kappa_1 b$. Finally, for $b_2 < b_3$, it must always be true for any possible agent sensitivities $s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]$ that $(1 + \kappa_1 s)b_2 < (1 + \kappa_1 s)b_3$, or that $\kappa_1 > -1/S_{\mathrm{U}}$.

Next, we show that degree-d monomial latency functions must be assigned degree-d tolling functions. The network in Figure 6.4(b) has two edges in parallel with latency functions $\ell_1(f_1) = \alpha(f_1)^d$ and $\ell_2(f_2) = \lambda\alpha(f_2)^d$, where $\alpha > 0$, $\lambda > 0$, and $d \geq 1$. For any such network, the unique un-influenced Nash flow is optimal; thus, a non-perverse T would need to induce a flow f^T that satisfies $f^T = f^{\mathrm{nf}} = f^{\mathrm{opt}}$. It can be shown that for any $r > 0$, this flow is

$$f_1^T = \frac{(\lambda\alpha)^{1/d} r}{(\alpha)^{1/d} + (\lambda\alpha)^{1/d}}, \quad f_2^T = \frac{(\alpha)^{1/d} r}{(\alpha)^{1/d} + (\lambda\alpha)^{1/d}}.$$

Since f^T is a nominal Nash flow, $\ell_1(f_1^T) = \ell_2(f_2^T)$; and $f^T = f^T$ implies that $\tau_1(f_1^T) = \tau_2(f_2^T)$. In the following, let $r = (\alpha)^{1/d} + (\lambda\alpha)^{1/d}$, so for all α, λ, $\tau_1\left((\lambda\alpha)^{1/d}\right) = \tau_2\left((\alpha)^{1/d}\right)$.

First let $\lambda = 2$, so that $\ell_2(f_2) = 2\ell_1(f_2)$. Then additivity ensures that $\tau_2(f_2) = 2\tau_1(f_2)$. That is, $\tau_1\left((2\alpha)^{1/d}\right) = 2\tau_1\left((\alpha)^{1/d}\right)$. Since this must hold for any α, it implies either that $\tau_1(f) \equiv 0$, or that $\tau_1(f) = \eta_1 f^d$ for some $\eta_1 > 0$.

To find η, we need only substitute f^T into $\tau_1(f_1^T) = \tau_2(f_2^T)$ and solve, yielding $\eta_1\lambda = \eta_2$. Due to the fact that η_1 cannot be a function of λ, the above is only satisfied when all tolling functions are given by $K\alpha_e(f_e)^d$ for some $K \geq 0$. K is a constant that does not depend on e (but may depend on d).

To find K, consider Figure 6.4(c). This network has $\ell_1(f_1) = \alpha(f_1)^d$ in parallel with a constant latency function $\ell_2(f_2) = 1$. Here, if $r \leq (1/(\alpha(d+1)))^{1/d}$, the uninfluenced Nash and optimal flow on this network is $(r, 0)$. Thus, $\tau_1(f)$ must be small enough that it does not incentivize *any* user to use edge 2 when r is low. Precisely, keeping in mind that $\tau_1(f_1) = K\alpha(f_1)^d$, we require that for all sensitivities $s \in [S_{\mathrm{L}}, S_{\mathrm{U}}]$, $\alpha(f_1)^d + sK\alpha(f_1)^d \leq$

[3] Of course, this is provided that we require T to be Lebesgue-measurable.

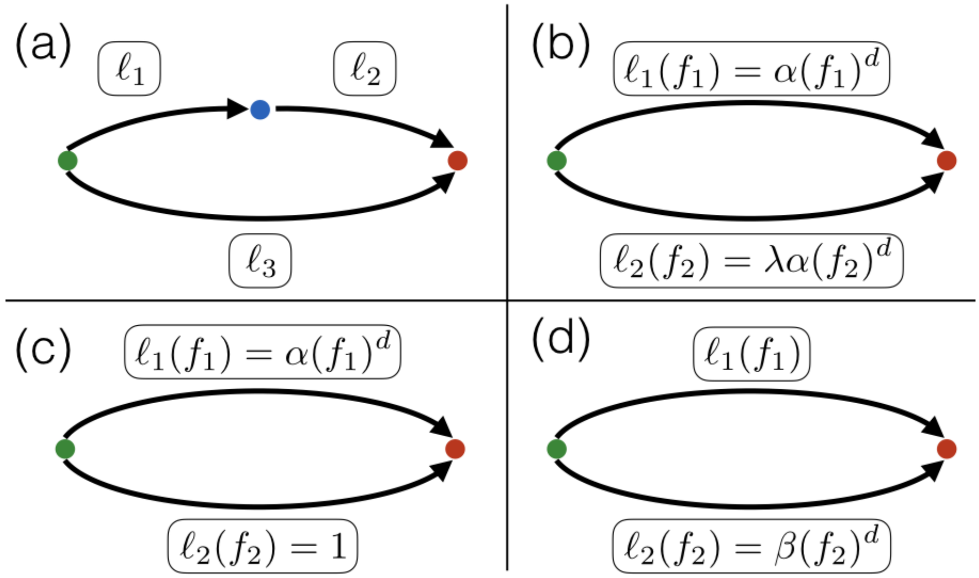

Figure 6.4: Example networks used to prove Lemma 6.1.1.

$1 + s\kappa_1$, or, substituting the appropriate f_1, that

$$\alpha(1 + sK)\left(\left(\frac{1}{\alpha(d+1)}\right)^{1/d}\right)^d \leq 1 + s\kappa_1. \tag{6.24}$$

This implies that $sK \leq s\kappa_1 d + s\kappa_1 + d$. This simplifies nicely if we write $K = \kappa_1 + \kappa_2 d$ (where $\kappa_2 \in \mathbb{R}$), in which case it follows that $\kappa_2 \leq \kappa_1 + 1/s$ for all d and s, or that $\kappa_2 \leq \kappa_1 + 1/S_{\mathrm{U}}$. Writing $\tau_1(f_1)$ in terms of κ_1 and κ_2 gives the nice decomposition in terms of latency function ℓ_1 and marginal-cost function $f_1 \cdot \ell_1'$:

$$\tau_1(f_1) = \kappa_1\ell_1(f_1) + \kappa_2 f_1 \cdot \ell_1'(f_1).$$

Finally, consider the network in Figure 6.4(d). This network has some arbitrary admissible latency function ℓ_1 on edge 1 and a monomial latency function $\ell_2(f_2) = \beta(f_2)^d$ on edge 2. We will choose ℓ_2 such that the optimal and Nash flows coincide on this network for some $r > 1$ when $f_2 = 1$. Due to the additivity of T, we can assume without loss of generality that $\ell_1(0) = 0$ because any nonzero intercept can be "canceled" by adding an equal constant term to ℓ_2.

Let $\beta = \ell_1(f_1)$ and $d = f_1\ell_1'(f_1)/\ell_1(f_1) \geq 1$. Then $\ell_1(f_1) = \ell_2(1)$ and $f_1\ell_1'(f_1) = \ell_2'(1)$; i.e., both the latencies and the marginal costs of the edges are equal, which means that $(f_1, 1)$ is both a Nash and an optimal flow. Since ℓ_2 is a monomial, we can write its tolling function as $\tau_2(f_2) = \kappa_1\beta(f_2)^d + \kappa_2 d\beta(f_2)^d$, where $\kappa_2 \leq \kappa_1 + 1/S_{\mathrm{U}}$. Using this, we can simply derive the first-link tolling function $T(\ell_1)(f_1)$ using the following:

$$\ell_1(f_1) + T(\ell_1)(f_1) = (\beta + \kappa_1\beta + \kappa_2 d\beta)\,(f_2)^d.$$

Substituting the definitions of β and d and canceling similar terms, we obtain that $T(\ell_1)$ satisfies (6.7) as desired. ■

Next, Lemma 6.7.1 shows that Nash flows on parallel networks behave very nicely under the influence of T^{gmc}. Specifically, Lemma 6.7.1 proves that the worst-case total latency on a parallel network with T^{gmc} is realized by a low-sensitivity homogeneous population.

Lemma 6.7.1 *Let s^{L} denote a homogeneous population in which every user has sensitivity $S_{\mathrm{L}} \geq 0$, and denote by T^{gmc} a taxation mechanism satisfying the conditions of Lemma 6.1.1. For any symmetric parallel network $G \in \mathcal{G}^{\mathrm{sp}}$ and any heterogeneous population s in which every user has a sensitivity no less than S_{L},*

$$\mathcal{L}^{\mathrm{nf}}\left(G, s^{\mathrm{L}}, T^{\mathrm{gmc}}\right) \geq \mathcal{L}^{\mathrm{nf}}\left(G, s, T^{\mathrm{gmc}}\right). \tag{6.25}$$

Proof: For every user x, T^{gmc} induces cost functions of the form

$$J_e^x(f_e) = (1 + s_x\kappa_1)\ell_e(f_e) + s_x\kappa_2 f_e\ell_e'(f_e). \tag{6.26}$$

Since we can scale these costs functions by any user-specific positive scalar without changing the underlying Nash flows, these cost functions are equivalent to the following:

$$J_e^x(f_e) = \ell_e(f_e) + \frac{s_x\kappa_2}{1 + s_x\kappa_1} f_e\ell_e'(f_e). \tag{6.27}$$

Given the conditions $\kappa_2 \geq 0$ and $\kappa_1 \geq \kappa_2 - 1/S_{\mathrm{U}}$, the expression $\frac{s_x \kappa_2}{1+s_x\kappa_1} \in [0,1]$ and is monotone increasing in s_x. Thus, analysis can be simplified by assuming that $\kappa_1 = 0$, $\kappa_2 = 1$ and cost functions are simply given by

$$J_e^x(f_e) = \ell_e(f_e) + s_x f_e \ell_e'(f_e), \tag{6.28}$$

where $s_x \in [0,1]$ for all x.

For convenience, we write $\ell_e^*(f_e) \triangleq f_e \ell_e'(f_e)$. When describing the cost experienced by a particular agent whose sensitivity is $s \in \mathbb{R}_+$, we write

$$\ell_e^s(f_e) \triangleq \ell_e(f_e) + s\ell_e^*(f_e), \tag{6.29}$$

and we write $\ell_e^{\mathrm{mc}}(f_e) \triangleq \ell_e^1(f_e)$ to denote the marginal-cost function associated with edge e. The following proposition gives important information about the structure of Nash flows induced by T^{gmc}.

Proposition 6.8 *If f^{nf} is a Nash flow on $G \in \mathcal{G}^{\mathrm{sp}}$ for population s under the influence of T^{gmc}, the following facts hold for any two paths satisfying $\ell_i(0) \leq \ell_j(0)$, $f_j^{\mathrm{nf}} > 0$, and where a user x is on p_i and user y on p_j:*

1. $\ell_i^{\mathrm{mc}}(f_i^{\mathrm{nf}}) \geq \ell_j^{\mathrm{mc}}(f_j^{\mathrm{nf}})$
2. $s_x \leq s_y$.

When $\ell_i(0) < \ell_j(0)$, inequality (1) is strict.

Proof: Order the paths so that $\ell_i(0) \leq \ell_{i+1}(0)$ for all $i < n$, and take two paths p_i and p_{i+1} such that $f_i^{\mathrm{nf}} > 0$ and $f_{i+1}^{\mathrm{nf}} > 0$. Because this is a Nash flow, any agent y using path p_{i+1} experiences a (weakly) lower cost than he would on path p_i, or

$$\ell_i(f_i) + s_y\ell_i^*(f_i) \geq \ell_{i+1}(f_{i+1}) + s_y\ell_{i+1}^*(f_{i+1}). \tag{6.30}$$

Any latency function can be uniquely decomposed into its 0-flow latency and its flow-varying part in the following way:

$$\ell(f) = \tilde{\ell}(f) + \ell(0). \tag{6.31}$$

It is always true that $f\tilde{\ell}_i'(f) = f\ell_i'(f)$, so (6.30) and $\ell_{i+1}(0) - \ell_i(0) \geq 0$ imply that

$$s_y \left(\ell_i^*(f_i) - \ell_{i+1}^*(f_{i+1})\right) \geq \tilde{\ell}_{i+1}(f_{i+1}) - \tilde{\ell}_i(f_i). \tag{6.32}$$

In the same Nash flow, consider some user x using path p_i. For this user, a similar argument shows that

$$s_x \left(\ell_i^*(f_i) - \ell_{i+1}^*(f_{i+1})\right) \leq \tilde{\ell}_{i+1}(f_{i+1}) - \tilde{\ell}_i(f_i). \tag{6.33}$$

Combining (6.32) and (6.33) yields

$$0 \leq (s_y - s_x) \left(\ell_i^*(f_i) - \ell_{i+1}^*(f_{i+1})\right), \tag{6.34}$$

meaning that $s_y \geq s_x$ implies that $\ell_i^*(f_i) \geq \ell_{i+1}^*(f_{i+1})$. That is, higher-sensitivity agents use higher-index paths (paths with higher zero-flow latencies), proving item (2).

This means that for each pair of paths, if we define s_i as the number satisfying

$$\ell_i(f_i) + s_i\ell_i^*(f_i) = \ell_{i+1}(f_{i+1}) + s_i\ell_{i+1}^*(f_{i+1}), \tag{6.35}$$

it will be the case that each $s_i \leq s_{i+1}$ and that $s_i \leq 1$. Finally, it follows from $\ell_i^*(f_i) \geq \ell_{i+1}^*(f_{i+1})$ and (6.35) that for any $s_i < 1$, we have $\ell_i(f_i) + \ell_i^*(f_i) > \ell_{i+1}(f_{i+1}) + \ell_{i+1}^*(f_{i+1})$, proving item (1). ■

The basic proof approach is to exploit this ordering of marginal costs, and show that reducing agents' sensitivities (thereby making the population "more homogeneous") shifts agents from low marginal-cost paths to high marginal-cost paths, increasing the total latency. Formally, we define a mapping $\Sigma : [0,1] \times \mathcal{S} \to \mathcal{S}$. For any starting population s^0 and any α, we will define $\Sigma\left(\alpha; s^0\right)$ as a right-shift of s^0 by α units. The sensitivity of user x in population $\Sigma(\alpha, s^0)$ is given by

$$\Sigma(\alpha, s^0)_x = \begin{cases} s_0(0) & \text{if } x \leq \alpha \\ s_0(x-\alpha) & \text{if } x > \alpha. \end{cases} \tag{6.36}$$

Because s is defined to be an increasing function, this is equivalent to converting a mass of α of the most-sensitive users to a mass α of the least-sensitive users.

Proposition 6.8 allows us to assume without loss of generality that any user population s has a finite number of sensitivity types; to see this, simply note that if users with distinct sensitivities are using the same path in a Nash flow, one sensitivity may be exchanged for the other without perturbing either agent's preferences. To be precise, given a Nash flow f^{nf}, we will assume for each path $p_i \in \mathcal{P} \setminus p_1$, each user has the *minimally-indifferent* sensitivity which satisfies (6.35).

For notational brevity, we will typically write $f^{\mathrm{nf}}(\alpha)$ to represent $f^{\mathrm{nf}}(\Sigma(\alpha; s_0))$. Our central goal will be to characterize the effect of marginal increases in α on the Nash flow. We express this marginal effect as $\frac{\partial}{\partial\alpha} f^{\mathrm{nf}}(\alpha)$.

The following definition will be helpful in the proof:

Definition 6.1 *In a Nash flow f^{nf}, paths p_i and p_j with $i < j$ are said to be* strategically coupled *if s_i satisfies $\ell_i^{s_i}(f_i^{\mathrm{nf}}) = \ell_j^{s_i}(f_j^{\mathrm{nf}})$. That is, agents on the lower-order path are indifferent between the two paths. We write $\mathcal{P}_i(f^{\mathrm{nf}})$ to denote the set of paths that are strategically coupled to path p_i in f^{nf}.*[4]

First, we show that the primary effect of an increase in α is to shift traffic from $\mathcal{P}_n$ to $\mathcal{P}_1$.

Proposition 6.9 *For every path $p_i \in \mathcal{P}_1$, $\frac{\partial}{\partial\alpha} f_i^{\mathrm{nf}}(\alpha) \geq 0$. For every path $p_j \in \mathcal{P}_n$, $\frac{\partial}{\partial\alpha} f_j^{\mathrm{nf}}(\alpha) \leq 0$.*

[4]When clear from context, we write $\mathcal{P}_i(f^{\mathrm{nf}})$ simply as $\mathcal{P}_i$.

Proof: Let s_1 denote the sensitivity of agents using p_1 in f^{nf}. Increasing α changes the sensitivity of a small fraction of high-sensitivity users to s_1. By Definition 6.1 and Proposition 6.8, these users strictly prefer the paths in $\mathcal{P}_1$ to any other paths, so a marginal increase in α induces a marginal increase in flow on $\mathcal{P}_1$. That is, at least one path in $p_i \in \mathcal{P}_1$ has $\frac{\partial}{\partial \alpha} f_i^{\mathrm{nf}}(\alpha) > 0$. An implication of Proposition 6.8 is that all paths in $\mathcal{P}_1$ have strictly flow-varying cost functions, so an increase on flow on p_i induces an increase in flow on all paths in $\mathcal{P}_1$, proving the first statement.

Next, let s_n denote the sensitivity of agents using p_n in f^{nf}; Definition 6.1 and Proposition 6.8 shows that these agents weakly prefer $\mathcal{P}_n$. Increasing α shifts some of these users to $\mathcal{P}_1$, so at least one path in $p_i \in \mathcal{P}_n$ has $\frac{\partial}{\partial \alpha} f_i^{\mathrm{nf}}(\alpha) < 0$. If $\mathcal{P}_n$ contains a path with a constant latency function, then this is the path which the flow leaves; otherwise, the flow would deviate to a non-Nash flow. On the other hand, if all paths in $\mathcal{P}_n$ are strictly flow-varying, then every path flow in $\mathcal{P}_n$ must decrease, proving the second statement. ■

Proposition 6.10 *For any α, if $p_j \notin \mathcal{P}_1(\alpha)$ and $p_j \notin \mathcal{P}_n(\alpha)$, it holds that $\frac{\partial}{\partial \alpha} f_j^{\mathrm{nf}}(\alpha) = 0$.*

Proof: First, let p_i be the lowest-index path such that $p_i \notin \mathcal{P}_1$ (that is, $p_{i-1} \in \mathcal{P}_1$). Definition 6.1 means that for any $p_j \in \mathcal{P}_1$, $\ell_j^{s_j}(f_j) < \ell_i^{s_j}(f_i)$. Since the inequality is strict, the fact from Proposition 6.9 that $\frac{\partial}{\partial \alpha} f_j^{\mathrm{nf}}(\alpha) \geq 0$ means that marginally no agent on $\mathcal{P}_1$ will switch to p_i.

However, since $f^{\mathrm{nf}}(\alpha)$ is a Nash flow, it is true that $\ell_j^{s_i}(f_j) \geq \ell_i^{s_i}(f_i)$. Here, $\frac{\partial}{\partial \alpha} f_j^{\mathrm{nf}}(\alpha) \geq 0$ implies that $\ell_j^{s_i}(f_j)$ can only increase, so no agent on p_i will be incentivized to switch to any path in $\mathcal{P}_1$. Thus, the flow on p_i is not influenced by the changes in flow on any lower-index path; if its flow changes, the influence must come from a higher-index path.

Now, let p_i be the highest-index path such that $p_i \notin \mathcal{P}_n$ (that is, $p_{i+1} \in \mathcal{P}_n$). Definition 6.1 means that for any $p_j \in \mathcal{P}_n$, $\ell_i^{s_i}(f_i) < \ell_j^{s_i}(f_j)$. Since the inequality is strict, the fact that $\frac{\partial}{\partial \alpha} f_j^{\mathrm{nf}}(\alpha) \leq 0$ means that (marginally) no agent on p_i will be incentivized to switch to any path in $\mathcal{P}_n$. However, since $f^{\mathrm{nf}}(\alpha)$ is a Nash flow, it is true that $\ell_j^{s_j}(f_j) \leq \ell_i^{s_j}(f_i)$. Here,

$\frac{\partial}{\partial\alpha} f_j^{\mathrm{nf}}(\alpha) \leq 0$ implies that $\ell_j^{s_j}(f_j)$ can only decrease, so no agent on any path in $\mathcal{P}_n$ will be incentivized to switch to p_i. Thus, the flow on p_i is not influenced by the changes in flow on any higher-index path.

This argument may then be repeated with all remaining paths that are not in $\mathcal{P}_1$ or $\mathcal{P}_n$ to show that the only path flows that may change in response to α are those in $\mathcal{P}_1$ and $\mathcal{P}_n$, obtaining the proof of the proposition. ■

Proof of Lemma 6.7.1: We can now quantify the effect of an increase in α on total latency. In the following, $\nabla_f L(f)$ represents the gradient vector of L with respect to flow f given by $\{\ell_p^{\mathrm{mc}}\}_{p\in\mathcal{P}}$, which by Proposition 6.8 is ordered descending. Let p_j be the highest-index path in $\mathcal{P}_1$, and p_k be the lowest-index path in $\mathcal{P}_n$:

$$\begin{aligned}
\frac{\partial}{\partial\alpha} L\left(f^{\mathrm{nf}}(\alpha)\right) &= \nabla_f L\left(f^{\mathrm{nf}}(\alpha)\right) \cdot \frac{\partial}{\partial\alpha} f^{\mathrm{nf}}(\alpha) \\
&= \sum_{i\in\mathcal{P}_i\cup\mathcal{P}_n} \ell_i^{\mathrm{mc}}\left(f_i^{\mathrm{nf}}(\alpha)\right) \frac{\partial}{\partial\alpha} f_i^{\mathrm{nf}}(\alpha) \\
&\geq \left[\ell_j^{\mathrm{mc}}\left(f_j^{\mathrm{nf}}(\alpha)\right) - \ell_k^{\mathrm{mc}}\left(f_k^{\mathrm{nf}}(\alpha)\right)\right] \geq 0.
\end{aligned}$$

Since at every Nash flow $f^{\mathrm{nf}}(\alpha)$ it is true that $\frac{\partial}{\partial\alpha} L\left(f^{\mathrm{nf}}(\alpha)\right) \geq 0$, the definition of $\Sigma(\alpha, s_0)$ implies that for any initial sensitivity distribution s_0,

$$\mathcal{L}\left(f^{\mathrm{nf}}\left(\Sigma\left(1, s_0\right)\right)\right) \geq \mathcal{L}\left(f^{\mathrm{nf}}\left(\Sigma\left(0, s_0\right)\right)\right), \tag{6.37}$$

or that $\mathcal{L}^{\mathrm{nf}}\left(G, s^{\mathrm{L}}, T^{\mathrm{gmc}}\right) \geq \mathcal{L}^{\mathrm{nf}}\left(G, s, T^{\mathrm{gmc}}\right).$ ■

Proof of Theorem 6.2

Let $G \in \mathcal{G}_p$ be a parallel network, s be any arbitrary sensitivity distribution, s^{L} be a homogeneous population in which all users have sensitivity S_{L}, and let taxation mechanism T^{gmc} satisfy (6.7). Lemma 6.7.1 ensures that

$$\mathcal{L}^{\rm nf}\left(G, s^{\rm L}, T^{\rm gmc}\right) \geq \mathcal{L}^{\rm nf}\left(G, s, T^{\rm gmc}\right). \tag{6.38}$$

Let s^0 denote a totally-insensitive homogeneous population; that is, all agents have sensitivity 0. Note that s^0 is itself a low-sensitivity homogeneous population and that s is a population in which all users have sensitivity no less than 0; thus, we may simply apply Lemma 6.7.1 a second time to obtain

$$\mathcal{L}^{\rm nf}\left(G, s^0, T^{\rm gmc}\right) \geq \mathcal{L}^{\rm nf}\left(G, s^{\rm L}, T^{\rm gmc}\right). \tag{6.39}$$

The left-hand side of (6.39) is simply the un-tolled total latency on G, so combining inequalitites (6.38) and (6.39), we obtain

$$\mathcal{L}^{\rm nf}(G, \emptyset) \geq \mathcal{L}^{\rm nf}\left(G, s, T^{\rm gmc}\right). \tag{6.40}$$

Since G and s were arbitrary, this implies that $T^{\rm gmc}$ has perversity index of 1 on $\mathcal{G}_p$. ■

Proof of Theorem 6.3: Let $\beta(s, (\kappa_1, \kappa_2)) := \frac{\kappa_2 s}{1+\kappa_1 s}$. The constraints on κ_1, κ_2 specified in Theorem 6.2 imply that for all $s \in [S_{\rm L}, S_{\rm U}]$, $\beta(s, (\kappa_1, \kappa_2)) \in [0, 1]$. Lemma 6.7.1 implies that on any $G \in \mathcal{G}^{\rm sp}$, worst-case routing is achieved by a homogeneous population with $s = S_{\rm L}$. Thus, the price of anarchy for heterogeneous populations is equal to that given by Lemma 6.4.1 for $\beta \leq 1$, proving the theorem. ■

Proof of Lemma 6.5.1: Let $\beta(s) := \frac{\kappa_2 s}{1+\kappa_1 s}$. It is a consequence of Lemma 6.7.1 that the perversity index can never be greater than 1 *due to a population with* $\beta(s) \leq 1$. When $\beta(s) > 1$, note that it is increasing in s whenever $\kappa_2 > 0$ and $\kappa_1 > -1/S_{\rm U}$; this implies that the price of anarchy *due to a population with* $\beta(s) > 1$ is achieved by one with $s = S_{\rm U}$. The lemma is proved if it can be shown that there is a perverse flow for this population whose perversity equals the corresponding price of anarchy. Accordingly, let $\beta := \frac{\kappa_2 S_{\rm U}}{1+\kappa_1 S_{\rm U}} > 1$. Consider the network in Figure 6.4(c) with $r = 1$ and $\alpha = \frac{(\beta(1+d))^d}{(1+d\beta)^{d+1}}$; this network is borrowed from [80]. The optimal and uninfluenced Nash flow on this network has $f_1 = 1$, but the tolled Nash flow for a population with sensitivity S_{U} has a vastly increased total latency which achieves the $\beta > 1$ PoA bound given in (6.12). Thus, the Lemma is proved. ■

Chapter 7

Discriminatory Pricing

As Chapters 2–6 have made clear, uninfluenced social systems can exhibit quite poor behavior, and incentive mechanisms to mitigate this are often nontrivial. Every pricing mechanism considered thus far has charged every agent on each edge the same price. It is almost trivial to see that a tax-designer with the ability to charge each agent an individualized price could apply an individualized marginal-cost pricing scheme and and inherit the optimality given by [42] while avoiding the unbounded tolls of Chapter 5 and the perversity of Chapter 6. Something like this individualized pricing approach is discussed in [32]. The goal of this chapter is to investigate the performance of discriminatory pricing schemes along these lines.

It is well-known that a monopolist can maximize profits with "first-degree" (also called "perfect") price-discrimination, in which prices are individualized for every customer [82], provided that goods cannot be re-sold. Another commonly-studied form of price-discrimination is often termed "second-degree" price discrimination or nonlinear pricing; here, different prices are charged for different *quantities* of a good or service [83], essentially inducing customers to partition themselves into different sensitivity classes. If a seller lacks direct access to customers' sensitivities, it may still be possible to indirectly disaggregate the customer population: senior, student, and corporate discounts are common means to this end [82].

This chapter is adapted, with permission, from [81], which was previously published in the proceedings of the 2016 American Control Conference.

Even in cases such as public utility pricing in which certain kinds of discriminatory pricing are prohibited by law, price-discrimination via volume discounts (known as "nonlinear" pricing) is a common practice [84]. Price discrimination has also been studied in the context of cloud computing [85] and the provision of network services [86].

In this chapter, we study a coarse version of first-degree discrimination, under the assumption that perfect price-discrimination is likely to be impractical in a real-world setting. Here, we partition the users into a small number of groups or "bins" on the basis of their price-sensitivity and charge a single price to members of each bin. It seems natural to assume that even coarse price-discrimination could improve the efficiency guarantees, since each agent would be charged a price that is close to the "correct" price for his particular price-sensitivity. We assume that the designer has no distributional information about users *within* each bin (similar to an approach outlined in [87]), but that each user is correctly categorized, however coarsely.

First, Theorem 7.1 demonstrates that network flows can be made arbitrarily close to optimal flows by binning users sufficiently finely. This validates the intuitive concept that if we charge each user an individualized price, we can enforce any network flow.

However, Theorem 7.1 gives no hint as to how many bins are required to achieve a particular efficiency target; accordingly, Theorem 7.2 shows a fundamental equivalence between discriminatory pricing and the tax-designer's uncertainty regarding the user population's price-sensitivity. Here, we show that discriminatory pricing on a poorly-characterized population is essentially identical to simple pricing on a well-characterized population.

Finally, Theorem 7.3 considers the class of affine-cost, parallel-network routing games and provides a methodology for deriving the optimal binning and bin taxes for any number of bins. We also prove bounds on the inefficiencies resulting from this congestion-minimizing binning.

7.0.1 Price Discrimination Model

Our proposed price-discrimination model is comprised of two components: a collection of bins represented by a partition of the interval $[S_{\mathrm{L}}, S_{\mathrm{U}}]$ into m sub-intervals, and a collection of taxation functions for each group.

The collection of bin boundaries is written $\{\beta_i\}_{i=0}^m$, $\beta_0 = S_{\mathrm{L}}$, $\beta_m = S_{\mathrm{U}}$, with $\beta_i < \beta_{i+1}\ \forall\ i < m$. For each edge e, all users in bin i (that is, all users whose sensitivities lie in the interval $[\beta_{i-1}, \beta_i]$) are charged a taxation function $\tau_e^i(f_e)$, yielding for each edge e a collection of m distinct taxation functions $\{\tau_e^i\}_{i=1}^m$. To make general statements about binnings as functions of m, we write B to represent a function that maps each $m \in \mathbb{N}$ to a particular partition $\{\beta_i\}_{i=0}^m$ and taxation functions $\left\{\{\tau_e^i\}_{e\in E}\right\}_{i=1}^m$, and write B^m to denote a binning for a specific value of m. Occasionally, we use the notation B^1 to denote a trivial "binning" in which all users are charged the same price.

The social behavior resulting from an m-binning is modeled as a Nash flow, or a flow f in which for all users x, where x belongs to commodity $c \in \mathcal{C}$ and bin i, we have

$$J_x(f) = \min_{p\in\mathcal{P}_c} \left\{ \sum_{e\in p} \left[\ell_e(f_e) + s_x^c \tau_e^i(f_e)\right] \right\}. \tag{7.1}$$

It is well-known that a Nash flow exists for any non-atomic game of the above form [54].

For a given routing problem $G \in \mathcal{G}$, we gauge the efficacy of a binning B^m by comparing the total latency of the resulting Nash flow and the total latency associated with the optimal flow, and then performing a worst-case analysis over all possible sensitivity distributions. Let $\mathcal{L}^*(G)$ denote the total latency associated with the optimal flow, and $\mathcal{L}^{\mathrm{nf}}(G, s, B^m)$ denote the total latency of the Nash flow resulting from binning B^m and user population s. The worst-case efficiency loss associated with this specific instance is captured by the *price of anarchy* which takes the general form

$$\mathrm{PoA}(G, S_{\mathrm{L}}, S_{\mathrm{U}}, B^m) = \sup_{s\in\mathcal{S}_G} \left\{ \frac{\mathcal{L}^{\mathrm{nf}}(G, s, B^m)}{\mathcal{L}^*(G)} \right\} \geq 1. \tag{7.2}$$

7.1 Universal Discriminatory Pricing

As discussed in the previous section, the state of the art gives us no obvious way to enforce optimal flows without either significant quantities of information or excessively-high tolls. In this paper, we show that price-discrimination may provide a third way. We begin with our most general result, in which we present a family of binnings which enforce arbitrarily-optimal flows for any routing problem. This optimality is asymptotic in the number of bins, which implies that it is always possible to enforce Nash flows within ϵ of optimal with a finite number of bins.

Theorem 7.1 *Define $\mathcal{B}$ as the family of binnings whose bin sizes shrink to 0 as m approaches infinity; for any $B^m \in \mathcal{B}$,*[1]

$$\lim_{m\to\infty} \left(\beta_i^m - \beta_{i-1}^m\right) = 0, \tag{7.3}$$

and the bin taxation functions of $B^m \in \mathcal{B}$ satisfy the following: for each bin i choose any $\kappa_i^m \in \left[\frac{1}{\beta_i^m}, \frac{1}{\beta_{i-1}^m}\right]$ and let edge tolls be given by

$$\tau_e^i\left(f_e\right) = \kappa_i^m f_e \cdot \frac{d}{df_e}\ell_e(f_e). \tag{7.4}$$

Then for any $G \in \mathcal{G}$ and any binning $B^m \in \mathcal{B}$,

$$\lim_{m\to\infty} \mathrm{PoA}\left(G, S_{\mathrm{L}}, S_{\mathrm{U}}, B^m\right) = 1 \tag{7.5}$$

and for all m, all tolling functions are bounded by $\max_e \ell_e'(1)/S_{\mathrm{L}}$.

Note that B^m need not depend on information about user sensitivity distributions or demands, and may be completely topology-independent.

[1] For the sake of precision, in this theorem we make explicit the dependence of each bin boundary on m by writing β_i^m.

Proof: For each $m \in \mathbb{N}$, let binning B^m be defined with bin boundaries according to (7.3) and bin tolls according to (7.4), so that $B^m \in \mathcal{B}$. For any routing problem $G \in \mathcal{G}$ and price-sensitivities $s \in \mathcal{S}$, let $f^m = \left(f_p^m\right)_{p \in \mathcal{P}}$ denote the Nash flow resulting from the binning B^m. For each commodity c, let $\mathcal{P}_c^m \subseteq \mathcal{P}_c$ denote the set of paths that have positive flow in f^m. For any $p \in \mathcal{P}_c^m$, there must be some user $x \in [0, r_c]$ using p; suppose this user has sensitivity $s_x^c \in \left[\beta_{i-1}^m, \beta_i^m\right]$, then the cost experienced by this user is given by

$$J_x\left(f^m\right) = \sum_{e \in p} \left[\ell_e(f_e) + s_x^c \kappa_i^m f_e \cdot \frac{d}{df_e} \ell_e(f_e)\right].$$

Write $\ell_e^*(f_e) = f_e \cdot \frac{d}{df_e} \ell_e(f_e)$; then for any other path $p' \in \mathcal{P}_c \setminus p$, user x must experience a lower cost on p than on p', or

$$\sum_{e \in p} \ell_e(f_e) - \sum_{e \in p'} \ell_e(f_e) \leq s_x^c \kappa_i^m \left[\sum_{e \in p'} \ell_e^*(f_e) - \sum_{e \in p} \ell_e^*(f_e)\right]. \tag{7.6}$$

Therefore, for any $m \geq 1$, f^m must satisfy some set of inequalities defined by (7.6). By definition,

$$\beta_{i-1}^m / \beta_i^m \leq s_x^c \kappa_i^m \leq \beta_i^m / \beta_{i-1}^m,$$

so (7.3) implies that $\lim_{m \to \infty} s_x^c \kappa_i^m = 1$. Thus, because all functions in (7.6) are continuous, f^m converges to a set F^* of feasible flows that satisfy

$$\sum_{e \in p} \ell_e(f_e) - \sum_{e \in p'} \ell_e(f_e) \leq \left[\sum_{e \in p'} \ell_e^*(f_e) - \sum_{e \in p} \ell_e^*(f_e)\right] \tag{7.7}$$

for all c, all $p \in \mathcal{P}_c^*$, and $p' \in \mathcal{P}_c$, where $\mathcal{P}_c^* \subseteq \mathcal{P}_c$ is some subset of paths. As in the proof of Theorem 5.1 in Chapter 5, any flow satisfying (7.7) must be optimal, which along with the continuity of $\mathcal{L}$ concludes the proof. ∎

7.2 General Effect of Discriminatory Pricing

Theorem 7.1 showed that we can enforce low-congestion routing using price discrimination, but it gave no hint as to how many bins are needed or how the price of anarchy evolves as a function of m. This is the purview of Theorem 7.2, in which we show that there is a general equivalence between fine discrimination and well-characterized sensitivity distributions. Here, we show that the price of anarchy resulting from discriminatory pricing for a poorly-characterized population is no worse than the price of anarchy resulting from non-discriminatory pricing for a well-characterized population.

Theorem 7.2 *Suppose for routing problem G that some taxation mechanism $T(S_{\mathrm{L}}, S_{\mathrm{U}})$ is known to have price of anarchy $\mathrm{PoA}\,(G, S_{\mathrm{L}}, S_{\mathrm{U}})$. For any $S'_{\mathrm{L}} > 0$, let $S'_{\mathrm{U}} = S'_{\mathrm{L}}\,(S_{\mathrm{U}}/S_{\mathrm{L}})^{1/m}$, and define the bin boundaries of B^m by*

$$\beta_i = S_{\mathrm{L}}^{\frac{m-i}{m}} S_{\mathrm{U}}^{\frac{i}{m}}. \tag{7.8}$$

Then if the bin taxes of B^m are given by $T_i = S'_{\mathrm{L}}/\beta_{i-1} \cdot T(S'_{\mathrm{L}}, S'_{\mathrm{U}})$, the following holds:

$$\mathrm{PoA}\,(G, S_{\mathrm{L}}, S_{\mathrm{U}}, B^m) \leq \mathrm{PoA}\left(G, S'_{\mathrm{L}}, S'_{\mathrm{L}}\left(\frac{S_{\mathrm{U}}}{S_{\mathrm{L}}}\right)^{1/m}, B^1\right). \tag{7.9}$$

In particular, we wish to point out two important facts regarding Theorem 7.2. First, it is natural to evaluate the uncertainty of our user-sensitivity estimate by the sensitivity ratio $S_{\mathrm{U}}/S_{\mathrm{L}}$: the higher the ratio, the less certainty we possess. Theorem 7.2 shows us that m-binning reduces the effective sensitivity ratio to $(S_{\mathrm{U}}/S_{\mathrm{L}})^{1/m}$. Thus, by applying price discrimination, we can dramatically reduce our uncertainty regarding the price-sensitivity of network users.

Second, note that the guarantees of Theorem 7.2 are independent of the specific taxation mechanism used; thus, this result can be used as a design tool to take any off-the-shelf taxation mechanism and apply it to a discriminatory setting.

However, it is important to understand that the price of anarchy provided by Theorem 7.2 need not be *optimal* in any sense. If a binning is designed more precisely with a particular taxation methodology in mind, it may be possible to guarantee even better network efficiencies. This is the focus of Theorem 7.3, in which we look at a restricted class of routing problems and taxation mechanisms and derive the optimal binning in that specific setting.

Proof: Consider routing problem G and population s; design bin boundaries and charge taxes according to the theorem statement. Note that for every i, (7.8) implies that

$$\frac{\beta_i}{\beta_{i-1}} = \frac{S'_{\mathrm{U}}}{S'_{\mathrm{L}}} = \left(\frac{S_{\mathrm{U}}}{S_{\mathrm{U}}}\right)^{1/m}; \tag{7.10}$$

that is, each bin has the same "width" as the entire emulated population $s' \in [S'_{\mathrm{L}}, S'_{\mathrm{U}}]$. Consider the cost experienced on any edge e by some agent x who happens to belong to bin i:

$$J_x(f) = \ell_e(f_e) + \kappa_i s_x \tau_e(f_e). \tag{7.11}$$

Inserting the definition of κ_i in the above, we obtain

$$J_x(f) = \ell_e(f_e) + \frac{s_x S'_{\mathrm{L}}}{\beta_{i-1}} \tau_e(f_e). \tag{7.12}$$

But $s_x \in [\beta_{i-1}, \beta_i]$ implies that

$$\ell_e(f_e) + S'_{\mathrm{L}} \tau_e(f_e) \leq J_x(f) \leq \ell_e(f_e) + S'_{\mathrm{U}} \tau_e(f_e), \tag{7.13}$$

or that agent x sees the exact same cost as he would if he were a member of population s'. Since this is true of every agent for every edge in every bin, it must be true that every Nash flow resulting from these binned tolls corresponds exactly to a Nash flow resulting from the nominal tolls $\{\tau\}$ and some sensitivity distribution in $[S'_{\mathrm{L}}, S'_{\mathrm{U}}]$. In particular, no binned Nash flow can have higher total latency than the worst-case flows for populations in $[S'_{\mathrm{L}}, S'_{\mathrm{U}}]$. Since $S'_{\mathrm{U}}/S'_{\mathrm{L}} = (S_{\mathrm{U}}/S_{\mathrm{L}})^{1/m}$, the theorem conclusion follows. ■

7.3 Optimal Binning for Simple Routing Problems

The principles proved in Theorem 7.2 are compelling, but in general may not result in optimal binning. That is, when investigating a restricted class of games, it may often be the case that we can design pricing that significantly outperforms the pricing described by Theorem 7.2.

In this section, we restrict attention to the class of parallel-network, affine-cost routing games. For the following results, let $\mathcal{G}^p \subseteq \mathcal{G}$ represent the class of all single-commodity, parallel-link routing problems with affine latency functions. That is, for all $e \in \mathcal{G}^p$, the latency function satisfies

$$\ell_e(f_e) = a_e f_e + b_e \tag{7.14}$$

where a_e and b_e are non-negative edge-specific constants. "Single-commodity" implies that all traffic has access to all network edges. Furthermore, we assume that every edge has positive flow in an un-tolled Nash flow. As in previous work in this dissertation, we study network-agnostic tolls here. Chapter 5 proved that affine tolls suffice to minimize the price of anarchy for affine congestion games; accordingly, in this chapter we limit our search to affine taxation mechanisms.

Definition 7.1 *The* Bounded Affine Taxation Mechanism *assigns tolls of*

$$\tau_e(f_e) = \kappa_1 a_e f_e + \kappa_2 b_e, \tag{7.15}$$

where a_e and b_e are the latency function coefficients in (7.14) and $\kappa_1 \leq \kappa_{\max}$ and $\kappa_2 \leq \kappa_{\max}$ are non-negative edge-independent constants upper-bounded by some $\kappa_{\max} \geq 0$.

Now, in Theorem 7.3, we show how to compute optimal bin boundaries and affine

Optimization Problem (P)

$$\max_{\{\beta_i\}_{i=1}^m} \gamma_{\mathrm{L}}$$

$$\text{s.t.} \quad \gamma_{\mathrm{L}} \le \frac{\kappa_{\max} + 1/\beta_i}{\kappa_{\max} + 1/\beta_{i-1}} \qquad \forall i \in \{1, \ldots, m\} \tag{7.19}$$

$$\beta_{i-1} \le \beta_i \qquad \forall i \in \{1, \ldots, m\}$$

$$\beta_0 = S_{\mathrm{L}}$$

$$\beta_m = S_{\mathrm{U}}$$

$$\beta_1 \ge \frac{1}{\kappa_{\max}^2 S_{\mathrm{L}}}. \tag{7.20}$$

Figure 7.1: As proved in Theorem 7.3, the solutions $\{\beta_i\}_{i=1}^m$ to this optimization problem are congestion-minimizing bin boundaries for any routing problem.

taxation function coefficients that minimize congestion for any $\kappa_{\max}$. Furthermore, we derive an upper bound for the price of anarchy that is independent of the number of network links and holds for any S_{U}.

Theorem 7.3 *For any $G \in \mathcal{G}^p$, for any set of bin boundaries $\{\beta_i\}_{i=1}^m$, the optimal bounded affine tolling coefficients are given by*

$$\kappa_1^i = \kappa_{\max} \tag{7.16}$$

$$\kappa_2^i = \max\left\{0, \frac{\left(\kappa_1^i\right)^2 \beta_{i-1}\beta i - 1}{\beta_{i-1} + \beta_i + 2\left(\kappa_1^i\right)\beta_{i-1}\beta_i}\right\}. \tag{7.17}$$

Furthermore, the congestion-minimizing bin boundaries $\{\beta_i\}_{i=1}^m$ can be found by solving Optimization Problem (P) (see Figure 7.1). Let $\lambda = \frac{\kappa_{\max} S_{\mathrm{L}}}{1+\kappa_{\max} S_{\mathrm{L}}}$. If $\kappa_{\max} \le 1/S_{\mathrm{L}}$, let $\mu = S_{\mathrm{L}}\kappa_{\max}$; otherwise, let $\mu = 1$. Then for all $S_{\mathrm{U}} \in [S_{\mathrm{L}}, \infty]$ and this binning B^m,

$$\mathrm{PoA}(\mathcal{G}^p, S_{\mathrm{L}}, S_{\mathrm{U}}, B^m) \le \frac{4}{3}\left(1 - \frac{\min\{\lambda^{1/m}, \mu\}}{\left(1 + \min\{\lambda^{1/m}, \mu\}\right)^2}\right). \tag{7.18}$$

A few words are in order regarding the price of anarchy bound in (7.18): First, this bound is tight for cases when $S_{\mathrm{U}} = \infty$; i.e., there is no upper bound on the price-sensitivities of the agents. When S_{U} is finite, the tools of Lemmas 7.3.1 and 7.3.2 can be used to determine an exact price of anarchy once the optimal bin boundaries have been derived. We are not aware of a convenient closed-form expression for the exact price of anarchy for finite S_{U}, but in Figure 7.3 we show that even for relatively low values of S_{U}, the gap between the finite-S_{U} and infinite-S_{U} prices of anarchy is not large.

Second, since we are dealing with bounded tolls, whenever $\kappa_{\max} < 1/S_{\mathrm{L}}$, it is not possible to guarantee perfectly-optimal network flows, even for arbitrarily-high m. This is because when $\kappa_{\max}$ is too low, a homogeneous sensitivity distribution with $s_x \equiv S_{\mathrm{L}}$ cannot be effectively influenced, and after a point, finer binning cannot remedy this. This is captured in the theorem statement by the parameter μ, which prevents extremely-fine binning from improving the price of anarchy when $\kappa_{\max}$ is too low.

See Figure 7.2 for a depiction of the congestion-minimizing bin boundaries for several values of m. In Figure 7.3, we depict the price of anarchy as a function of m, contrasting the guarantees provided by Theorems 7.2 and 7.3.

As a first step towards proving Theorem 7.3, we introduce Lemma 7.3.1, in which we present a powerful tool with which we can analyze the price of anarchy of parallel affine congestion games under various types of tolls.

Lemma 7.3.1 *Suppose that there exists a function $\gamma : [0,1] \to [\gamma_L, \gamma_U]$ with $0 \le \gamma_L \le 1$ and $\gamma_U \le 1/\gamma_L$ such that in every $G \in \mathcal{G}^p$, the cost function of user x on edge e is given by*

$$J_x^e(f_e) = (1 + \gamma(x))\, a_e f_e + b_e. \tag{7.21}$$

Then the price of anarchy of $\mathcal{G}^p$ is tightly upper-bounded by

$$\mathrm{PoA}(\mathcal{G}^p) \le \frac{4}{3}\left(1 - \frac{\gamma_L}{(1+\gamma_L)^2}\right). \tag{7.22}$$

Proof: The simplest proof of this relates (7.21) to the analysis of scaled marginal-cost

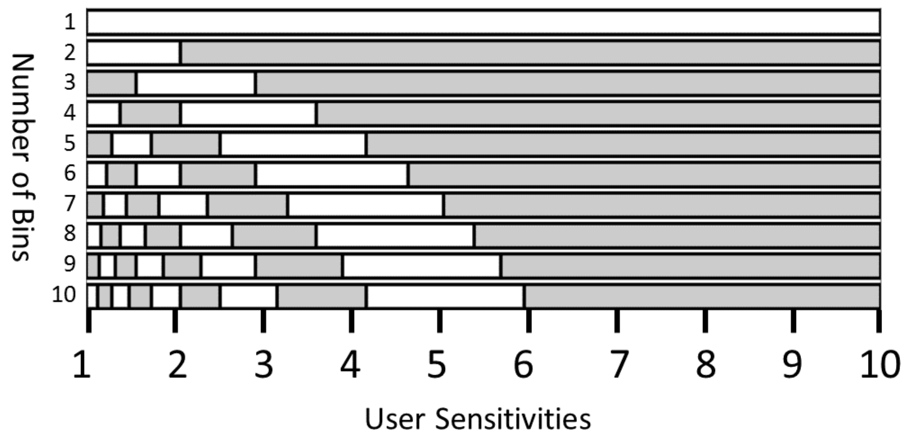

Figure 7.2: Optimal bin boundary locations computed by Optimization Problem (P) for $m \in \{1, \dots, 10\}$, with $\kappa_{\max} = S_{\mathrm{L}} = 1$ and $S_{\mathrm{U}} = 10$. These correspond exactly to the price of anarchy plot shown in Figure 7.3.

tolls presented in Chapter 4. There, in Lemma 4.2.2, we show that for $\kappa \leq 1/\sqrt{S_{\mathrm{L}} S_{\mathrm{U}}}$, of all the Nash flows induced by edge tolls $\tau_e(f_e) = \kappa a_e f_e$, the worst congestion always occurs for a homogeneous population in which all users' sensitivities are equal to S_{L}. To prove Lemma 7.3.1, we shall compute a virtual sensitivity distribution s^v and tolling coefficient κ^v which will induce Nash flows that precisely mimic the behavior of Nash flows induced by γ (that is, Nash flows for cost functions (7.21)).

Given the γ function of the statement of Lemma 7.3.1, let $\kappa^v = 1$, and let $s_x^v = \gamma(x)$ for all $x \in [0, 1]$. By this definition, any Nash flow induced by γ has a corresponding marginal-cost-tolled Nash flow induced by a sensitivity distribution given by s^v; we can accordingly use marginal-cost toll arguments to argue about γ-induced Nash flows. The upper and lower bounds of our virtual sensitivity distribution are thus given by $S_{\mathrm{L}}^v = \gamma_L$ and $S_{\mathrm{U}}^v = \gamma_U$, respectively. By the properties of γ_L and γ_U, it is always true that $\kappa^v \leq 1/\sqrt{S_{\mathrm{L}}^v S_{\mathrm{U}}^v}$, so the congestion-maximal Nash flows occur for a virtual homogeneous population with all users' sensitivities equal to S_{L}^v.

This implies that a Nash flow with all cost functions equal to $J_x^e(f_e) = (1 + \gamma_L)\, a_e f_e + b_e$ will have higher total congestion than a Nash flow induced by γ, and we can use techniques from Chapter 4 (namely, Lemma 4.2.3) to derive (7.22). The tightness of the bound in (7.22) is due to the constructive nature of the proof of Lemma 4.2.3. ■

Next, in Lemma 7.3.2, for any arbitrary bin boundaries, we derive specific tolling coefficients which should be charged in each bin.

Lemma 7.3.2 *For any $G \in \mathcal{G}^p$, for any bin boundaries $\{\beta_i\}$, the optimal tolling coefficients can be obtained for each bin i by choosing*

$$\kappa_1^i = \kappa_{\max} \tag{7.23}$$

$$\kappa_2^i = \max\left\{0, \frac{\left(\kappa_1^i\right)^2 \beta_{i-1}\beta i - 1}{\beta_{i-1} + \beta_i + 2\left(\kappa_1^i\right)\beta_{i-1}\beta_i}\right\}. \tag{7.24}$$

With these coefficients, in the language of Lemma 7.3.1, it holds that

$$\gamma_{\mathrm{L}} = \min_i \left\{\min\left\{\beta_{i-1}\kappa_{\max}, \frac{\kappa_{\max} + 1/\beta_i}{\kappa_{\max} + 1/\beta_{i-1}}\right\}\right\}. \tag{7.25}$$

Proof: Since uniform scaling by a constant factor does not change underlying Nash flows, we can say without loss of generality that the effective cost to agent $x \in [\beta_{i-1}, \beta_i]$ in bin i for edge e is given by

$$J_x(f_e) = \frac{1 + \kappa_1^i s_x}{1 + \kappa_2^i s_x} a_e f_e + b_e, \tag{7.26}$$

and when $\kappa_1^i \geq \kappa_2^i$, it is evident that[2]

$$\frac{1 + \kappa_1^i \beta_{i-1}}{1 + \kappa_2^i \beta_{i-1}} a_e f_e + b_e \leq J_x(f_e) \leq \frac{1 + \kappa_1^i \beta_i}{1 + \kappa_2^i \beta_i} a_e f_e + b_e. \tag{7.27}$$

For each i, for each $x \in [\beta_{i-1}, \beta_i]$, define $\gamma(x) \triangleq \frac{s_x(\kappa_1^i - \kappa_2^i)}{1 + \kappa_2^i s_x}$. Then any agent $x \in [0, 1]$ has

[2] In Chapter 5, it is shown that assuming $\kappa_1^i \geq \kappa_2^i$ is without loss of generality.

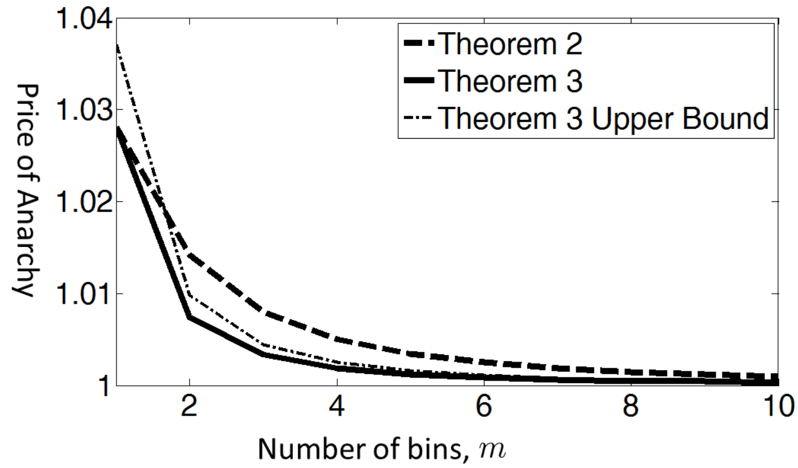

Figure 7.3: Comparison between the price of anarchy resulting from the generic price-discrimination approach of Therorem 7.2 (dashed line), the specific congestion-minimizing price-discrimination approach of Theorem 7.3 (solid line), and the general upper-bound given in Theorem 7.3 (dash-dot line). For the first two curves, we apply affine tolls to the class of parallel-network, affine-cost congestion games, with $\kappa_{\max} = S_{\mathrm{L}} = 1$ and $S_{\mathrm{U}} = 10$. The dash-dot line represents the general upper-bound proved in Theorem 7.3 that holds for this value of $\kappa_{\max}$ and S_{L}, and any S_{U}. Note how close this upper bound is to the instance-specific solid line for $m > 1$.

the following cost for edge e:

$$J_x^e(f_e) = (1 + \gamma(x))\, a_e f_e + b_e, \tag{7.28}$$

just as in (7.21). It is evident that in accordance with the assumptions of Lemma 7.3.1 we have

$$\begin{aligned} &\gamma(x) \geq \gamma_{\mathrm{L}} \triangleq \min_i \{\beta_{i-1}(\kappa_1^i - \kappa_2^i)/(1 + \kappa_2^i \beta_{i-1})\} \text{ and} \\ &\gamma(x) \leq \gamma_{\mathrm{U}} \triangleq \max_i \{\beta_i(\kappa_1^i - \kappa_2^i)/(1 + \kappa_2^i \beta_i)\} \end{aligned} \tag{7.29}$$

Equation (7.22) in Lemma 7.3.1 shows that we minimize the price of anarchy by maximizing γ_{L}, subject to $\gamma_{\mathrm{L}} \leq 1/\gamma_{\mathrm{U}}$.

First, assume we are given a fixed, arbitrary feasible set of bin boundaries $\{\beta_i\}$, a nonnegative value of κ_1^i for each bin, and that κ_2^i can take any real value. Because the problem is otherwise unconstrained, the constraint $\gamma_{\mathrm{L}} \leq 1/\gamma_{\mathrm{U}}$ will bind, or $\gamma_{\mathrm{L}} = 1/\gamma_{\mathrm{U}}$. Suppose γ_{L} is maximal with respect to the relevant constraints, and that bin i is the source of γ_{L}; i.e.,

$$\gamma_{\mathrm{L}} = \frac{\beta_{i-1}(\kappa_1^i - \kappa_2^i)}{1 + \kappa_2^i \beta_{i-1}}. \tag{7.30}$$

Simultaneously, κ_2^i must satisfy the following (the only other constraint on κ_2^i):

$$\frac{\beta_i(\kappa_1^i - \kappa_2^i)}{1 + \kappa_2^i \beta_i} \geq \gamma_{\mathrm{U}}. \tag{7.31}$$

It is clear that γ_{L} is decreasing in κ_2^i; if γ_{L} is indeed optimal, κ_2^i must be binding the constraint in (7.31). Thus, a single bin generates both γ_{L} and γ_{U}, and since $\gamma_{\mathrm{L}} = 1/\gamma_{\mathrm{U}}$, we have that

$$\frac{\beta_{i-1}(\kappa_1^i - \kappa_2^i)}{1 + \kappa_2^i \beta_{i-1}} = \frac{\beta_i(\kappa_1^i - \kappa_2^i)}{1 + \kappa_2^i \beta_i}, \tag{7.32}$$

which implies that

$$\kappa_2^i = \frac{\left(\kappa_1^i\right)^2 \beta_{i-1}\beta i - 1}{\beta_{i-1} + \beta_i + 2\left(\kappa_1^i\right)\beta_{i-1}\beta_i} \tag{7.33}$$

Now, if we re-introduce the constraint that $\kappa_2^i \geq 0$, we find that it may no longer be possible to satisfy (7.33), but that since γ_{L} is decreasing in some κ_2^i, simply saturating κ_2^i at 0 still yields a maximal γ_{L} while respecting $\gamma_{\mathrm{L}} \leq 1/\gamma_{\mathrm{U}}$. Thus, for any binning, choosing κ_2^i according to (7.24) is sufficient to ensure an optimal price of anarchy.

Finally, with (7.24), it is simple to show that γ_{L} is nondecreasing in each of the κ_1^i coefficients, so letting $\kappa_1^i = \kappa_{\max}$ suffices to minimize the price of anarchy. ■

Proof of Theorem 7.3

First, note that Optimization Problem (P) is largely a re-statement of Lemmas 7.3.1 and 7.3.2, with the sole exception of constraint (7.20). By the definition of γ_{L} in Lemma 7.3.1, it is clear why we desire to maximize γ_{L}. By the fact regarding γ_{L} in Lemma 7.3.2, it is clear that any optimal binning will satisfy constraint (7.19).

The only curious element of the optimization problem is constraint (7.20), which we show here does not reduce the optimality of the solutions, even when the constraint is active. This constraint plays a role in cases when $\kappa_{\max}$ is small, and avoids the non-smoothness of κ_2^i in (7.24) and γ_{L} in (7.25). Let $\{\beta_i^*\}$ be an optimal solution to the optimization problem in which constraint (7.20) binds. Then $\beta_1^* = 1/(\kappa_{\max}^2 S_{\mathrm{L}})$, or $\kappa_{\max} = 1/\sqrt{S_{\mathrm{L}}\beta_1^*}$. Note that at this point, according to (7.24), $\kappa_2^1 = 0$, and this is the precise breakpoint at which the expression for κ_2^1 "switches over" from 0 to the non-constant function of $\kappa_{\max}$. Thus, the first effect of constraint (7.20) is that it ensures that κ_2^1 will always be a smooth function of the bin boundaries.

Second, considering (7.25), note that $\beta_1 = 1/(\kappa_{\max}^2 S_{\mathrm{L}})$ is also the precise breakpoint at which γ_{L} "switches over" from $\beta_{i-1}\kappa_{\max}$ to $(\kappa_{\max} + 1/\beta_i)/(\kappa_{\max} + 1/\beta_{i-1})$. Thus, the second effect of constraint (7.20) is that it ensures that γ_{L} itself will always be a consistent function of the bin boundaries.

To see that including (7.20) does not reduce the optimality of solutions, note that if it were true that $\beta_1^* < 1/(\kappa_{\max}^2 S_{\mathrm{L}})$, it would also necessarily be true that the price of anarchy would simply be determined by $\gamma_{\mathrm{L}} = S_{\mathrm{L}}\kappa_{\max}$. Thus, β_1 has no impact on the price of anarchy until it reaches the $1/(\kappa_{\max}^2 S_{\mathrm{L}})$ threshold.

The essential uniqueness of the solution of (P) follows from the fact that the only way to increase γ_{L} is to raise the lower-boundary of some bin, which simultaneously raises the upper-boundary of the next-lower bin, serving to decrease γ_{L}. If (7.20) is active at an optimizer of (P), $\{\beta_i\}_{i=2}^m$ are not unique, but all optimizers yield the same price of anarchy.

To show the price of anarchy bound, first note that for a fixed m and S_{L}, the price of anarchy will always be increasing in S_{U}, since this represents increasing the sensitivity-uncertainty of our population. Our arguments thus involve investigating the limiting price of anarchy as $S_{\mathrm{U}} \to \infty$.

First, suppose that $\kappa_{\max} < 1/S_{\mathrm{L}}$, so that for some binnings we activate constraint (7.20).

When (7.20) is active, the price of anarchy is determined only by the lowest-index bin; in the language of Lemma 7.3.1, $\gamma_{\mathrm{L}} = S_{\mathrm{L}}\kappa_{\max}$. This can be considered a worst-case situation, so we include it in the price of anarchy expression (7.18) via the μ argument.

However, even when $\kappa_{\max} < 1/S_{\mathrm{L}}$, it is possible that constraint (7.20) will not be active, so we must consider the case when constraint (7.19) is active. In this case, each bin has the same "width," or for $i = (m-1)$, it is true that

$$\gamma_{\mathrm{L}} = \frac{\kappa_{\max} + 1/\beta_{m-1}}{\kappa_{\max} + 1/\beta_{m-2}} = \frac{\kappa_{\max} + 1/S_{\mathrm{U}}}{\kappa_{\max} + 1/\beta_{m-1}}.$$

In general, closed-form expressions for bin boundaries resulting from this are quite complicated, but considering a high-S_{U} case can simplify things considerably:

$$\lim_{S_{\mathrm{U}}\to\infty} \gamma_{\mathrm{L}} = \frac{\kappa_{\max} + 1/\beta_{m-1}}{\kappa_{\max} + 1/\beta_{m-2}} = \frac{\kappa_{\max}}{\kappa_{\max} + 1/\beta_{m-1}}. \tag{7.34}$$

Since the above is true for any $m > 1$, we can use it to inductively deduce the structure of an optimal binning for any positive S_{L} and infinite S_{U}. First, letting $m = 2$, fixing β_1 implies the following unique value for S_{L}:

$$S_{\mathrm{L}} = \kappa_{\max}/\left(\left(\kappa_{\max} + 1/\beta_1\right)^2 - \kappa_{\max}^2\right).$$

Inductively, it can be shown that for arbitrary m and fixed β_{m-1}, the unique implied value of S_{L} is given by

$$S_{\mathrm{L}} = \kappa_{\max}^{m-1}/\left(\left(\kappa_{\max} + 1/\beta_{m-1}\right)^m - \kappa_{\max}^m\right).$$

Solving this for β_{m-1}, we have that for any S_{L}, $\kappa_{\max}$, m, and infinite S_{U},

$$\beta_{m-1} = \left(\kappa_{\max}\left[\left(\frac{1+\kappa_{\max}S_{\mathrm{L}}}{\kappa_{\max}S_{\mathrm{L}}}\right)^{1/m} - 1\right]\right)^{-1}. \tag{7.35}$$

This β_{m-1} represents the highest bin boundary in an optimal binning for an infinite-S_{U} population. To compute the corresponding $\gamma_{\mathrm{L}}^{\infty}$, we can simply substitute (7.35) into (7.34) and simplify:

$$\gamma_{\mathrm{L}}^{\infty} = \left(\frac{\kappa_{\max} S_{\mathrm{L}}}{1 + \kappa_{\max} S_{\mathrm{L}}}\right)^{1/m}, \tag{7.36}$$

the source of λ in the theorem statement. Any finite S_{U} will yield a better price of anarchy than an infinite one, so for any game, $\gamma_{\mathrm{L}} \geq \min\{\kappa_{\max} S_{\mathrm{L}}, \gamma_{\mathrm{L}}^{\infty}\}$, so by Lemma 7.3.1 the expression (7.18) is valid for any S_{U}. ■

Part II

Game Theory for Distributed Control

Chapter 8

Informational Fragility in Distributed Learning

In recent years, game theory has received considerable attention as an overarching framework for the study of multi-agent engineered systems such as distributed power generators or swarms of autonomous vehicles [4]. A central line of reasoning in this setting involves posing the problem of interest as a distributed optimization problem, computing appropriate utility functions for agents, endowing agents with a simple learning rule, and then appealing to results in game-theoretic learning to prove that the applied learning rule causes system-optimal states to emerge [4,22,52,88]. As an example, the associated game may be designed to be a potential game whose potential function is exactly the main optimization objective, and thus the system-optimal states are potential-maximizing Nash equilibria of the associated game [89]. When this is the case, many results from the literature on learning in games suggest that there exist learning rules which allow the agents to solve the original problem in a distributed fashion [90–94].

This line of reasoning typically relies on an implicit assumption that the agents have a full specification of their utility functions *and* can accurately observe the action choices of other agents [49–51].[1] As a simple example, for an agent to apply a best-response rule, the agent must have some means of determining which of her actions maximize her utility function given the actions of other agents – implicitly requiring that she has sufficient information about other agents' actions to do so. In practice, this requirement may not always be met; how should an agent react on-the-fly to an unexpected loss of information about other agents' actions? In this chapter, we ask a question of robustness: if an agent cannot observe some other (relatively) unimportant agent's action, is there a way to endow that agent with a simple adaptation policy such that the learning dynamics would still converge to (or close to) the nominal system optimum?

To illustrate these concepts concretely, we present the following 3-player, 2-action game, for some small $\delta > 0$. Note that "players" and "agents" are used interchangeably.

		Player 2			Player 2	
		A	B		A	B
Player 1	A	$1-\delta,\ 1$	$0,\ 0$	A	$0,\ 2\delta$	$2\delta,\ 0$
	B	$\delta,\ 0$	$1,\ 0$	B	$3\delta,\ \delta$	$\delta,\ 2\delta$
Player 3		A			B	

The rows indicate the action choices available to Player 1, the columns indicate the action choices available to Player 2, and Player 3 selects between the left and right matrices. Let Player 2 and Player 3 have identical utility functions, so each cell in the matrices depicts the payoffs for Player 1 and Player 2/3, respectively as a function of joint action choices.

The optimal action profile (measured by the sum of the players' payoffs) is the upper-left, and all best-response paths lead to this unique Nash equilibrium.[2] If Player 1 does not know the action of Player 2 (and also does not know Player 2's utility function), how should

[1] An alternative approach involves the so-called "payoff-based" learning rules; here, explicit communication between agents or observation of actions is avoided by the assumption that each agent merely realizes the output of its utility function as a payoff, and attempts to play actions which are historically correlated with high payoffs. However, these payoffs are still functions of the actions of other agents [95, 96].

[2] A best-response path is a sequence of unilateral payoff-maximizing action updates by players, formally defined in Section 8.3.3; a Nash equilibrium is a joint action profile from which no agent can unilaterally deviate and improve payoff, defined in (8.1).

Player 1 evaluate her own action choices? In this chapter we assume that Player 1 knows what payoffs she *could* receive for any action choice by Player 2; for example, If Player 1 and Player 3 are both playing A, Player 1 knows that her payoff is either $1-\delta$ or 0. Given this information, Player 1 needs to assign a "proxy payoff" to this situation as a function of the payoffs she could be receiving; we call such a policy for computing proxy payoffs an *evaluator*, formally defined in Definition 8.1 (Section 9.1). One simple evaluator is to choose the maximum payoff from each row, yielding this effective payoff matrix:

		Player 2			Player 2	
		A	B		A	B
Player 1	A	$1-\delta,\ 1$	$1-\delta,\ 0$	A	$2\delta,\ 2\delta$	$2\delta,\ 0$
	B	$1,\ 0$	$1,\ 0$	B	$3\delta,\ \delta$	$3\delta,\ 2\delta$
Player 3		A			B	

Note that Player 2 and Player 3's payoffs are unchanged, but Player 1's experienced payoffs are now independent of Player 2's action. In this modified game, Player 1 now always prefers action B over action A, and this causes all best-response paths to lead to the inefficient lower-right action profile. Furthermore, note that this conclusion stands for many different evaluators chosen by Player 1: if she had chosen the minimum payoff in each row rather than the maximum, the result is unchanged – and the same holds for mean and sum.

However, it appears clear from the nominal payoffs that Player 2 was important to Player 1's decision-making, and this made it impossible for Player 1 to compute helpful proxy payoffs when information about Player 2 was lost. Indeed, the communication failure amounted to a large, essentially-arbitrary perturbation of the payoff matrix, and it is well-known that emergent behavior in games can be significantly altered by adversarial manipulation of payoffs [97]. However, if Player 2 had been "inconsequential" to Player 1 in some sense, would that have lessened the harm of the lost information? Our goal is to explore this issue and determine when and how it is possible to endow agents with policies for computing proxy payoffs to protect against losses of information about inconsequential agents. Specifically, some questions addressed in this chapter are

1. Is there a method of computing proxy payoffs which can provide some guarantee that emergent behavior is good despite lost information?

2. Are there particular problem structures that confer greater degrees of robustness than others?

This chapter's main contribution is to show that *regardless of how proxy payoffs are chosen,* in several settings, loss of action information about even weakly-coupled agents can cause arbitrarily-low-quality states to emerge as various game solution concepts – even for games that are quite well-behaved nominally. Definition 8.2 formalizes a notion of weak coupling that we term *inconsequentiality*: if Player 2 is ϵ-inconsequential to Player 1 this means that for any joint action, a unilateral action change by Player 2 can cause no more than an ϵ change in Player 1's payoff (See Definition 8.2). That is, even if Player 2's action is unknown, Player 1 can always estimate her own payoff to within ϵ.[3] We call a situation in which one agent cannot observe another agent's action choice a *communication failure*, and if the unobserved agent is ϵ-inconsequential to the observing agent, we call it an ϵ-*failure.*

We next consider the well-studied class of potential games [100]. It is known that in any potential game possessing a unique Nash equilibrium, discrete best-response dynamics converge to that Nash equilibrium (thus maximizing the potential function). However, Theorem 8.3 here shows that for any $\epsilon > 0$, there exist games in this class for which an ϵ-failure causes best response dynamics to converge to states which essentially minimize the potential function, *regardless of which payoff evaluator is applied.* This result holds even if the payoff evaluator is selected as a function of the full game, and even if the nominal game possesses a unique equilibrium.

Subsequently, Section 8.2.2 shows via Proposition 8.6 that the class of identical interest

[3] This notion of inconsequentiality is tightly connected with the notion of *influence* in [98] and the definition of a game's Lipschitz constant in [99].

games is slightly better-behaved than general potential games, in the sense that even under a communication failure, every identical interest game always possesses at least one efficient Nash equilibrium. Next, we appeal to the well-studied stochastic learning rule known as log-linear learning, which is known to stabilize high-quality states of games in many situations [94, 101]. Unfortunately, Theorem 8.7 shows that in the presence of ϵ-failures, noisy learning dynamics can actually destabilize efficient Nash equilibria and stabilize inefficient ones. Here, there is no general way to compute proxy payoffs that can prevent inefficient action profiles from emerging as stochastically stable states of log-linear learning.

Section 8.3 presents a pair of positive results showing features which can limit the harm of communication failures when agents update their actions stochastically according to log-linear learning. Theorem 8.9 shows that if *every* agent loses information about a single inconsequential agent's action, identical-interest games retain their desirable properties and high-quality states remain stochastically stable. Furthermore, we show in Theorem 8.10 that the damage a communication failure can cause in a potential game is limited by the total number of action profiles in a game.

Lastly, Section 8.3.3 proposes a way to certify whether a set of proxy payoffs is susceptible to the types of pathologies demonstrated in this chapter. This certificate looks for some notion of alignment between the chosen proxy payoffs and the nominal potential function of the game; if these are sufficiently aligned, then losing communication with inconsequential players can cause no harm.

8.1 Model

8.1.1 Game theoretic preliminaries

We model a multiagent system as a finite strategic-form game with player set $\mathcal{I} = \{1, \ldots, n\}$ where each player $i \in \mathcal{I}$ selects an action from action set $\mathcal{A}_i$. The preferences of each player over his actions are encoded by a utility function $U_i : \mathcal{A} \to \mathbb{R}$ where $\mathcal{A} =$

$\mathcal{A}_1 \times \cdots \times \mathcal{A}_n$; that is, each player's payoff is a function of his own action and the actions of all other players. A game is specified by the tuple $G = (\mathcal{I}, \mathcal{A}, \{U_i\}_{i\in\mathcal{I}})$.

For an action profile $a = (a_1, a_2, \ldots, a_n) \in \mathcal{A}$, let a_{-i} denote the profile of player actions other than player i; i.e.,

$$a_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n).$$

Similarly, a_{-ij} denotes the profile of player actions other than players i and j:

$$a_{-ij} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_{j-1}, a_{j+1}, \ldots, a_n).$$

With this notation, we will sometimes write an action profile a as (a_i, a_{-i}). Similarly, we may write $U_i(a)$ as $U_i(a_i, a_{-i})$. Let $\mathcal{A}_{-i} = \Pi_{j\neq i}\mathcal{A}_j$ denote the set of possible collective actions of all players other than player i. We define player i's *best response set* for an action profile $a_{-i} \in \mathcal{A}_{-i}$ as $B_i(a_{-i}) := \arg\max_{a_i\in\mathcal{A}_i} \; U_i\left(a_i, a_{-i}\right)$. An action profile $a^{\mathrm{ne}} \in \mathcal{A}$ is known as a *pure Nash equilibrium* if for each player i,

$$a_i^{\mathrm{ne}} \in B_i\left(a_{-i}^{\mathrm{ne}}\right). \tag{8.1}$$

That is, all players are best-responding to each other. The set of pure Nash equilibria of game G is denoted PNE(G).

8.1.2 Distributed optimization and game classes

We assume that there is a global welfare function (i.e., objective function) $W : \mathcal{A} \to [0, 1]$ that the players are trying to maximize. We assume without loss of generality that at least one action profile achieves the welfare maximum; i.e., in every game there exists $\bar{a} \in \mathcal{A}$ such that $W(\bar{a}) = 1$.

A common game-theoretic formulation of distributed optimization problems involves assigning utility functions to players that are derived from W in some way that ensures that $\arg\max_a W(a) \subseteq \mathrm{PNE}(G)$. One way to assign utility functions that accomplishes this is simply to give each player i the full objective function, or let

$$U_i(a_i, a_{-i}) := W(a_i, a_{-i}).$$

This induces an *identical-interest game*, or one in which all players have the same utility function. In this chapter, if a game is stated to be an identical-interest game, it is assumed that the common utility function is the welfare function W.

A somewhat more nuanced method of assigning player utility functions is known as "marginal-contribution" utility design [89]. Here, for each player i, arbitrarily select a baseline action $a_i^b \in \mathcal{A}_i$, and assign the player a utility function of

$$U_i(a_i, a_{-i}) := W(a_i, a_{-i}) - W(a_i^b, a_{-i}).$$

Assigning utility functions in this way ensures that in the resulting game, whenever a player unilaterally changes actions in a way that improves her utility, this improves W by the same amount. Formally, for every player $i \in \mathcal{I}$, for every $a_{-i} \in \mathcal{A}_{-i}$, and for every $a_i', a_i'' \in \mathcal{A}_i$,

$$U_i\left(a_i', a_{-i}\right) - U_i\left(a_i'', a_{-i}\right) = W\left(a_i', a_{-i}\right) - W\left(a_i'', a_{-i}\right). \tag{8.2}$$

A game that satisfies (8.2) for some W is known as a *potential game* with *potential function* W [100]. In a potential game, there is a strong notion of alignment between the interests of the various players, and action profiles which maximize W are always pure Nash equilibria: $\arg\max_a W(a) \subseteq \mathrm{PNE}(G)$. Further, many distributed learning algorithms converge in potential games to potential-maximizing action profiles [20].

Throughout this chapter, if a game is stated to be a potential game, it is assumed that its potential function is equal to the welfare function W. For non-potential games, the welfare function will be specified as needed.

8.1.3 Online learning and its associated solution concepts

Once the utility functions have been assigned offline, how should the agents choose their actions online? In this chapter, we have agents update their actions by stochastic asynchronous processes that proceed as follows. Starting at some initial joint action $a(0)$, at each time $t \in \mathbb{N}$ an agent i (called the *updating agent*) is selected uniformly at random from $\mathcal{I}$ to choose an action $a_i(t+1)$ to play at time $t+1$, and all other agents simply repeat their previous action: $a_{-i}(t+1) = a_{-i}(t)$.

Asynchronous best reply process

The prototypical learning rule that we will subsequently build upon is the *asynchronous best reply process*, where at each time step the updating agent i selects an action from their best response set uniformly at random. Formally, conditional on agent i being the updating agent at time t, the probability that agent i selects action a_i' in the next time step is given by

$$\Pr\left[a_i(t+1) = a_i' \mid a_{-i}(t)\right] = \begin{cases} \frac{1}{|B_i(a_{-i}(t))|} & \text{if } a_i' \in B_i(a_{-i}(t)), \\ 0 & \text{otherwise.} \end{cases} \tag{8.3}$$

For any strategic-form game, the asynchronous best reply process defines a Markov process P^{br} over joint action profiles $\mathcal{A}$; as such, we will frequently refer to a joint action profile a as a "state" of P^{br}, and denote a sample path of P^{br} as $a^{\text{br}}(t)$. A *recurrent class* of P^{br} is a set of action profiles $A \subseteq \mathcal{A}$ such that if the process is started in A it remains in A (i.e., if $a^{\text{br}}(0) \in A$, then $a^{\text{br}}(t) \in A$ for all $t > 0$), and for any states $a, a' \in A$, the probability that the process started at a eventually visits a' is positive: $\Pr\left[a^{\text{br}}(t) = a' \text{ for some } t \geq 1 \,\middle|\, a^{\text{br}}(0) = a\right] > 0$. We denote the set of recurrent classes of the asynchronous best reply process of a game G under P^{br} by $\text{ABR}(G)$; since the state space is finite, $\text{ABR}(G)$ is never empty. With a slight abuse of notation, we will occasionally write $a \in \text{ABR}(G)$ to mean that $a \in A$ for some $A \in \text{ABR}(G)$.

The recurrent classes $\text{ABR}(G)$ fully characterize the long-run behavior of P^{br} in the sense that for any initial joint action $a^{\mathrm{br}}(0)$, standard Markov results show that the process eventually enters some recurrent class almost surely. In general, a game may have many recurrent classes, but if G is a potential game with a unique pure Nash equilibrium a^{ne}, then it holds that $\{a^{\mathrm{ne}}\}$ is the unique recurrent class of P^{br} and thus $a^{\mathrm{br}}(t)$ converges to a^{ne} almost surely. However, even a potential game may have multiple strict Nash equilibria, in which case each equilibrium forms its own recurrent class – and it is not possible to predict *a priori* which of these will eventually capture the process. Moreover, it is possible for P^{br} to have a cycle as a recurrent class.

Log-linear learning preliminaries

Motivated in part by this ambiguity, a rich body of literature has developed to study "noisy" best-reply processes, in which a nominal best-reply process is perturbed by a small amount of noise to make its associated Markov chain ergodic. The typical setup is that players are randomly offered opportunities to choose a new action, and with some small probability, they choose suboptimal actions. In the context of a social system, this "choice of suboptimal action" is viewed as the player mistakenly choosing an action that is not a best response; if the game is meant to model an engineered system, the suboptimal action is used as a means for the agents to explore the state space [90–93,101,102]. One such noisy update process is the *log-linear learning* rule [103], defined as follows.[4]

At time t, updating agent i chooses its next action probabilistically as a function of the payoffs associated with the current action profile: the probability of choosing some action a_i' in the next time step is given by

$$\Pr\left[a_i(t+1) = a_i' \mid a_{-i}(t)\right] = \frac{e^{\beta U_i(a_i', a_{-i}(t))}}{\sum_{a_i \in \mathcal{A}_i} e^{\beta U_i(a_i, a_{-i}(t))}},$$

[4]This chapter investigates log-linear learning in particular for reasons of parsimony and concreteness; several of our results can be readily extended to more general noisy best-response processes. Formally making this extension would considerably increase the notational density of the chapter while adding little substance.

where $\beta > 0$ is a parameter indicating the degree to which agents desire to select their best response. If $\beta = 0$, agents select actions uniformly at random; as $\beta \to \infty$, the limiting process is the asynchronous best-reply process of Section 9.1.

For any $\beta > 0$, log-linear learning induces an ergodic Markov process on $\mathcal{A}$; denote its unique stationary distribution by π^β. It is well-known that the limiting distribution $\pi \triangleq \lim_{\beta\to\infty} \pi^\beta$ exists, and that it is a stationary distribution of the asynchronous best-reply process. Let $\pi(a)$ denote the probability of joint action $a \in \mathcal{A}$ being played in the limiting distribution π; if $\pi(a) > 0$, then action profile a is said to be *stochastically stable* under log-linear learning. It is known that for any game G,

$$a \in \mathrm{SS}(G) \implies a \in \mathrm{ABR}(G). \tag{8.4}$$

Furthermore, existing results give us that in a potential game with potential function W, the set of stochastically stable action profiles under log-linear learning is equal to the set of potential function maximizers [94]:

$$\mathrm{SS}(G) = \arg\max_{a\in\mathcal{A}} W(a). \tag{8.5}$$

8.1.4 Communication failures

We define a *communication failure* as a situation in which a single player loses access to information about the action of a single other player; without loss of generality, let these players be Player 1 and Player 2, respectively. We will commonly say that "Player 2 is *hidden* from Player 1." Note that more general formulations are possible, but in this chapter we investigate this special case as it suffices to highlight several important issues. Furthermore, it seems likely that allowing more than one communication failure in a game would worsen our results.

When Player 1 cannot observe Player 2's action, this means that the utility function U_1

must be modified in some way so that it no longer depends on the action choice of Player 2. That is, Player 1 must adopt a *proxy payoff function* $\tilde{U}_1(a_1, a_{-12})$ that becomes the new basis for decision-making, but that does not depend on the action choice of Player 2.

Computing $\tilde{U}_1$ means assigning a value to each (a_1, a_{-12}), taking into account that the true utility is a function of the unobservable action of Player 2 as well. That is, the true utility is some unknown number in the set $\{U_1(a_1, a_2, a_{-12}) : a_2 \in \mathcal{A}_2\}$. To compute proxy payoffs, we assign Player 1 an *evaluator* f, which is a function that for each (a_1, a_{-12}), takes the set of possible payoffs and returns a proxy payoff:

$$\tilde{U}_1\left(a_1, a_{-12})\right) = f\left(\{U_1(a_1, a_2, a_{-12}) : a_2 \in \mathcal{A}_2\}\right). \tag{8.6}$$

The space of feasible evaluators is large, and we restrict it in only two simple ways, as indicated by the following definition:

Definition 8.1 *An* acceptable evaluator *f is a mapping from sets of numbers to $\mathbb{R}$ satisfying the following properties. Let $S = (s_i)_{i=1}^k$ and $S' = (s_i')_{i=1}^k$, where S and S' are assumed to be ordered increasing:*

1. *If $s_i > s_i'$ for each i, then $f(S) > f(S')$,*

2. *If $s_i = s_i'$ for each i, then $f(S) = f(S')$.*

If an acceptable evaluator further satisfies $f(S) \in [\min(S), \max(S)]$, it is called a bounded acceptable evaluator.

If Player 1 uses acceptable evaluator f to compute proxy payoffs for the case when Player 2's action is unobservable, we say that Player 1 *applies* f to Player 2. Proposition 8.1 gives a partial list of evaluators which are acceptable by Definition 8.1; its proof is included in Section 8.4.

Proposition 8.1 *The following functions are acceptable evaluators; numbers (2) through (4) are also bounded:*

1. *Sum:* $f_{\text{sum}}(S) = \sum_{s \in S} s$

2. *Maximum element:* $f_{\max}(S) = \max_{s \in S} s$

3. *Minimum element:* $f_{\min}(S) = \min_{s \in S} s$

4. *Mean:* $f_{\text{mean}}(S) = \frac{1}{|S|} \sum_{s \in S} s$

Given a nominal game G and evaluator f, we write G_f to denote the *reduced* game generated when Player 1 applies evaluator f to Player 2.

8.1.5 Assessing the quality of an evaluator

We consider an acceptable evaluator f to be effective if for any nominal game G, the reduced game G_f induced by f is similar to the nominal game, where this similarity is measured by the welfare of the games' equilibria. For a game G, let $\mathcal{E}(G)$ be some set of equilibria associated with G; for example, $\mathcal{E}(G)$ could represent the set of recurrent classes of the asynchronous best-reply process for G. In the forthcoming, we write $\mathcal{G}$ to denote a given class of games.

This chapter presents a number of negative and positive results; to show the negative results, we use an optimistic measure of quality given by

$$Q_{\mathcal{E}}^{-}(\mathcal{G}, f) \triangleq \inf_{G \in \mathcal{G}} \frac{\max\limits_{a \in \mathcal{E}(G_f)} W(a)}{\min\limits_{a \in \mathcal{E}(G)} W(a)}. \tag{8.7}$$

Note that here, by checking the ratio of maximum to minimum welfare, this quality measure is designed to produce the highest values possible. If (8.7) is close to 0, this indicates that for some game $G \in \mathcal{G}$, the *best* equilibria induced by f can perform far worse than the *worst* equilibria of the nominal game. Accordingly, all of our negative results show situations in which this form of the quality metric can be close to 0. We sometimes wish to evaluate (8.7) on an individual game (i.e., a singleton class of games) and denote this with $Q_{\mathcal{E}}^{-}(G, f) := Q_{\mathcal{E}}^{-}(\{G\}, f)$.

Alternatively, to show the positive results, we use a pessimistic measure of quality given by

$$Q_{\mathcal{E}}^{+}(\mathcal{G}, f) \triangleq \inf_{G\in\mathcal{G}} \frac{\min\limits_{a\in\mathcal{E}(G_f)} W(a)}{\max\limits_{a\in\mathcal{E}(G)} W(a)}. \tag{8.8}$$

In contrast to (8.7), here by checking the ratio of minimum to maximum welfare, the quality measure is designed to produce the lowest values possible. That is, if (8.8) is large (close to 1), this indicates that for every game in $\mathcal{G}$, the *worst* equilibria induced by f are nearly as good as the *best* equilibria of the nominal game. Accordingly, all of our positive results show situations in which this form of the quality metric can be close to 1.

We can now state the main goal of this chapter. We wish to find payoff evaluators f which can ensure that these measures of quality are high for meaningful classes of games. That is, given some $\mathcal{G}$ and $\mathcal{E}$, we wish to find f to maximize $Q_{\mathcal{E}}^{+}(\mathcal{G}, f)$.

8.2 Resilience against communication failures is challenging

In multiagent systems, it is an attractive goal to endow agents with local policies which allow them to react on-the-fly to losses of information about other agents' actions. Unfortunately, the results in this section illustrate that general methods for doing so may be elusive. Unless otherwise stated, proofs of all theorems appear in Section 8.4.

First, note that if no restriction is placed on which types of game we are considering, a single communication failure can easily and catastrophically degrade performance. The following proposition assumes that Player 1 loses information about the action choice of Player 2, and thus must compute proxy payoffs for the case when Player 2's action is unobservable.

Proposition 8.2 *Let $\mathcal{G}$ be the set of all games. If Player 1 applies any acceptable evaluator f to Player 2, then for all $\epsilon > 0$, it holds that*

$$Q_{\mathrm{ABR}}^{-}(\mathcal{G}, f) \leq \epsilon. \tag{8.9}$$

Here, by showing that the optimistic measure $Q^{-}_{\mathrm{ABR}}(\mathcal{G}, f)$ is close to 0, we see that in general games, no evaluator can prevent best-response processes from selecting arbitrarily-inefficient action profiles.

Proof: This is shown using the game presented in the introduction paired with welfare function $W(a) = \sum_i U_i(a)$ (normalized to have a range of $[0, 1]$). In that game, if Player 2 is hidden from Player 1 there exists no acceptable evaluator which can prevent the bottom-right action profile from being the unique strict Nash equilibrium to which all best-response paths lead. This is because for any acceptable evaluator f, action B is a strictly dominant strategy for Player 1. The upper-left action profile has welfare $3 - \delta$, but the lower-right action profile has welfare 5δ. Thus, for any $\epsilon > 0$, letting $0 < \delta < 3\epsilon/(5 + \epsilon)$ yields the proof. ■

8.2.1 Results for recurrent classes of the asynchronous best reply process

It is clear from Proposition 8.2 that some structure is needed in order to prevent communication failures from causing harm. Potential games offer a natural starting point for studying resilience in this context; intuitively, since the payoffs of agents in a potential game are nicely aligned, this might offer a degree of protection.

Furthermore, we wish to study games in which hidden players are in some sense only weakly important to the players which cannot observe them. We introduce the following notion of weak interrelation: we say Player 2 is "inconsequential" to Player 1 if Player 2 can never cause a large change in Player 1's payoff by changing actions. We make this notion precise in Definition 8.2:

Definition 8.2 *Player j is* ϵ-inconsequential *to player i if for all $a_i \in \mathcal{A}_i$, all $a_{-ij} \in \mathcal{A}_{-ij}$, and all $a_j, a_j' \in \mathcal{A}_j$,*

$$\left|U_i\left(a_i, a_j, a_{-ij}\right) - U_i\left(a_i, a_j', a_{-ij}\right)\right| \leq \epsilon. \tag{8.10}$$

Now, let $\mathcal{G}^{\mathrm{PG}}_{\epsilon}$ denote the class of potential games for which Player 2 is no more than ϵ-incon-sequential to Player 1. Our standing assumption will be that in each game, Player 1 loses information about the action of Player 2, and thus must apply an acceptable evaluator to Player 2. That is, for each game $G \in \mathcal{G}_{\epsilon}^{\mathrm{PG}}$, we are assured that even if Player 1 cannot observe Player 2's action, a unilateral deviation by Player 2 can have only a small impact on Player 1's payoff. One might hope that by imposing the additional structure provided by potential games and inconsequentiality that the severe pathologies of Proposition 8.2 could be avoided. Indeed, it can be readily shown that for the special case of $\epsilon = 0$, for any game $G \in \mathcal{G}_{\epsilon=0}^{\mathrm{PG}}$, if Player 1 applies any acceptable evaluator f to Player 2 we have that $\mathrm{ABR}(G) = \mathrm{ABR}\left(G_f\right)$.

Unfortunately, Theorem 8.3 demonstrates that whenever $\epsilon > 0$, even a single communication failure can cause significant harm to emergent behavior in a game. Note that Theorem 8.3 uses the optimistic form of the quality measure from (8.7) with $\mathcal{E} = \mathrm{ABR}$; that is, it evaluates the best state in any recurrent class (of the best-reply proces) of the reduced game against the worst state in any recurrent class of the nominal game.

Theorem 8.3 *For any $\epsilon > 0$, let $\mathcal{G}_{\epsilon}^{\mathrm{PG}}$ be the set of potential games in which Player 2 is ϵ-inconsequential to Player 1. There exists a game $G \in \mathcal{G}_{\epsilon}^{\mathrm{PG}}$ such that if Player 1 applies any acceptable evaluator f to Player 2, it holds that*

$$Q^{-}_{\mathrm{ABR}}(G, f) \leq \epsilon. \tag{8.11}$$

That is, losing information about another agent (even an inconsequential one) can have devastating consequences. Note that Theorem 8.3 even allows Player 1 to select the evaluator f *after* the pathological game is realized, indicating that even knowledge about the particular game instance is not sufficient to allow an agent to select an effective evaluator. Theorem 8.3 leads to the following immediate corollary regarding the performance of the class of potential games as a whole:

Corollary 8.4 *For any $\epsilon > 0$, let $\mathcal{G}_{\epsilon}^{\mathrm{PG}}$ be the set of potential games in which Player 2 is ϵ-inconsequential to Player 1. For any acceptable evaluator f that Player 1 applies to*

Player 2, it holds that

$$Q^{-}_{\mathrm{ABR}}\left(\mathcal{G}^{\mathrm{PG}}_{\epsilon}, f\right) \leq \epsilon. \tag{8.12}$$

Note that another corollary of this is that log-linear learning is susceptible in potential games to the same pathologies as asynchronous best-reply processes. In other words, the following corollary is a consequence of the above and (8.4):

Corollary 8.5 *For any $\epsilon > 0$, let $\mathcal{G}^{\mathrm{PG}}_{\epsilon}$ be the set of potential games in which Player 2 is ϵ-inconsequential to Player 1. For any acceptable evaluator f that Player 1 applies to Player 2, it holds that*

$$Q^{-}_{\mathrm{SS}}\left(\mathcal{G}^{\mathrm{PG}}_{\epsilon}, f\right) \leq \epsilon. \tag{8.13}$$

8.2.2 Identical interest games are also susceptible

Why are potential games subject to such severe pathologies as in Theorem 8.3? This is partially because in a potential game, each agent's utility function is only locally aligned with the global welfare function; it may not give the agent any information about the absolute quality of a particular action, and gives the agent no information about the utility functions of other agents. As such, it is relatively easy to construct games in which communication failures cause Player 1 to make potential-decreasing moves while "believing" she is ascending the potential function.

However, in identical interest games, this does not appear to be a concern: each player has access to the full welfare function, and thus knows both the relative quality of each action *and* the utility functions of all other players. Intuitively, it seems that this additional structure may be enough to prevent pathologies. Proposition 8.6 shows that when considering the special case of pure Nash equilibria, identical interest games are indeed immune to the worst of the pathologies of Theorem 8.3, provided that the max evaluator is applied (see Proposition 8.1). However, we also show that this positive result does not take us far:

Proposition 8.6 *For any $\epsilon > 0$, let $\mathcal{G}^{\mathrm{II}}$ be the set of all identical-interest games. For each $G \in \mathcal{G}^{\mathrm{II}}$ let $a^* \in \arg\max_{a\in\mathcal{A}} W(a)$. Let $f_{\max}$ denote the* max *evaluator, and let Player 1 apply $f_{\max}$ to Player 2. Then it is always true that*

$$a^* \in \mathrm{PNE}\left(G_{f_{\max}}\right) \qquad \textit{and} \qquad \mathrm{PNE}\left(G_{f_{\max}}\right) \subseteq \mathrm{PNE}(G). \tag{8.14}$$

The intuition behind (8.14) is simple: the max evaluator can be viewed as an attempt to be optimistic; Player 1 is assuming that Player 2 is maximizing U_1. In an identical interest game $U_1 = U_2$, so this is equivalent to assuming that Player 2 is maximizing her own utility function U_2. Thus, at a pure Nash equilibrium, other players are best-responding to each others' actions – and the optimism of max becomes a self-fulfilling prophecy. However, the positive nature of Proposition 8.6 is tenuous: a game may have many Nash equilibria – and (8.14) gives no guarantee that these are optimal. Furthermore, the second part of (8.14) depends strongly on the assumption that only one communication failure occurs; if Player 2 also applies $f_{\max}$ to Player 1, the resulting reduced game may have many more equilibria that are not present in the nominal game.

8.2.3 Noisy dynamics in identical interest games

Proposition 8.6 showed that when the max evaluator is applied, a reduced identical-interest game always has at least one good pure Nash equilibrium. Is the optimal equilibrium always a stochastically-stable state of log-linear learning? Unfortunately, the answer is no. Theorem 8.7 uses the optimistic form of the quality measure from (8.7); this time, we set $\mathcal{E} = \mathrm{SS}$. That is, for log-linear learning, it evaluates the best stochastically-stable state of the reduced game against the worst stochastically-stable state of the nominal game.

Theorem 8.7 *For any $\epsilon > 0$, let $\mathcal{G}^{\mathrm{II}}_\epsilon$ be the set of identical interest games in which Player 2 is ϵ-inconsequential to Player 1. There exists a game $G \in \mathcal{G}^{\mathrm{II}}_\epsilon$ such that if Player 1 applies* *any acceptable evaluator f to Player 2, it holds that*

$$Q^{-}_{\mathrm{SS}}(G, f) \leq \epsilon. \tag{8.15}$$

Requiring stochastic stability is now too much – and Theorem 8.7 shows that games exist for which even inconsequential communication failures induce stochastically-stable states which are arbitrarily less efficient than those of the nominal game. That is, the high-quality Nash equilibria guaranteed to exist by Proposition 8.6 need not be stochastically stable. Once again, we state the following immediate corollary:

Corollary 8.8 *For any $\epsilon > 0$, let $\mathcal{G}^{\mathrm{II}}_{\epsilon}$ be the set of identical-interest games in which Player 2 is ϵ-inconsequential to Player 1. For every acceptable evaluator f that Player 1 applies to Player 2, it holds that*

$$Q^{-}_{\mathrm{SS}}\left(\mathcal{G}^{\mathrm{II}}_{\epsilon}, f\right) \leq \epsilon. \tag{8.16}$$

8.3 Limiting the harm of communication failures

How can a game designer mitigate the pathologies of the previous sections? Nominally the negative results appear quite formidable, as they appear to rule out several well-behaved classes of games, as well as learning dynamics that are generally thought to provide good efficiency guarantees. In this section we investigate more closely what is causing these pathologies, and show preliminary results on how to avoid them. To do so, we will apply the equilibrium selection properties of log-linear learning.

8.3.1 Resilience to failures via an informational paradox

Here, we investigate the cause of the identical interest pathologies of Theorem 8.7. In the example game enabling (8.15), Player 1 essentially needs a third player's "help" to drive the system to a low-welfare state (see the proof of Theorem 8.7 and Figure 8.3 in Section 8.4). However, in an identical interest game, if Player 2 is inconsequential to Player 1, then

Player 2 is universally inconsequential; that is, also inconsequential to all other players. This suggests that if Player 1 cannot observe Player 2's action, then perhaps other players should not either. Theorem 8.9 confirms this intuition; here we are using the pessimistic form of the quality metric from (8.8), lending strength to our result showing it can be close to 1.

Theorem 8.9 *For all $\epsilon \in [0,1]$, and let $\mathcal{G}_\epsilon^{\mathrm{II}}$ be the set of identical-interest games in which Player 2 is inconsequential to Player 1. Let all players other than Player 2 apply the* max *evaluator $f_{\max}$ to Player 2. Then it holds that*

$$Q_{\mathrm{SS}}^{+}\left(\mathcal{G}_\epsilon^{\mathrm{II}}, f_{\max})\right) = 1 - \epsilon. \tag{8.17}$$

Note that Theorem 8.9 presents a curious paradox when compared with Theorem 8.7, showing that performance *improves* when *less* communication is allowed. That is, if Player 2's action is hidden from *every* player, the performance can be dramatically better than when it is hidden from a *single* player. Though more study is needed, one possible implication of this is as follows: suppose in a multiagent system that Agent A has a high risk of losing communication with Agent B. Theorem 8.9 seems to suggest that if the agents' utility functions define an identical interest game, it could be desirable to preemptively sever communications between Agent B and *all* other agents. Naturally, without a somewhat more detailed investigation, this should not be taken explicitly as design advice – but nonetheless it seems to warrant more research.

8.3.2 Large games are more susceptible

What drives the negative results of Theorem 8.3 for potential games? In the game used to prove the foregoing theorems, the number of action profiles was conditioned on the size of ϵ (see proof of Theorem 8.3 in Section 8.4). When ϵ was very close to 0, the proof of Theorem 8.3 required many action profiles to generate the pathology. Is this simply an artifact of the proof technique, or is it indicative of a deeper principle? Here, we show that there is indeed a connection between the size of a game and the degree to which communication failures can create pathologies. This is because for small ϵ, the game resulting from a bounded evaluator is *close* to the nominal game in a formal sense defined in [51], provided that the number of total action profiles is small – and thus log-linear learning selects action profiles close to the potential-maximizing states of the nominal game.

To show positive results, here we take the pessimistic form of the quality metric (8.8), and let $\mathcal{E} = \mathrm{SS}$. Recall that this compares the worst stochastically-stable state of the reduced game with the best of the nominal game, setting the stage for stronger positive results.

Theorem 8.10 *For any $\epsilon \geq 0$, let $\mathcal{G}_\epsilon^{\mathrm{P}k}$ be the set of potential games in which Player 2 is ϵ-inconsequential to Player 1 and such that $|\mathcal{A}| \leq k$. For every bounded acceptable evaluator f, it holds that*

$$Q_{\mathrm{SS}}^{+}\left(\mathcal{G}_\epsilon^{\mathrm{P}k}, f\right) \geq \max\left\{0, 1 - 8\epsilon(k-1)\right\}. \tag{8.18}$$

That is, if ϵ is small relative to the number of action profiles in G, communication failures cause limited harm.

8.3.3 A safety certificate: coarse potential alignment

Though this work has shown that in general no straightforward method exists for computing good proxy payoffs from the nominal game payoffs, it would be valuable to develop tools which could certify a set of proxy payoffs as safe for a particular setting. In this section we propose one such certificate. This section requires the following definition: a *best-reply path* is a sequence of action profiles $\{a^1, a^2, \ldots, a^m\}$ such that for each $k \in \{1, \ldots, m-1\}$,

i) $a^{k+1} = (a_i, a_{-i}^k)$ for some agent $i \in \mathcal{I}$ with $a_i \neq a_i^k$, and

ii) $a_i^{k+1} \in B_i(a^k)$.

That is, each successive action profile differs from the previous in the action of a single agent, and the updating agent chooses a best response. In a potential game, it is known that all best-reply paths terminate at a pure Nash equilibrium. A *weakly-acyclic game under best replies* is a generalization of a potential game in which for every joint action profile $a \in \mathcal{A}$, there exists a best-reply path $\{a^1, a^2, \ldots, a^m\}$ where $a^1 = a$ and a^m is a Nash equilibrium.

Many simple learning rules are known that converge almost surely to Nash equilibria in weakly-acyclic games under best replies; in particular, this is true of our asynchronous best reply process defined in Section 9.1. Thus, when one of these games has a *unique* pure Nash equilibrium, this equilibrium may be considered highly likely to arise in game play [93,101].

Our certificate pertains to a setting in which the nominal game is a potential game with a unique pure Nash equilibrium. In these games, pathologies could be constructed because a single player's payoffs could be specified so that any proxy payoffs would cause that player to best-respond *against* the potential gradient. To rule out such pathologies, proxy payoffs must be appropriately aligned with the potential function. In the following theorem, we give one such characterization of "appropriately aligned."

Suppose that Player 1 cannot know Player 2's action. Given a potential game G with potential function W, let the *reduced* potential function $\tilde{W}$ be

$$\tilde{W}\left(a_1, a_{-12}\right) = \max_{a_2 \in \mathcal{A}_2} W\left(a_1, a_2, a_{-12}\right). \tag{8.19}$$

That is, $\tilde{W}$ is the nominal potential function with the maximum evaluator applied to Player 2. Using this definition, we can state the following result, which holds even if we do not require inconsequentiality.

Proposition 8.11 *Let G be a potential game with $n \geq 3$ players and a unique pure Nash equilibrium a^*. If Player 1 is assigned proxy payoff functions $\tilde{U}_1 : \mathcal{A}_1 \times \mathcal{A}_{-12} \to \mathbb{R}$ satisfying*

$$\arg\max_{a_1 \in \mathcal{A}_1} \tilde{U}_1\left(a_1, a_{-12}\right) \subseteq \arg\max_{a_1 \in \mathcal{A}_1} \tilde{W}_1\left(a_1, a_{-12}\right), \tag{8.20}$$

then the reduced game $\tilde{G}$ associated with $\tilde{U}$ is weakly-acyclic under best replies and has a unique pure Nash equilibrium a^*.

A consequence of this proposition is that for every potential game with a unique pure Nash equilibrium, there do exist safe proxy payoffs. However, these proxy payoffs may not be computable. That is, a player may not have access to the potential function to be able to compute these proxy payoffs, and in an arbitrary potential game, Player 1's utility function need not contain enough information about the potential function for this to be a valid approach. This highlights a crucial issue: in a potential game, despite the fact that individual players' utility functions are locally aligned with the potential function, even slight perturbations of these utility functions can essentially discard the information that is required to ascend the potential function.

8.3.4 Fine potential alignment

This second certificate, though more difficult to verify in general, is a guarantee for any potential game that under any evaluator, the reduced game is an exact potential game with potential function equal to the nominal potential function. Here, we examine more closely the intuition behind our notion of inconsequentiality. Note that ϵ-inconsequentiality implies the inability to affect another player's payoffs by more than ϵ, but it says nothing about the ability to affect another player's *relative preferences* between pairs of actions. For any $\epsilon > 0$, it is possible to construct games where a player's optimal decision is always conditioned on an "inconsequential" player's action. Thus, inconsequentiality may not adequately capture the coupling between relative preferences and actions.

To see the distinction, consider the following identical-interest game:

		Player 2	
		b_1	b_2
Player 1	a_1	1	ϵ
	a_2	$1-\epsilon$	0

At any action profile, Player 2 can deviate and cause a large change in Player 1's payoff: in our parlance, Player 2 is not at all inconsequential to Player 1. Nonetheless, Player 2's action is irrelevant to Player 1's relative preferences, since Player 1 always values a_1 exactly ϵ more than a_2. Thus, if Player 1 applies any linear acceptable evaluator to Player 2, the resulting proxy payoffs will correctly value a_1 over a_2. This concept is formalized in the following proposition.

Proposition 8.12 *Let G be a potential game with potential function W and $n \geq 2$, and suppose that Player 1 cannot observe Player 2's action. If for each pair of actions of Player 1 (a_1, a_1'), and for each joint action $a_{-12} \in \prod_{i \notin \{1,2\}} \mathcal{A}_i$, there exists a constant $C(a_1, a_1', a_{-12}) \in \mathbb{R}$ such that for all a_2,*

$$U_1\left(a_1, a_2, a_{-12}\right) - U_1\left(a_1', a_2, a_{-12}\right) = C(a_1, a_1', a_{-12}), \tag{8.21}$$

then for any bounded acceptable evaluator f among those specified in Proposition 8.1, G_f is a potential game with potential function W.

Proof: The conditions of Proposition 8.12 guarantee that the payoff gain (or loss) of any deviation by Player 1 in G is not conditioned on the action choice of Player 2. Thus, for any bounded evaluator among those specified in Proposition 8.1, the reduced payoffs resulting from the evaluator will satisfy

$$\tilde{U}_1(a_1, a_{-12}) - \tilde{U}_1(a_1', a_{-12}) = C(a_1, a_1', a_{-12}). \tag{8.22}$$

Since $C(a_1, a_1', a_{-12})$ is equal to Player 1's payoff gain (or loss) in G as well, this means that

G_f must be a potential game with the same potential function as G. ■

This strict condition may be of questionable value in practical settings, particularly when there are a large number of action profiles. In addition, it is not difficult to show that straightforward relaxations of this condition allow for the possibility that pathologies can occur, similar to those studied in Theorems 8.3 and 8.7.

8.4 Chapter Proofs

Proof of Proposition 8.1

To see that each satisfies Definition 8.1, let $S \in \mathbb{R}^k$ and $S' \in \mathbb{R}^k$ satisfy the first assumption of Definition 8.1. Arrange S and S' in ascending order and denote the i-th element of S and S' as s_i and s'_i, respectively so that $\min_{s \in S} = s_1$ and $\max_{s \in S} = s_k$. Thus, $f_{\text{sum}}(S) = \sum_{i=1}^{k} s_i > \sum_{i=1}^{k} s'_i > f_{\text{sum}}(S')$. Since f_{sum} satisfies Definition 8.1, it must be true that f_{mean} does as well. To see that $f_{\max}$ and $f_{\min}$ satisfy Definition 8.1, simply note that $f_{\max}(S) = s_k > s'_k = f_{\max}(S')$ and $f_{\min}(S) = s_1 > s'_1 = f_{\min}(S')$. For these evaluators, the second axiom of Definition 8.1 is obvious. ■

Proof of Theorem 8.3

We will construct a potential game $G \in \mathcal{G}_{\epsilon}^{\text{PG}}$ with a unique Nash equilibrium (and thus a unique efficient recurrent class of the asynchronous best reply process) and show that for any f its reduced variant G_f is a weakly-acyclic game under best replies (see Section 8.3.3 for definition) with a unique Nash equilibrium with welfare within ϵ of 0. Standard results in learning theory then imply that the Nash equilibrium of the reduced game is the unique recurrent class of the asynchronous best reply process, completing the proof.

Our constructed game has 3 players. Let $M \in \mathbb{N}$ be the positive integer satisfying $1/\epsilon - 8 \leq M < 1/\epsilon - 7$.[5] Player 1 has actions $\mathcal{A}_1 = \{0, 1, \ldots, M+1\}$; Player 2 has actions

[5]Strictly, we also require $\epsilon < 1/7$; for larger ϵ, similar examples can be constructed to show the same bound but we omit these for reasons of space and because the result is considerably more interesting for

$\mathcal{A}_2 = \{0, 1, 2\}$; Player 3 has actions $\mathcal{A}_3 = \{0, 1, \ldots, M\}$. The game is built of $M + 1$ two-player games in which the action of Player 3 selects which game is played between Players 1 and 2. We refer to the actions of Players 1, 2, and 3 as "rows," "columns," and "levels," respectively.

The payoff matrix which comprises Level 0 (that is, $a_3 = 0$) is depicted in Figure 8.1. The two matrices in Figure 8.1 depict payoffs and potential values resulting from the actions of Players 1 and 2 when $a_3 = 0$. For each (a_1, a_2), the upper matrix represents the value of Player 1's payoffs $U_1(\cdot)$; the lower matrix depicts the value of the potential function $W(\cdot)$. Let Players 2 and 3 have payoffs equal to W so for any a, $U_2(a) = U_3(a) = W(a)$. For the action profiles not depicted for $a_1 > 2$, the payoffs and potential are equal to those when $a_1 = 2$. That is, for any $m > 2$, $U_1(m, a_2, 0) = U_1(2, a_2, 0)$ and $W(m, a_2, 0) = W(2, a_2, 0)$.

For $a_3 > 0$, consider the matrices in Figure 8.2. Note that these matrices are similar to those for $a_3 = 0$; $u_1(a_1, a_2, k) = u_1(0, a_2, 0)$ for $a_1 \in \{k, k + 1\}$ and $W(a_1, a_2, k) = W(0, a_2, 0) - k\epsilon$. They each contain the additional row $k - 1$, which is simply the payoffs/potential from row k plus $\epsilon/2$. For both , all rows not depicted are identical to row $k + 2$.

In the nominal game, $a^{\mathrm{ne}} = (0, 0, 0)$ is a unique pure Nash equilibrium with potential $W(a^{\mathrm{ne}}) = 1$. This can be proved by induction; the base case is depicted in Figure 8.1, which contains only $(0, 0, 0)$ as a Nash equilibrium. If there is another pure Nash equilibrium, it must be associated with some $a_3 > 0$. For the inductive step, consider $a_3 = k$ as in Figure 8.2. Here, the only possible Nash equilibrium is $(k - 1, 0, k)$. At this action profile, Player 3 has a payoff of $1 + \epsilon/2 - k\epsilon$, but he can deviate to $a_3 = k - 1$ (i.e., the next-lower level) to obtain an improved payoff of $1 - k\epsilon + \epsilon$, so this cannot be a Nash equilibrium. Therefore, $a^{\mathrm{ne}} = (0, 0, 0)$ is unique, so all better-reply paths in G terminate at a^{ne}.

Now, note that Player 2 is 6ϵ-inconsequential to Player 1. Let Player 1 apply an acceptable evaluator f to Player 2 to obtain the reduced game G_f. Considering Fig-

small ϵ.

$U_1(\cdot)$		Player 2		
		0	1	2
Player 1	0	1	$1-3\epsilon$	$1-6\epsilon$
	1	$1-2\epsilon$	$1+\epsilon$	$1-5\epsilon$
	2	0	4ϵ	-2ϵ

$W(\cdot)$	Player 2		
	0	1	2
0	1	$1-7\epsilon$	$1-4\epsilon$
1	$1-2\epsilon$	$1-3\epsilon$	$1-3\epsilon$
2	0	0	0

Figure 8.1: Left: Player 1 payoff function $u_1(\cdot)$ for game used to prove Theorem 8.3. Right: Potential function W values (and Player 2/3 payoffs) for the same game. Both are depicted with Player 3 playing action 0. When Player 1 applies an acceptable evaluator to Player 2, he prefers Action 1 to Action 0. When Player 1 plays action 1, Player 2's best response is to play action 0, making $(1, 0, 0)$ the only pure Nash equilibrium in this simplified game. See Figure 8.2 for a depiction of the game's payoffs for action profiles when Player 3 is playing $a_3 > 0$.

ure 8.2, let S_{k-1}, S_k, S_{k+1}, and S_{k+2} denote the depicted rows in Player 1's payoff matrix. Order each row nondecreasing and match elements so Definition 8.1 implies that $f(S_{k+1}) > f(S_{k-1}) > f(S_k) > f(S_{k+2})$ whenever $M < 1/\epsilon - 3$. Thus, for any $a_2 \in \mathcal{A}_2$, Player 1's reduced payoff function $\tilde{U}_1(a)$ satisfies

$$\tilde{U}_1(k+1, a_2, k) > \tilde{U}_1(k-1, a_2, k) > \tilde{U}_1(k, a_2, k).$$

That is, when Player 3 is playing k, Player 1's best response is $a_1 = k+1$, regardless of the action of Player 2. Then, Player 2's best response to $a_1 = k+1$ is to choose $a_2 = 0$, to obtain the gray-shaded action profile in Figure 8.2.

At this action profile $(k+1, 0, k)$, Player 3 can improve his payoff from $1 - 2\epsilon - k\epsilon$ to $1 - \epsilon - k\epsilon$ by deviating to action $a_3 = k+1$ (i.e., "moving up" one level). Thus, all action profiles have a best-reply path which terminates at the unique Nash equilibrium $\tilde{a}^{\text{ne}} = (M+1, 0, M)$, with potential $W(\tilde{a}^{\text{ne}}) = 1 - 2\epsilon - M\epsilon$. That is, G_f is a weakly-acyclic game under best replies. Because $M \geq 1/\epsilon - 8$,

$$W(\tilde{a}^{\text{ne}}) \leq 1 - 2\epsilon - (1/\epsilon - 8)\,\epsilon = 6\epsilon. \tag{8.23}$$

Since Player 2 is 6ϵ-inconsequential to Player 1, and the potential maximum is 1, the

$U_1(\cdot)$		Player 2		
		0	1	2
Pl. 1	$k-1$	$1+\frac{\epsilon}{2}$	$1-2.5\epsilon$	$1-5.5\epsilon$
	k	1	$1-3\epsilon$	$1-6\epsilon$
	$k+1$	$1-2\epsilon$	$1+\epsilon$	$1-5\epsilon$
	$k+2$	$k\epsilon$	$4\epsilon+k\epsilon$	$-2\epsilon+k\epsilon$

$W(\cdot)$	Player 2		
	0	1	2
$k-1$	$1+\frac{\epsilon}{2}-k\epsilon$	$1-6.5\epsilon-k\epsilon$	$1-3.5\epsilon-k\epsilon$
k	$1-k\epsilon$	$1-7\epsilon-k\epsilon$	$1-4\epsilon-k\epsilon$
$k+1$	$1-2\epsilon-k\epsilon$	$1-3\epsilon-k\epsilon$	$1-3\epsilon-k\epsilon$
$k+2$	0	0	0

Figure 8.2: Left: Player 1 payoff function $U_1(\cdot)$ for game used to prove Theorem 8.3. Right: Potential function W values (and Player 2/3 payoffs) for the same game. Both are depicted with Player 3 playing action $k > 0$. All rows not depicted have the same payoffs (or potential values) as row $(k+2)$. Here, given this action of Player 3, the nominal game has a unique Nash equilibrium at $(k-1, 0)$. When Player 1 applies an acceptable evaluator to Player 2, he prefers Action $k+1$ to Action k; Player 2's best response to this is to play action 0, making the shaded action profile $(k+1, 0, k)$ the only pure Nash equilibrium in this simplified game. Because the full game is staggered, at the action profile $(k+1, 0, k)$, Player 3 has an incentive to switch to action $k+1$.

theorem is proved. ■

Proof of Proposition 8.6

Let G be any identical interest game and let $f = \max$ so that G_f is the reduced variant. Let $a^* \in \arg\max_{a\in\mathcal{A}} W(a)$; this is both the maximum-potential action profile and since the game is identical-interest, $a^* \in \mathrm{PNE}(G)$. First we show that $a^* \in \mathrm{PNE}(G_f)$. Since $W(a^*) \geq W(a)$ for any $a \in \mathcal{A}$, the fact that $f = \max$ implies that for Player 1 and any $a_1 \in \mathcal{A}_1$,

$$\tilde{U}_1\left(a_1^*, a_{-1}^*\right) = W(a^*) \geq \tilde{U}_1\left(a_1, a_{-1}^*\right), \tag{8.24}$$

where the inequality stems from the fact that $W(a^*)$ (and thus $U_1(a^*)$) is maximal with respect to any other action profile. Thus, player 1 cannot unilaterally deviate and improve utility. Since a^* is an equilibrium of G, every player j with is already playing a best response to a^*. Thus, a^* is a pure Nash equilibrium of G_f, implying (8.14).

Next, let $\bar{a} \in \mathrm{PNE}(G_f)$; we wish to show that $\bar{a} \in \mathrm{PNE}(G)$. Since all other players

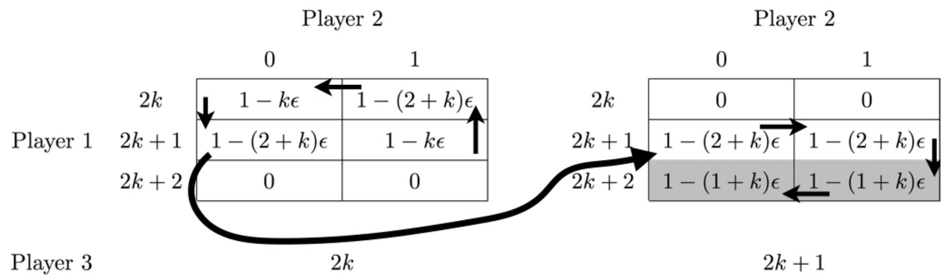

Figure 8.3: Generic modular block for the identical-interest game described in the proof of Theorem 8.7, displaying best-reply path resulting when Player 1 applies any acceptable evaluator to Player 2. If Player 1 applies any acceptable evaluator to Player 2, note that he will be indifferent between actions $2k$ and $2k+1$ when Player 3 is playing $2k$ (i.e., the left matrix). Thus, there is a best-reply path from every action profile depicted here to $\mathcal{A}_k^\dagger$. Once in $\mathcal{A}_k^\dagger$, a best-response process cannot escape except by Player 3 incrementing his action to $2k+2$.

$j \neq 1$ know Player 1's action, they can best-respond with respect to the true W. Thus,

$$\tilde{U}_1(\bar{a}_1, \bar{a}_{-1}) = U_1(\bar{a}_1, \bar{a}_{-1}). \tag{8.25}$$

Since $f = \max$, for all $a_1' \neq \bar{a}_1$ we have

$$\tilde{U}_1(a_1', \bar{a}_{-1}) \geq U_1(a_1', \bar{a}_{-1}). \tag{8.26}$$

Combining (8.25) and (8.26) with the fact that $\bar{a}$ is a Nash equilibrium of G_f, we have for all $a_1' \neq \bar{a}_1$ that

$$U_1(\bar{a}_1, \bar{a}_{-1}) \geq U_1(a_1', \bar{a}_{-1}). \tag{8.27}$$

That is, Player 1 has no incentive to deviate from $\bar{a}$ in the nominal game. Since every other has the same payoffs in the nominal game as in the reduced game, $\bar{a}$ must be an equilibrium of the nominal game as well. ∎

Proof of Theorem 8.7

We shall construct a 3-player game that exhibits the pathology described in Theorem 8.7. For $\epsilon > 0$, let $M \in \mathbb{N}$ be the positive integer satisfying $1/\epsilon - 3 \leq M < 1/\epsilon - 2$.[6] In this game, $\mathcal{A}_1 = \{0, 1, \ldots, 2M+2\}$, $\mathcal{A}_2 = \{0, 1\}$, and $\mathcal{A}_3 = \{0, 1, \ldots, 2M+1\}$.

The game is comprised of M fundamental blocks indexed by $k \leq M$, a generic one of which is depicted in Figure 8.3. Given a block k, Player 3 has two actions: $2k$ and $2k+1$, depicted in Figure 8.3 as the left and right payoff matrices, respectively. Note that in each block, the potential-maximizing states are both in the left matrix: $(2k, 0, 2k)$ and $(2k+1, 1, 2k)$, each with a potential of $1 - k\epsilon$. Since these states' potential are maximal when $k = 0$, the potential-maximizing states of the overall game are $(0, 0, 0)$ and $(1, 1, 0)$ with potential of 1. Thus, these are also the only stochastically-stable states of log-linear learning in the nominal game.

Player 2 is 2ϵ-inconsequential to Player 1; let Player 1 apply any acceptable evaluator f to Player 2. Now, when $a_3 = 2k$ (left matrix), Definition (8.1) implies that $f(1 - k\epsilon, 1 - (2+k)\epsilon) = f(1 - (2+k)\epsilon, 1 - k\epsilon)$, so Player 1 receives equal payoff for actions $2k$ and $2k+1$. Nevertheless, when $a_3 = 2k+1$ (right matrix), Player 1 still strictly prefers action $2k+2$ to action $2k+1$. To prove Theorem 8.7, we will show that for any $k \leq M$, there is a best-reply path from every action profile to the gray-shaded action profiles in Figure 8.3, but that the only best-reply path *leaving* those action profiles has Player 3 incrementing his action to $2k+2$ (thus pushing the system state into the next-higher block, where the process repeats).

Let $k < M$, and let the gray-shaded action profiles in Figure 8.3 be denoted $\mathcal{A}_k^\dagger = \{(2k+2, 0, 2k+1),\ (2k+2, 1, 2k+1)\}$. Let a be an action profile such that $a_3 \in \{2k, 2k+1\}$ (i.e., a is in block k). If $W(a) = 0$, note that $\mathcal{A}_k^\dagger$ can be reached in one step; either by Player 3 deviating to action $2k+1$ or by Player 1 deviating to action $2k+2$. On the other

[6] We also require $\epsilon < 1/3$. Note that this is without loss of generality: when $\epsilon' > \epsilon$, it is true that ϵ'-inconsequentiality implies ϵ-inconsequentiality.

hand, if $W(a) > 0$, then a lies on the best-reply path depicted in Figure 8.3. That is, for every a, there is a best-reply path from a to $\mathcal{A}_k^\dagger$.

Next, note that a deviation by Player 2 cannot escape $\mathcal{A}_k^\dagger$, and Player 1 strictly prefers the states in $\mathcal{A}_k^\dagger$ to any other states that he can reach with a single deviation. Thus, the only way for a best-reply path to escape $\mathcal{A}_k^\dagger$ is for Player 3 to increment his action to $a_3 = 2k+2$; when $k < M$, this can occur when the state is $(2k+1, 1, 2k) \in \mathcal{A}_k^\dagger$.

Thus, for every action profile in block $k < M$, there is a best-reply path leading to an action profile in block $k+1$. When $k = M$, there is a best-reply path leading to an action profile in $\mathcal{A}_M^\dagger$, but no best-reply path *leaves* $\mathcal{A}_M^\dagger$. This implies that $\mathrm{ABR}(G) = \mathcal{A}_M^\dagger$. We then apply (8.4) to obtain $\mathrm{SS}(G) \subseteq \mathcal{A}_M^\dagger$. By the definition of M, it follows for any $a \in \mathrm{SS}(G) \subseteq \mathcal{A}_M^\dagger$ that $W(a) \leq 2\epsilon$, proving the theorem (since Player 2 is 2ϵ-inconsequential).■

Proof of Theorem 8.9

Suppose that Player 2 is ϵ-inconsequential to Player 1; because this is an identical interest game, this implies that Player 2 is ϵ-inconsequential to all players $\mathcal{I} \setminus \{2\}$. If all players other than 2 ignore the actions of Player 2 by applying an acceptable evaluator f to their utility functions, then the action choice of Player 2 has no effect on the decisions of players other than 2. This means that the reduced game can be analyzed as an $(n-1)$-player identical interest game, and Player 2 can be modeled as a simple optimizer.

Let all players $\mathcal{I}\setminus\{2\}$ apply the maximum evaluator $f = \max$, so that the reduced payoff functions $\tilde{U}$ and corresponding reduced potential function $\tilde{W}$ for each player $i \in \mathcal{I} \setminus \{2\}$ are given by $\tilde{W}(a) = \tilde{U}_i(a) = \max_{a_2 \in \mathcal{A}_2} U_i\left(a_{-2}, a_2\right)$. The reduced game, denoted $\tilde{G}$, will be an identical-interest game played between all players in $\mathcal{I} \setminus \{2\}$.

Let $\tilde{\mathcal{A}}^* = \arg\max_{a \in \mathcal{A}} \tilde{W}(a)$ be the set of potential-maximizing states of $\tilde{G}$; because $\tilde{G}$ is an identical-interest game, $\tilde{\mathcal{A}}^*$ must also be the set of stochastically-stable states of log-linear learning. All stochastically-stable states for the reduced game G_f must also be stochastically-stable for $\tilde{G}$; thus, to prove the theorem, it will suffice to establish a tight lower bound on potential for states in $\tilde{\mathcal{A}}^*$.

Since the players are applying the maximum evaluator, it must be the case for any state $\tilde{a} \in \tilde{\mathcal{A}}^*$ that there exists some a_2 such that $W(a_2, \tilde{a}_{-2}) = 1$. Thus, ϵ-inconsequentiality provides the lower bound of $W(\tilde{a}) \geq W(a_2, \tilde{a}_{-2}) - \epsilon$. The game in Figure 8.3 with $k = 0$ shows the bound to be tight. ■

Proof of Theorem 8.10

The proof of Theorem 8.10 relies on the following definition:

Definition 8.3 ([51]) *Let G and $\hat{G}$ be two games with players $\mathcal{I}$, action set $\mathcal{A}$, and utility functions $\{U_i\}_{i\in\mathcal{I}}$ and $\{\hat{U}_i\}_{i\in\mathcal{I}}$ respectively. Let $\Delta_i(G, \hat{G}, a, a_i')$ denote*

$$\left| \left(U_i(a_i, a_{-i}) - U_i(a_i', a_{-i})\right) - \left(\hat{U}_i(a_i, a_{-i}) - \hat{U}_i(a_i', a_{-i})\right) \right|. \tag{8.28}$$

The maximum pairwise difference (MPD) *between G and $\hat{G}$ is defined as*

$$d(G, \hat{G}) \triangleq \max_{a\in\mathcal{A},\ i\in\mathcal{I},\ a_i'\in\mathcal{A}_i} \Delta_i(G, \hat{G}, a, a_i') \tag{8.29}$$

Let $G \in \mathcal{G}_\epsilon^{\mathrm{P}k}$, and let f be a bounded acceptable evaluator f generating reduced utility function $\tilde{U}_1$. Let $U_1^*(a_1) := U_1(a_1, a_{-1})$. For any $a \in \mathcal{A}$ and $a_i' \in \mathcal{A}_i$, we have that

$$\begin{aligned} \Delta_i(G, G_f, a, a_i') &= \left| U_i^*(a_i) - U_i^*(a_i') - \tilde{U}_i^*(a_i) + \tilde{U}_i^*(a_i') \right| \\ &\leq \left| U_i^*(a_i) - \tilde{U}_i^*(a_i) \right| + \left| \tilde{U}_i^*(a_i') - U_i^*(a_i') \right| \leq \epsilon + \epsilon. \end{aligned}$$

The first inequality is the triangle inequality; the second follows from the ϵ-inconsequentiality of Player 2 and the boundedness of f, implying that $d(G, G_f) \leq 2\epsilon$. Corollary 4.3 of [51] states that the stochastically stable states of $\hat{G}$ under log-linear learning are within $4d(G, \hat{G})(|\mathcal{A}| - 1)$ of the potential function maxima of G; applying this completes the proof of the theorem.

■

Proof of Proposition 8.11

Let G and $\tilde{G}$ satisfy the assumptions of Proposition 8.11, and let $a \in \mathcal{A}$ be any action profile. To show that $\tilde{G}$ is weakly acyclic under best replies, it suffices to construct a best-reply path from a to a^* for $\tilde{G}$. Denote this best-reply path by $\{a^1, a^2, \ldots, a^m\}$, where $a^1 = a$.

First, let Player 1 choose a best response; second, let Player 2 choose a best response.[7] Thus, by the definition of $\tilde{U}_1$, a^3 satisfies

$$\left(a_1^3, a_2^3\right) \in \arg\max_{a_1,\ a_2} W\left(a_1, a_2, a_{-12}^1\right), \tag{8.30}$$

and it must be true that $W(a^3) > W(a)$.

If a^3 is a pure Nash equilibrium, we are done. Otherwise, there exists a player $j \in \{3, \ldots, n\}$ that can strictly improve its payoff with a best reply; such a deviation strictly increases the value of the potential function. In this case, let player j deviate, and then repeat the process with Players 1 and 2 as before. In this way, since there are a finite number of action profiles and the potential function is strictly increasing along this best-reply path, it can easily be seen that a best-reply path can be found from a to some Nash equilibrium a^m of $\tilde{G}$.

Since a^m satisfies (8.30), it must also be a Nash equilibrium of G; since G has a unique equilibrium, it must be that $a^m = a^*$ and the proof is obtained. ∎

[7] Assume without loss of generality that at least one of Player 1 and Player 2 are not already playing a best response when it is their chance to deviate.

Chapter 9

Robust Influence by an Adversary

We use graphical coordination games, introduced in [105, 106], to study the impact of adversarial manipulation. The foundation of a graphical coordination game is a simple two agent coordination game, where each agent must choose one of two alternatives, $\{x, y\}$, with payoffs depicted by the following payoff matrix which we denote by $u(\cdot)$:

$$\begin{array}{c|c|c|} & x & y \\ \hline x & 1+\alpha,\, 1+\alpha & 0,\, 0 \\ \hline y & 0,\, 0 & 1,\, 1 \\ \hline \end{array} \tag{9.1}$$

Both agents prefer to agree on a convention, i.e., (x, x) or (y, y), than disagree, i.e., (x, y) or (y, x), with a preference for agreeing on (x, x). The parameter $\alpha > 0$ indicates that (x, x) has an intrinsic advantage has over (y, y); we refer to α as the *payoff gain*. Nonetheless, unilaterally deviating from (y, y) for an individual agent incurs an immediate payoff loss of 1 to 0; hence, myopic agents may be reluctant to deviate, stabilizing the inefficient equilibrium (y, y).

This two player coordination game can be extended to an n-player *graphical coordination game* [107–109], where the interactions between the agents $N = \{1, 2, \ldots, n\}$ are described

This chapter is adapted, with permission, from [104], scheduled to be published by IEEE Transactions on Network Control.

by an undirected graph $G = (N, E)$, where an edge $(i, j) \in E$ for some $i \neq j$ indicates that agent i is playing the two-player coordination game (9.1) with agent j. An agent's total payoff is the sum of payoffs it receives in the two-player games played with its neighbors $\mathcal{N}_i = \{j \in N : (i, j) \in E\}$, i.e., for a joint action $a = (a_1, \ldots, a_n) \in \{x, y\}^n$, the utility function of agent i is

$$U_i(a_1, \ldots, a_n) = \sum_{j \in \mathcal{N}_i} u(a_i, a_j), \tag{9.2}$$

where $u(\cdot)$ is chosen according to payoff matrix (9.1). Joint actions $\vec{x} := (x, x, \ldots, x)$ and $\vec{y} := (y, y, \ldots, y)$, where either all players choose x or all players choose y, are Nash equilibria of the game for any graph; other equilibria may also exist depending on the structure of graph G.

The system operator's goal is to endow agents with decision-making rules to ensure that for any realized graph G and payoff gain α, the emergent behavior maximizes the sum of agents' utilities. Log-linear learning [91, 110] is one distributed decision making rule that selects the efficient equilibrium in this setting.

We assume that an adversary wishes to influence agents to play the less efficient Nash equilibrium $\vec{y}$. We model the adversary as additional nodes/edges in the graph, where the new node plays a fixed action y in an effort to influence agents' utilities and provide them an additional incentive to play y. Thus, the adversary's core problem can be thought of as analogous to more general incentive design problems.

We investigate the tradeoffs between the amount of information available to an adversary, the policies at the adversary's disposal, and the adversary's resulting ability to stabilize the alternative Nash equilibrium $\vec{y}$. We perform this analysis by specifying three distinct styles of adversarial behavior:

- *Uniformly Random:* This type of adversary influences a random subset of agents at each time step. Uniformly random adversaries have the least amount of information available, essentially only requiring knowledge of n.

- *Fixed Intelligent*: This type of adversary chooses a subset of agents to influence; this subset is fixed for all time. Fixed intelligent adversaries know the graph structure G.

- *Mobile Intelligent*: This type of adversary can choose which agents to influence as a function of the current joint action profile. Thus, mobile intelligent adversaries know the graph structure, and at each time step, the action choices of all agents.

Our results include an initial study on the influence of uniformly random and fixed intelligent agents on general graphs, as well as a complete characterization of each adversary's ability to stabilize $\vec{y}$ in a ring graph.

9.1 Model

9.1.1 Model of agent behavior

Suppose agents in N interact according to the graphical coordination game above, specified by the tuple (G, α), with underlying graph $G = (N, E)$, alternatives $\{x, y\}$, and payoff gain $\alpha \in \mathbb{R}$. We denote the joint action space by $\mathcal{A} = \{x, y\}^n$, and we write $(a_i, a_{-i}) = (a_1, a_2, \ldots, a_i, \ldots, a_n) \in \mathcal{A}$ when considering agent i's action separately from other agents' actions.

A special type of graph considered in some of our results is a *ring* graph, defined as follows. Let $G = (N, E)$ with $N = \{1, 2, \ldots, n\}$ and $E = \{\{i, j\} : j = i + 1 \bmod n\}$, i.e., G is a ring (or cycle) with n nodes.[1] We denote the set of all ring graphs by $\mathcal{G}^{\mathrm{r}}$.

Now, suppose agents in N update their actions according to the *log-linear learning* algorithm as specified in Section 8.1.3 in Chapter 8. We take *strict stochastic stability* as our solution concept of interest, defined in Section 8.1.3.

Joint action $\vec{x}$ is strictly stochastically stable under log-linear learning for any graphical coordination game whenever $\alpha > 0$ [91]. We will investigate conditions when an adversary can destabilize $\vec{x}$ and stabilize the alternative coordinated equilibrium $\vec{y}$.

[1]When considering ring graphs, all addition and subtraction on node indices is assumed to be mod n.

9.1.2 Model of adversarial influence

Consider the situation where agents in N interact according to the graph G and update their actions according to log-linear learning, and an adversary seeks to convert as many agents in N to play action y as possible. At each time $t \in \mathbb{N}$ the adversary influences a set of agents $S(t) \subseteq N$ by posing as a friendly agent who always plays action y. Agents' utilities, $\tilde{U} : \mathcal{A} \times 2^N \to \mathbb{R}$, are now a function of adversarial and friendly behavior, defined by:

$$\tilde{U}_i((a_i, a_{-i}), S) = \begin{cases} U_i(a_i, a_{-i}) & \text{if } i \notin S \\ U_i(a_i, a_{-i}) & \text{if } a_i = x \\ U_i(a_i, a_{-i}) + 1 & \text{if } i \in S, a_i = y \end{cases} \tag{9.3}$$

where $(a_i, a_{-i}) \in \mathcal{A}$ represents friendly agents' joint action, and *influence set* $S \subseteq N$ represents the set of agents influenced by the adversary. If $i \in S(t)$, agent i receives an additional payoff of 1 for coordinating with the adversary at action y at time $t \in \mathbb{N}$; that is, to agents in $S(t)$, the adversary appears to be a neighbor playing action y. By posing as a player in the game, the adversary can manipulate the utilities of agents belonging to S, providing an extra incentive to choose the inferior alternative y.

Throughout, we write k to denote the number of friendly agents the adversary can connect to, called the adversary's *capability*. Given k, $\mathcal{S}_k := \left\{S \in 2^N, |S| = k\right\}$ denotes the set of all possible influence sets. In this chapter, we consider three distinct models of adversarial behavior, which we term fixed intelligent (FI), mobile intelligent (MI), and uniformly random (UR). To denote a situation in which influence sets $S(t)$ are chosen by an adversary of $\textbf{type} \in \{\text{FI}, \text{MI}, \text{UR}\}$ for a given k, we write $S \in \textbf{type}(k)$.

If an adversary is fixed intelligent (FI), this means that the influence set S is a function only of the graph structure and α. That is, the adversary must commit to an influence set S that is fixed for all time (in the following, note that S is always implicitly assumed to be a function of G):

$$S \in \mathrm{FI}(k) \implies S(t) = S. \tag{9.4}$$

If an adversary is mobile intelligent (MI), this means that the influence set $S(a)$ is a function of the graph structure, α, *and* $a(t)$, the state at time t:

$$S \in \mathrm{MI}(k) \implies S(t) = S(a(t)). \tag{9.5}$$

Note that type-MI adversaries have the freedom to choose a mapping $S : \mathcal{A} \to \mathcal{S}_k$, whereas a type-FI adversary must choose that mapping to be a constant function of state a. Finally, if the adversary is uniformly random (UR), the influence set S at each time t is chosen uniformly at random from $\mathcal{S}_k$, independently across time:

$$S \in \mathrm{UR}(k) \implies S(t) \sim \mathrm{unif}\{\mathcal{S}_k\}. \tag{9.6}$$

9.1.3 Susceptibility

Given nominal game (G, α), adversary policy $S \in \mathbf{type}(k)$ for some **type** and $k \geq 1$, we write the set of stochastically stable states associated with log-linear learning as

$$\mathrm{SS}\left(G, \alpha, S\right). \tag{9.7}$$

We say a game (G, α) is *susceptible* to adversarial influence of $\mathbf{type}(k)$ if there exists a policy $S \in \mathbf{type}(k)$ for which $SS(G, \alpha, S) = \vec{y}$. A quantity measuring the susceptibility of a particular graph structure G is then

$$\alpha^{\mathrm{sus}}(G, \mathbf{type}(k)) \triangleq \sup\left\{\alpha : (G, \alpha) \text{ is suscep. to } \mathbf{type}(k)\right\}, \tag{9.8}$$

so that whenever $\alpha < \alpha^{\mathrm{sus}}(G, \mathbf{type}(k))$, then by employing the right policy, a $\mathbf{type}(k)$-adversary can ensure that $\vec{y}$ is strictly stochastically stable.

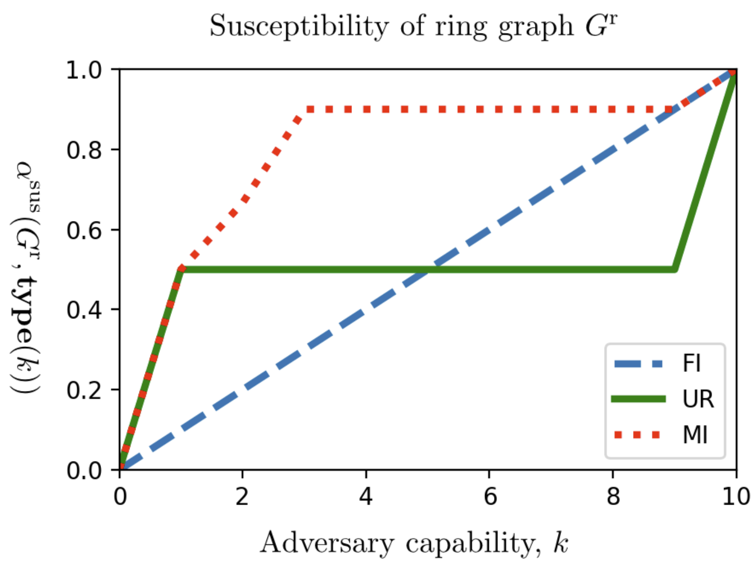

Figure 9.1: Values of α below which each type of adversary can stabilize joint action $\vec{y}$ in a 10-agent ring graph as a function of adversary capability k. For uniformly random (UR) adversaries, this shows that $k = 1$ yields the same susceptibility as $k = n - 1$, but that for the intelligent types of adversaries, susceptibility depends strongly on k.

9.2 Comparing Adversary Models

9.2.1 Fixed intelligent adversarial influence

In the fixed intelligent model of adversarial behavior with capability k, the adversary chooses a fixed subset $S \in \mathcal{S}_k$ of agents to influence, as in (9.4). In a sense, this is the type of adversary that is most limited, as the adversary has no ability to react to changing conditions as the agents update their actions. Nonetheless, as can be seen from Figure 9.1, fixed intelligent adversaries actually can outperform uniformly-random adversaries if k is sufficiently large.

Define

$$[i, j] := \{i, i + 1, \ldots, j\} \subseteq N,$$

and recall that in ring graphs, all addition and subtraction on node indices is assumed to be mod n. Theorem 9.1 gives the susceptibility of any ring graph influenced by a fixed intelligent adversary.

Theorem 9.1 *Let $G^{\mathrm{r}} \in \mathcal{G}^{\mathrm{r}}$ be a ring graph that is influenced by a fixed intelligent adversary with capability $k \leq n$. Then*

$$\alpha^{\mathrm{sus}}\left(G^{\mathrm{r}}, \mathrm{FI}(k)\right) = \frac{k}{n}. \tag{9.9}$$

This can be realized by an adversary distributing its influence set S as evenly as possible around the ring, so that

$$|S \cap [i, i+t]| \leq \left\lceil \frac{kt}{n} \right\rceil \tag{9.10}$$

for any set of nodes $[i, i+t] \subseteq N$, with $i \in N$, $t \leq n$.

The proof of Theorem 9.1 is in Section 9.3.2.

9.2.2 Uniformly random adversarial influence

General graphs

It may be difficult to characterize the exact susceptibility of an arbitrary graph to random adversarial influence, but the following theorem gives an important piece of the puzzle. Here, we show perhaps counterintuitively that that the susceptibility of every graph to a uniformly random adversary is independent of the adversary's capability k.

Theorem 9.2 *Let G be any graph. For any $k \in \{2, \ldots, n-1\}$,*

$$\alpha^{\mathrm{sus}}\left(G, \mathrm{UR}(k)\right) = \alpha^{\mathrm{sus}}\left(G, \mathrm{UR}(1)\right). \tag{9.11}$$

The proof of Theorem 9.2 appears in the Section. This result means that each graph has a universal threshold such that if α falls below this, then even a single uniformly-random adversary will eventually influence all agents to play $\vec{y}$. Note that larger k likely allows the adversary to achieve $\vec{y}$ *more quickly,* but that the threshold value itself for α is independent of k.

Exact susceptibility for ring graphs

When more graph structure is known, it may be possible to derive precise expressions for α^{sus}. In this section, we consider an adversary which influences a ring graph uniformly at random according to (9.6).

Theorem 9.3 *Let $G^{\text{r}} \in \mathcal{G}^{\text{r}}$ be a ring graph that is influenced by a uniformly random adversary with capability $k \leq n$. The susceptibility is given by*

$$\alpha^{\text{sus}}\left(G^{\text{r}}, \text{UR}(k)\right) = \begin{cases} \frac{1}{2} & \textit{if } k \in \{1, \ldots, n-1\}, \\ 1 & \textit{if } k = n. \end{cases} \tag{9.12}$$

Theorem 9.3 is proved in Section 9.3.4.

Consider Theorems 9.1 and 9.3 from the point of view of an adversary, and suppose that an adversary cannot choose k, but can choose whether to be fixed intelligent or uniformly random. Theorems 9.1 and 9.3 suggest that if the adversary's capability is low, it is better to employ a uniformly-random strategy than a fixed one; on the other hand, the conclusion is reversed if the capability is high.

9.2.3 Mobile intelligent adversarial influence on ring graphs

Finally we consider type MI(k) adversaries on ring graphs. Recall that mobile intelligent adversaries choose influence set S as a function of the current state; thus they are always at least as effective as fixed intelligent adversaries (since any fixed influence set can be implemented as a mobile adversary's policy). However, it is not clear *a priori* how mobile intelligent adversaries will compare to those of the uniformly random variety. In this section, we show in Theorem 9.4 that there exist policies which allow mobile intelligent adversaries to render $\vec{y}$ strictly stochastically stable *much* more easily than the other types, even for relatively low values of k.

Theorem 9.4 *Let $G^{\mathrm{r}} \in \mathcal{G}^{\mathrm{r}}$ be a ring graph that is influenced by a mobile intelligent adversary with capability $k \leq n$. Then*

$$\alpha^{\mathrm{sus}}\left(G^{\mathrm{r}}, \mathrm{MI}(k)\right) = \begin{cases} \frac{k}{k+1} & \text{if } k \in \{1, 2\}, \\ \frac{n-1}{n} & \text{if } k \in \{3, \ldots, n-1\}, \\ 1 & \text{if } k = n. \end{cases} \tag{9.13}$$

The proof of Theorem 9.4 is included in Section 9.3.5; this section also presents in Definition 9.2 a family of adversary policies which realize the susceptibilities given in Theorem 9.4. Recall that a uniformly random adversary with $k \geq 1$ can stabilize $\vec{y}$ any time $\alpha < 1/2$; an adversary who can intelligently influence a different single agent in N each day can stabilize $\vec{y}$ under these same conditions. If the mobile intelligent adversary has capability $k \geq 3$, it can stabilize $\vec{y}$ when $\alpha < (n-1)/n$, i.e., under the same conditions as a fixed intelligent adversary with capability $k = n - 1$.

9.3 Chapter Proofs

9.3.1 Log-linear learning and its underlying Markov process

For each model of adversarial behavior, log-linear learning dynamics define a Markov chain, P_β over state space $\mathcal{A}$ with transition probabilities parameterized by $\beta > 0$ [91]. These transition probabilities can readily be computed according to the definition of log-linear learning given in Section 8.1.3 in Chapter 8, taking into account the specifics of the adversarial model in question. Since P_β is aperiodic and irreducible for any $\beta > 0$, it has a unique stationary distribution, π_β, with $\pi_\beta P_\beta = \pi_\beta$.

As $\beta \to \infty$, this converges to a unique limiting distribution $\pi := \lim_{\beta \to \infty} \pi_\beta$. If $\pi(a) = 1$, then joint action a is *strictly stochastically stable* [111].

As $\beta \to \infty$, transition probabilities $P_\beta(a \to a')$ of log-linear learning converge to the transition probabilities, $P(a \to a')$, of a best response process. Distribution π is one of possibly multiple stationary distributions of a best response process over game G.

9.3.2 Stability in the presence of a fixed intelligent adversary

When a fixed intelligent adversary influences set S, the corresponding influenced graphical coordination game is a potential game [100]. Given action profile a, define $S_y(a) = \{j \in S : a_j = y\}$ to denote the set of influenced agents who are playing y in a. Then this game has potential function

$$\Phi^S(a_i, a_{-i}) = \frac{1}{2} \sum_{i \in N} \left(U_i(a_i, a_{-i}) + 2 \cdot \mathbb{1}_{S_y(a)}(i) \right). \tag{9.14}$$

It is well known that $a \in \mathcal{A}$ is strictly stochastically stable if and only if $\Phi^S(a) > \Phi^S(a')$ for all $a' \in \mathcal{A}$, $a' \neq a$ [91].

Proof of Theorem 9.1: Suppose $\alpha < k \,/\, n$. Then

$$\Phi^S(\vec{y}) = n + k > n + \alpha n = \Phi^S(\vec{x})$$

for any $S \subseteq N$ with $|S| = k$. Then, to show that $\vec{y}$ is stochastically stable for influenced set S satisfying (9.10), it remains to show that $\Phi^S(\vec{y}) > \Phi^S(\vec{y}_T, \vec{x}_{N \setminus T})$ for any $T \subset N$ with $T \neq \emptyset$ and $T \neq N$. Suppose the graph restricted to set T has p components, where $p \geq 1$. Label these components as $T_1, T_2, \ldots, T_p$ and define $t := |T|$ and $t_i := |T_i|$. For any $T \subset N$

with $T \neq N, T \neq \emptyset$, and $0 < t < n$,

$$\begin{aligned}\Phi^S(\vec{y}_T, \vec{x}_{N\setminus T}) &= (1+\alpha)(n-t-p) + t - p + \sum_{j=1}^{p} |S \cap T_j| \\ &< n + k \\ &= \Phi^S(\vec{y}),\end{aligned}$$

where the inequality follows from $\alpha < k/n$ and (9.10) combined with the fact that $\lceil z \rceil \leq z+1$ for any $z \in \mathbb{R}$. That is, $\alpha^{\text{sus}}(G, \text{FI}(k)) \geq k/n$. The matching upper bound is given by manipulations to Theorem 3 in [112] by selecting $T = \emptyset$. ■

9.3.3 Resistance trees for stochastic stability analysis

When graphical coordination game G is influenced by a uniformly-random adversary, it is no longer a potential game; resistance tree tools defined in this section enable us to determine stochastically stable states for uniformly-random and mobile intelligent adversaries.

The Markov process P_β defined by log-linear learning dynamics over a normal form game is a *regular perturbation* of a best response process. In particular, log-linear learning is a regular perturbation of the best response process defined in Section 9.3.1, where the size of the perturbation is parameterized by $\epsilon = e^{-\beta}$. The following definitions and analysis techniques are taken from [93].

Definition 9.1 (Regular Perturbed Process [93]) *A Markov process with transition matrix M_ϵ defined over state space Ω and parameterized by perturbation $\epsilon \in (0, a]$ for some $a > 0$ is a* regular perturbation *of the process M_0 if it satisfies:*

1. *M_ϵ is aperiodic and irreducible for all $\epsilon \in (0, a]$.*

2. *$\lim_{\epsilon \to 0^+} M_\epsilon(\xi, \xi') \to M(\xi, \xi')$ for all $\xi, \xi' \in \Omega$.*

3. *If $M_\epsilon(\xi,\xi') > 0$ for some $\epsilon \in (0,a]$ then there exists $r(\xi,\xi')$ such that*

$$0 < \lim_{\epsilon \to 0^+} \frac{M_\epsilon(\xi,\xi')}{\epsilon^{r(\xi,\xi')}} < \infty, \tag{9.15}$$

where $r(\xi,\xi')$ is referred to as the resistance *of transition $\xi \to \xi'$.*

Let Markov process M_ϵ be a regular perturbation of process M_0 over state space Ω, where perturbations are parameterized by $\epsilon \in (0,a]$ for some $a > 0$. Define graph $\Gamma = (\Omega, \mathcal{E})$ to be the directed graph with $(\xi,\xi') \in \mathcal{E} \iff M_\epsilon(\xi,\xi') > 0$ for some $\epsilon \in (0,a]$. Edge $(\xi,\xi') \in \mathcal{E}$ is weighted by the resistance $r(\xi,\xi')$ defined in (9.15). The resistance of path $p = (e_1, e_2, \ldots, e_k)$ is the sum of the resistances of the associated state transitions:

$$r(p) := \sum_{\ell=1}^{k} r(e_\ell). \tag{9.16}$$

Now let $\Omega_1, \Omega_2, \ldots, \Omega_m$ denote the $m \geq 1$ recurrent classes of process M_0. In graph G, these classes satisfy:

1. For all $\xi \in \Omega$, there is a zero resistance path in Γ from ξ to Ω_i for some $i \in \{1,2,\ldots,m\}$.
2. For all $i \in \{1,2,\ldots,m\}$ and all $\xi,\xi' \in \Omega_i$ there exists a zero resistance path in Γ from ξ to ξ' and from ξ' to ξ.
3. For all ξ,ξ' with $\xi \in \Omega_i$ for some $i \in \{1,2,\ldots,m\}$, and $\xi' \notin \Omega_i$, $r(\xi,\xi') > 0$.

Define a second directed graph $\mathcal{G} = (\{\Omega_i\}, \mathcal{E})$ over the m recurrent classes in Ω. This is a complete graph; i.e., $(i,j) \in \mathcal{E}$ for all $i,j \in \{1,2,\ldots,m\}$, $i \neq j$. Edge (i,j) is weighted by $R(i,j)$, the total resistance of the lowest resistance path in Γ starting in Ω_i and ending in Ω_j:

$$R(i,j) := \min_{i \in \Omega_i, j \in \Omega_j} \min_{p \in \mathcal{P}(i \to j)} r(p), \tag{9.17}$$

where $\mathcal{P}(i \to j)$ denotes the set of all simple paths in Γ beginning at i and ending at j.

Let $\mathcal{T}_i$ be the set of all spanning trees of $\mathcal{G}$ rooted at i. Denote the *resistance* of tree $T \in \mathcal{T}_i$ by $R(T) := \sum_{e \in T} R(e)$, and define

$$\gamma_i := \min_{T \in \mathcal{T}_i} R(T) \tag{9.18}$$

to be the *stochastic potential* of Ω_i. We use the following theorem due to [93] in our analysis:

Theorem 9.5 (From [93]) *State $\xi \in \Omega$ is stochastically stable if and only if $\xi \in \Omega_i$ where*

$$\gamma_i = \min_{j \in \{1,2,\dots,m\}} \gamma_j, \tag{9.19}$$

i.e., x belongs to a recurrent class with minimal stochastic potential. Furthermore, ξ is strictly stochastically stable if and only if $\Omega_i = \{\xi\}$ and $\gamma_i < \gamma_j, \quad \forall j \neq i$.

9.3.4 Stability in the presence of a uniformly random adversary

The following lemma applies to any graphical coordination game in the presence of a uniformly random adversary with capability $k \leq n-1$. It states that a uniformly random adversary decreases the resistance of transitions when an agent in N changes its action from x to y, but does not change the resistance of transitions in the opposite direction. Intuitively, this means that viewed through the lens of transition resistances, a uniformly-random adversary spreads y throughout the network optimally, but cannot slow the spread of x.

Lemma 9.5.1 *Suppose agents in N update their actions according to log-linear learning in the presence of a uniformly random adversary with capability k, where $1 \leq k \leq n-1$. Then the resistance of a transition where agent $i \in N$ changes its action from x to y is:*

$$r((x, a_{-i}) \to (y, a_{-i}) = \max \{U_i(x, a_{-i}) - U_i(y, a_{-i}) - 1, 0\} \tag{9.20}$$

and the resistance of a transition where agent $i \in N$ changes its action from y to x is:

$$r((y, a_{-i}) \to (x, a_{-i})) = \max\left\{U_i(y, a_{-i}) - U_i(x, a_{-i}), 0\right\}. \tag{9.21}$$

Here $U_i : \mathcal{A} \to \mathbb{R}$, defined in (9.2), is the utility function for agent i in the uninfluenced game, G.

Proof: In the presence of a uniformly random adversary,

$$P_\beta\left((x, a_{-i}) \to (y, a_{-i})\right) = \frac{1}{n}\left(\frac{k}{n} \cdot \frac{\exp(\beta(U_i(y, a_{-i}) + 1))}{\exp(\beta(U_i(y, a_{-i}) + 1)) + \exp(\beta U_i(x, a_{-i}))}\right.$$
$$\left. + \frac{n-k}{n} \cdot \frac{\exp(\beta U_i(y, a_{-i}))}{\exp(\beta U_i(y, a_{-i})) + \exp(\beta U_i(x, a_{-i}))}\right)$$

Define $P_\epsilon\left((x, a_{-i}) \to (y, a_{-i})\right)$ by making the substitution $\epsilon = e^{-\beta}$. Then, in reference to (9.15), algebraic manipulations yield

$$0 < \lim_{\epsilon \to 0^+} \frac{P_\epsilon\left((x, a_{-i}) \to (y, a_{-i})\right)}{\epsilon^{U_i(x, a_{-i}) - U_i(y, a_{-i}) - 1}} < \infty,$$

implying

$$r((x, a_{-i}) \to (y, a_{-i})) = \max\left\{U_i(x, a_{-i}) - U_i(y, a_{-i}) - 1, 0\right\}.$$

Equation (9.25) may be similarly verified. ■

Proof of Theorem 9.2: By Lemma 9.5.1, for any graphical coordination game with graph G, the resistance graph associated with log linear learning is the same for all $k \leq n-1$. Thus it follows that the set of stochastically stable states is independent of k, and thus the susceptibility $\alpha^{\text{sus}}(G, \text{UR}(k))$ is as well. ■

Proof of Theorem 9.3: For any $\alpha \in (0, 1)$, we first show that $\vec{x}$ and $\vec{y}$ are the only recurrent classes of the unperturbed process P. Note that any state transition out of $\vec{x}$ and $\vec{y}$ has positive resistance: $r(\vec{y} \to a) = 2$, and $r(\vec{x} \to a) = 1 + 2\alpha$. Next, consider any state $a \notin \{\vec{x}, \vec{y}\}$, and let i be such that $a_i = x$ and $a_{i+1} = y$. Lemma 9.5.1 gives that the resistance associated with agent i switching to action y is 0, and also that the resistance associated with agent $i+1$ switching to action x is 0. By repeating this argument, it can readily be shown that $R(a \to \vec{x}) = 0$ and $R(a \to \vec{y}) = 0$.

That is, there is a sequence of 0-resistance transitions from any $a \notin \{\vec{x}, \vec{y}\}$ to both $\vec{x}$ and $\vec{y}$, but every transition out of either $\vec{x}$ or $\vec{y}$ has positive resistance – implying that they are the only recurrent classes of P. That is, $R(\vec{x} \to \vec{y}) = 1 + 2\alpha$ and $R(\vec{y} \to \vec{x}) = 2$. Thus,

$$\alpha < 1/2 \implies R(\vec{x} \to \vec{y}) < R(\vec{y} \to \vec{x}) \tag{9.22}$$

and

$$\alpha > 1/2 \implies R(\vec{x} \to \vec{y}) > R(\vec{y} \to \vec{x}), \tag{9.23}$$

yielding the proof. ■

9.3.5 Stability in the presence of a mobile intelligent adversary

Similar to Lemma 9.5.1, in Lemma 9.5.2 we provide a characterization of the resistances of state transitions in the presence of a mobile intelligent adversary. Naturally, these resistances are a function of the adversary's policy, and thus unlike the resistances due to a uniformly random adversary, they do depend implicitly on k.

Lemma 9.5.2 *Suppose agents in N update their actions according to log-linear learning in the presence of a mobile intelligent adversary with capability $k \in \{1, \ldots, n-1\}$ and policy $S : \mathcal{A} \to \mathcal{S}_k$. Then the resistance of a transition where agent $i \in N$ changes its action from x to y in the presence of policy S is:*

$$r^S((x, a_{-i}) \to (y, a_{-i})) = \max\left\{\tilde{U}_i(x, a_{-i}, S(a)) - \tilde{U}_i(y, a_{-i}, S(a)), 0\right\} \tag{9.24}$$

and the resistance of a transition where agent $i \in N$ changes its action from y to x in the *presence of policy S is:*

$$r^S((y,a_{-i}) \to (x,a_{-i})) = \max\left\{\tilde{U}_i(y,a_{-i},S(a)) - \tilde{U}_i(x,a_{-i}S(a)),0\right\}. \tag{9.25}$$

Here $\tilde{U}_i : \mathcal{A} \to \mathbb{R}$, defined in (9.3), is the utility function for agent i in the influenced game $\tilde{G}$.

Proof: This proof proceeds in a very similar fashion to that of Lemma 9.5.1. The main difference lies in the fact that for the UR adversary, for any state $a \in \mathcal{A}$ and any agent $i \in N$, there was a positive probability (*ex ante*, before S is drawn) *either* that $i \in S$ or that $i \notin S$. In contrast, the impact of a MI adversary on agents' utilities is deterministic since S is a function of a. Thus, the asymmetry between $x \to y$ and $y \to x$ transitions that appeared in Lemma 9.5.1 vanishes. ■

For ring graphs, Table 9.1 enumerates several of the key transition resistances experienced under the influence of mobile intelligent adversaries. Next, for the special case of $k = 2$, the following lemma fully characterizes the recurrent classes of the unperturbed best response process $P^{\bar{s}}$ associated with an arbitrary mobile intelligent adversary's policy $\bar{S}$.

Lemma 9.5.3 *For any ring graph $G \in \mathcal{G}^{\mathrm{r}}$, let $k = 2$, let $\bar{S}$ be any adversary policy, and $P^{\bar{S}}$ be the unperturbed best response process on G associated with $\bar{S}$. Every recurrent class of $P^{\bar{S}}$ is a singleton. Furthermore, an action profile $\bar{a}$ (other than $\vec{x}$ and $\vec{y}$) is a recurrent class of $P^{\bar{S}}$ if and only if the following two conditions are both satisfied:*

1. *$\bar{a}$ has a single contiguous chain of agents $[i,j]$ playing x: $\bar{a}_i = \bar{a}_{i+1} = \cdots = \bar{a}_j = x$ such that $j - i \in [1, n-3]$.*
2. *$\bar{S}(\bar{a}) = \{i-1, j+1\}$.*

Proof: At a state $\bar{a}$ and policy satisfying 1) and 2), either agent i or j can switch to y with a resistance of α, or agent $i-1$ or $j+1$ can switch to x with a resistance of $1-\alpha$. All other transitions have higher resistance than these; this demonstrates that $\bar{a}$ is a singleton

Table 9.1: Summary of transition resistances derived from Lemma 9.5.2.

# nbrs with $a_j = a_i$	$i \in S$	$r^S(x \to y)$	$r^S(y \to x)$
0	no	$2 + 2\alpha$	2
	yes	$1 + 2\alpha$	3
1	no	α	0
	yes	0	$1 - \alpha$
2	no	0	0
	yes	0	0

recurrent class. We will show the contrapositive to complete the proof of 1) and 2). If $\bar{a}$ has more than one contiguous chain of agents playing x, then either there is an agent playing y with two neighbors playing x, or there is an uninfluenced agent playing y with a neighbor playing x. Thus, that agent deviating to x is a 0-resistance transition to another state (see Table 9.1) so $\bar{a}$ cannot be a singleton recurrent class. If $\bar{a}$ has a unique contiguous chain of x of length 1 or $n-1$, then there is a 0-resistance transition to $\vec{y}$ or $\vec{x}$, respectively. If 1) is satisfied but 2) is not, then there is an uninfluenced agent playing y with a neighbor playing x; again a 0-resistance transition to another state.

To see that every recurrent class is a singleton, simply note that any state other than $\vec{x}$, $\vec{y}$, or one satisfying 1) can be transformed into either $\vec{x}$ or $\vec{y}$ by a finite sequence of 0-resistance transitions. ■

Naturally, there are a vast number of policies which a mobile intelligent adversary can employ; we propose a family of such policies for ring graphs; we say that any policy in this family is a *balanced policy.*

Definition 9.2 (Balanced Policy) *Let $G \in \mathcal{G}^{\mathrm{r}}$ be a ring graph with n nodes. For any state $a \in \mathcal{A}$, let $[i,j]$ be the longest chain of agents playing x (break ties lexicographically). For a type* MI(k) *agent, policy $S : \mathcal{A} \to \mathcal{S}_k$ is* balanced *if it satisfies the following conditions.*[2]

1. *If $j - i > 1$, let $i \in S(a)$. If additionally $k \geq 3$, let $\{i-1, i, j+1\} \subseteq S(a)$.*

[2]Note that Definition 9.2 does not always specify the location of every adversary. Thus, there is a large family of policies satisfying Definition 9.2.

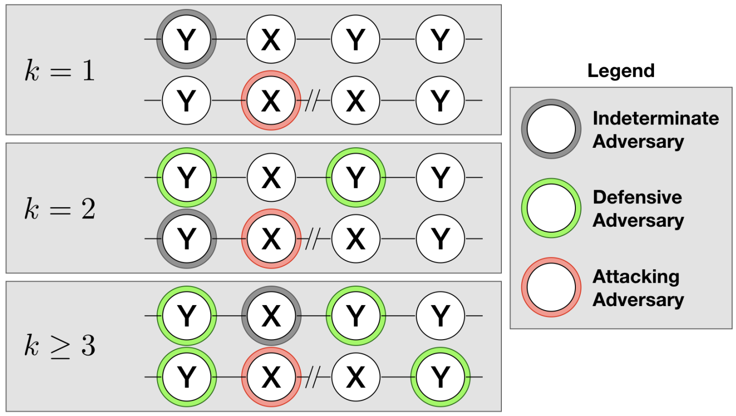

Figure 9.2: Graphical depiction of a balanced mobile intelligent adversary policy of Definition 9.2. There are essentially three strategies of adversarial influence: Indeterminate (gray), Defensive (green), or Attacking (red). For each k depicted above, the upper chain of agents depicts the special case that a single agent is playing x; the lower depicts the longest contiguous chain in the graph of agents playing x. Note that for all $k \geq 1$, a balanced policy requires one of the k adversaries to be attacking whenever the longest chain of x is longer than 1. For $k = 2$, in the special case that the longest chain of x is length 1, both adversaries are defensive. For $k \geq 3$, there are always enough adversaries that two can be defensive and one attacking.

2. *If* $k \geq 2$ *and* $j = i$, *let* $\{i-1, i+1\} \subseteq S(a)$.

Figure 9.2 is a graphical depiction of this policy. The key idea is that there should always be one adversary "attacking" (red circles in Figure 9.2) an x who has a y neighbor, and if there are enough adversaries, the longest contiguous chain of x's should always be surrounded by a pair of "defensive" adversaries (green circles in Figure 9.2). That is, this policy specifies the placement of no more than 3 adversaries; the placement of any remaining "indeterminate" adversaries (dark gray circles in Figure 9.2) is of no importance to the results of Theorem 9.4. We can now proceed with the proof of the theorem.

Proof of Theorem 9.4: We show that under a balanced policy, if α is less than each of the susceptibilities shown in (9.13) that $\vec{y}$ is strictly stochastically stable; subsequently, we show that no policy can outperform a balanced policy. In the following, we write $R^S(a, a')$ to denote the resistance of a transition from a to a' in the presence of adversary policy $S : \mathcal{A} \to \mathcal{S}_k$.

Susceptibility to a balanced policy

Let $G = (N, E)$ be a ring graph influenced by a mobile intelligent adversary with capability k using a balanced policy $S_{\rm b}$ satisfying Definition 9.2. As in the proof of Theorem 9.3, only $\vec{x}$ and $\vec{y}$ are recurrent in the unperturbed process $P^{S_{\rm b}}$, because at every state $a \in \mathcal{A}$, there is a sequence of 0-resistance transitions leading to either $\vec{x}$ or $\vec{y}$, as can be verified by the resistances listed in Table 9.1.

First consider $R^{S_{\rm b}}(\vec{x} \to \vec{y})$. We call a state transition an *incursion* if it is either $\vec{x} \to a$ or $\vec{y} \to a$. Lemma 9.5.2 gives that the incursion of y into state $\vec{x}$ (at an agent being influenced by an adversary) has resistance $r^{S_{\rm b}}(\vec{x} \to a) = 1 + 2\alpha$. Then, for any state $a \notin \{\vec{x}, \vec{y}\}$, there is always at least one agent playing x who has either two neighbors playing y, or a neighbor playing y and is connected to an adversary. Thus, there is always a sequence of transitions from a to $\vec{y}$ with a total resistance of 0, so $R^{S_{\rm b}}(\vec{x} \to \vec{y}) = 1 + 2\alpha$.

Next consider $R^{S_{\rm b}}(\vec{y} \to \vec{x})$. Whenever $k < n$, there is always at least one agent that is not being influenced by the adversary; thus Lemma 9.5.2 gives that an incursion of x into $\vec{y}$ has a resistance of 2. If $k = 1$, a balanced policy does not allow the adversary to "play defense"; so there is a sequence of subsequent $y \to x$ transitions that each have 0 resistance. Thus when $k = 1$ with a balanced policy, the situation is identical to that of the uniformly-random adversary, and $\vec{y}$ is strictly stochastically stable whenever $\alpha < 1/2$.

If $k \geq 2$, the first $y \to x$ transition *after* the incursion now has a positive resistance of at least $1 - \alpha$. If $k = 2$, the policy does not allow the adversary to protect against further spread of x, so we have that $R^{S_{\rm b}}(\vec{y} \to \vec{x}) = 3 - \alpha$. That is, whenever $\alpha < 2/3$, we have that $R^{S_{\rm b}}(\vec{x} \to \vec{y}) < R^{S_{\rm b}}(\vec{y} \to \vec{x})$ so $\vec{y}$ is strictly stochastically stable.

If $k \geq 3$, the policy always defends against further spread of the longest contiguous chain of agents playing x. Consider the $\vec{y} \to \vec{x}$ path given by $\{1, 2, \ldots, n\}$; under a balanced policy, this path has resistance $2 + (n-2)(1-\alpha)$. Note that any alternative $\vec{y} \to \vec{x}$ path has resistance no less than 4, as it would require at least two $y \to x$ transitions of agents who have no neighbors playing x. For any $\alpha < 3/2$, $R^{S_{\rm b}}(\vec{x} \to \vec{y}) < 4$; thus, the only relevant situation is when $R^{S_\mathrm{b}}(\vec{y} \to \vec{x}) = 2 + (n-2)(1-\alpha)$. When this is the case, whenever $\alpha < (n-1)/n$, it follows that $R^{S_\mathrm{b}}(\vec{x} \to \vec{y}) < R^{S_\mathrm{b}}(\vec{y} \to \vec{x})$ so $\vec{y}$ is strictly stochastically stable.

A balanced policy is optimal

Whenever $k < n$, for any ring graph G, it is clear that $\alpha^{\mathrm{sus}}(G, \mathrm{MI}(k)) \leq \alpha^{\mathrm{sus}}(G, \mathrm{MI}(n-1))$ because adding additional adversaries can only decrease $R(\vec{x} \to \vec{y})$ and/or increase $R(\vec{y} \to \vec{x})$. Furthermore, it always holds that $\alpha^{\mathrm{sus}}(G, \mathrm{MI}(n-1)) \leq \alpha^{\mathrm{sus}}(G, \mathrm{FI}(n-1)) = (n-1)/n$. This is because mobile intelligent adversaries are strictly more capable than fixed intelligent. Thus, the susceptibility in (9.13) is tight for the case of $k \geq 3$.

When $k = 1$, for any adversary policy S_1 of type MI(1), the only recurrent classes are $\vec{x}$ and $\vec{y}$. This is because the adversary does not have enough capability to "play defense": at every state a with some agent playing x, then regardless of the adversary's policy, there is an agent playing y that can switch to x with 0 resistance (see Table 9.1). That is, there is a 0-resistance transition to a state with strictly more agents playing x, showing that $R^{S_1}(\vec{y} \to \vec{x}) = R^{S_\mathrm{b}}(\vec{y} \to \vec{x})$. It has already been shown that $R^{S_1}(\vec{x} \to \vec{y}) \geq R^{S_\mathrm{b}}(\vec{x} \to \vec{y})$, and thus it is proved that a balanced policy with $k = 1$ is optimal.

When $k = 2$, the situation is more challenging, because there are too few adversaries for us to appeal to Theorem 9.1's upper bound, but enough adversaries that the unperturbed process may have a multiplicity of recurrent classes. Note that to show that $\alpha^{\mathrm{sus}}(G, \mathrm{MI}(2)) \leq 2/3$, it suffices to show that when $\alpha = 2/3$, $\vec{y}$ can never be strictly stochastically stable for any adversary policy.

Let S_2 be any adversary policy with $k = 2$, and suppose there are m states satisfying both conditions of Lemma 9.5.3 (we call these the "mixed" recurrent classes; in each, some agents are playing x and some y). When $\alpha = 2/3$, the minimum-resistance tree rooted at $\vec{x}$ has total resistance no more than $2 + (m+1)(1-\alpha)$: the resistance of leaving $\vec{y}$ is 2, the resistance of leaving each of the m mixed recurrent classes is $1 - \alpha$, and there can be up to an additional resistance-$1-\alpha$ transition from $\vec{y}$ to a state with 2 agents playing x. Thus, the stochastic potential of $\vec{x}$ as a function of α is

$$\gamma_{\vec{x}}(\alpha) = 2 + (m+1)(1-\alpha). \tag{9.26}$$

Let $\bar{a}$ denote a mixed recurrent class, and with abuse of notation, also the state associated with that class. When $\alpha = 2/3$, for any other state a' accessible from $\bar{a}$, Table 9.1 gives that $r^{S_2}(\bar{a} \to a') \geq 1-\alpha$. Thus, for any other recurrent class $a^\dagger$, we have

$$R^{S_2}\left(\bar{a}, a^\dagger\right) \geq 1-\alpha. \tag{9.27}$$

Now, let T be the minimum-resistance tree rooted at $\vec{y}$. For any recurrent class $a^\dagger$, $R^{S_2}(\vec{x}, a^\dagger) \geq 1+2\alpha$, so it follows from (9.27) that the stochastic potential of $\vec{y}$ as a function of α is lower bounded by

$$\gamma_{\vec{y}}(\alpha) = R^{S_2}(T) \geq 1 + 2\alpha + m(1-\alpha). \tag{9.28}$$

It can be readily seen that for all $\alpha > 2/3$, $\gamma_{\vec{x}}(\alpha) < \gamma_{\vec{y}}(\alpha)$, indicating that $\vec{x}$ is strictly stochastically stable, and showing that $\alpha^{\mathrm{sus}}(G^{\mathrm{r}}, \mathrm{MI}(2)) \leq 2/3$. ■

Chapter 10

Conclusions

In summary, we have presented the following main contributions to the field of robust game design for sociotechnical systems. Recall the central questions posed in Chapter 1:

1. How robust are existing behavior-influencing methodologies to variations in underlying system parameters?

In Chapter 3, we show that fixed tolls are not robust in many ways: variations in latency functions, traffic rate, and network structure can not only prevent fixed tolls from incentivizing optimal flows, but can also render them perverse. Chapter 4 studies well-known marginal-cost tolls and shows that they are more robust, but can still create perverse incentives if a routing problem is sufficiently complex.

2. What features of a behavior-influencing mechanism directly confer robustness?

We show that the universal taxation mechanism of Chapter 5 can incentivize optimal routing, requiring no information about network structure or user sensitivities, provided that it can charge very large prices. In this sense, we might say that "large" prices have more robustness than small ones. Furthermore, Chapter 5 directly compares flow-varying tolls to fixed tolls (see Theorem 5.4), and shows that a great deal of robustness can be gained by using a flow-varying approach.

3. How can a planner systematically avoid perverse incentives?

Our contribution here comes in Chapter 6, where we show that provided the networks under consideration are simple enough (symmetric parallel networks are sufficient), a planner can provably avoid perverse incentives by applying a conservative generalized marginal-cost taxation approach. On the other hand, we show that on general networks, perverse incentives are always a risk.

4. How does this connect with the literature on the design of strategic games?

We investigate this in Chapter 8, and show a setting where there exists no robust incentive design which can prevent arbitrarily-poor states from emerging for a distributed multiagent system. Additionally, Chapter 9 considers the case that an adversary wishes to influence emergent behavior in a distributed system; the results here may help inform the design of more resilient distributed systems.

10.1 Open Questions and Future Work

There is much work to be done here. In the area of incentives for selfish routing, a great deal of our work has focused on the case of so-called *network-agnostic* tolling functions; i.e., tolling functions which do not depend on the structure of the overall network. A recurring theme of this dissertation is that the network-agnostic paradigm is inherently limiting outside of relatively simple settings. Thus, future work should focus on studying the exact form of this limitation. Network agnosticity requires that a taxation mechanism takes the specific form of a mapping from latency functions to taxation functions, and thus cannot depend on any information about anything elsewhere in the network. However, suppose a taxation mechanism were allowed to know exactly which latency functions exist in a network, but that the mechanism could not observe how the respective edges were connected. Would this small additional piece of information improve the effectiveness of a taxation mechanism?

Another area which this dissertation does not consider is knowledge of the latency functions themselves. What if an edge's latency function were only known to be a polynomial with coefficients selected from some known polytope? Would marginal-cost pricing still be an effective approach, perhaps in some conservative sense?

In the area of distributed control, as considered in Chapter 8, how might the pathologies in this dissertation be avoided? A negative result as shown here can be seen as the identification of a bottleneck: what additional information or methodologies would widen the bottleneck? One avenue of investigation here may be to see how allowing limited information-sharing between agents may mitigate these pathologies.

Bibliography

[1] P. N. Brown and J. R. Marden, *Studies on Robust Social Influence Mechanisms: Incentives for Efficient Network Routing in Uncertain Settings*, *IEEE Control Systems Magazine* **37** (2017), no. 1 98–115.

[2] J. Wardrop, *Some Theoretical Aspects of Road Traffic Research*, in *Proc. of the Institution of Civil Engineers, Part II, Vol. 1, No. 36*, pp. 352–362, 1952.

[3] L. Pavel, *Game Theory for Control of Optical Networks.* Springer Science & Business Media, 2012.

[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] N. E. Friedkin, A. V. Proskurnikov, R. Tempo, and S. E. Parsegov, *Network science on belief system dynamics under logic constraints*, *Science* **354** (2016), no. 6310 321–326.

[6] W. Mei, N. E. Friedkin, K. Lewis, and F. Bullo, *Dynamic models of appraisal networks explaining collective learning*, in *2016 IEEE 55th Conference on Decision and Control, CDC 2016*, pp. 3554–3559, 2016. arXiv:1609.0954.

[7] V. Amelkin, F. Bullo, and A. K. Singh, *Polar Opinion Dynamics in Social Networks*, *IEEE Transactions on Automatic Control* **62** (2017), no. 11 5650–5665.

[8] T. Roughgarden, *Selfish Routing and the Price of Anarchy.* MIT Press, 2005.

[9] R. Johari and J. N. Tsitsiklis, *Efficiency Loss in a Network Resource Allocation Game*, *Mathematics of Operations Research* **29** (aug, 2004) 407–435.

[10] H. Youn, M. Gastner, and H. Jeong, *Price of Anarchy in Transportation Networks: Efficiency and Optimality Control*, *Physical Review Letters* **101** (sep, 2008) 128701.

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

[12] E. Nikolova, N. E. Stier-Moses, and T. Lianeas, *The Burden of Risk Aversion in Mean-Risk Selfish Routing*, 2015.

[13] N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, eds., *Algorithmic Game Theory.* Cambridge University Press, 2007.

[14] 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*, pp. 266–272, 2011.

[15] S. Banerjee, R. Johari, and C. Riquelme, *Pricing in Ride-Sharing Platforms : A Queueing-Theoretic Approach*, in *16th ACM Conference on Economics and Computation*, pp. 639–639, 2015.

[16] M. Esmaeili, M.-B. Aryanezhad, and P. Zeephongsekul, *A game theory approach in sellerbuyer supply chain*, *European Journal of Operational Research* **195** (2009), no. 2 442–448.

[17] C. Silva, *Application of Mechanism Design to Electric Power Markets (Republished)*, *IEEE Trans. on Power Systems* **16** (2001), no. 4 862–869.

[18] M. Gairing, *Covering games: Approximation through non-cooperation*, in *Internet and Economics*, pp. 184–195, 2009.

[19] H.-L. Chen, T. Roughgarden, and G. Valiant, *Designing Network Protocols for Good Equilibria*, *SIAM Journal on Computing* **39** (2010), no. 5 1799–1832.

[20] J. R. Marden and A. Wierman, *Distributed Welfare Games*, *Operations Research* **61** (2013) 155–168.

[21] 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* **8** (2014), no. 6 1111–1126.

[22] R. Gopalakrishnan, J. R. Marden, and A. Wierman, *Potential games are necessary to ensure pure nash equilibria in cost sharing games*, *Mathematics of Operations Research* **39** (2014), no. 4 1252–1296, [arXiv:1402.3610].

[23] S. J. O'Hare, R. D. Connors, and D. P. Watling, *Mechanisms that govern how the Price of Anarchy varies with travel demand*, *Transportation Research Part B: Methodological* **84** (2016) 55–80.

[24] C. Wang, X. V. Doan, and B. Chen, *Price of anarchy for non-atomic congestion games with stochastic demands*, *Transportation Research Part B: Methodological* **70** (2014) 90–111, [1310.4874].

[25] G. Karakostas, T. Kim, A. Viglas, and H. Xia, *On the degradation of performance for traffic networks with oblivious users*, *Transportation Research Part B: Methodological* **45** (2011), no. 2 364–371.

[26] A. C. Pigou, *The Economics of Welfare.* Macmillan, New York, 1920.

[27] D. Braess, *Uber ein Paradoxon aus der Verkehrsplanung*, *Unternehmensforschung* (1968) 258–268.

[28] D. Braess, A. Nagurney, and T. Wakolbinger, *On a Paradox of Traffic Planning*, *Transportation Science* **39** (nov, 2005) 446–450.

[29] R. Cole, Y. Dodis, and T. Roughgarden, *How Much Can Taxes Help Selfish Routing?*, *Journal of Computer and System Sciences* **72** (may, 2006) 444–467.

[30] M. J. Smith, *The Marginal Cost Taxation of a Transportation Network*, *Transportation Research Part B: Methodological* **13** (1979), no. 3 237–242.

[31] S. A. Morrison, *A survey of road pricing*, *Transportation Research Part A: General* **20** (1986), no. 2 87–97.

[32] R. B. Dial, *Network-Optimized Road Pricing: Part I: A Parable and a Model*, *Operations Research* **47** (jan, 1999) 54–64.

[33] L. Engelson and P. Lindberg, *Congestion pricing of road networks with users having different time values*, *Applied Optimization* **101** (2006) 81–104.

[34] G. Christodoulou, K. Mehlhorn, and E. Pyrga, *Improving the price of Anarchy for Selfish Routing via coordination mechanisms*, *Algorithmica* **69** (2014), no. 3 619–640, [arXiv:1202.2877].

[35] W. Sandholm, *Negative Externalities and Evolutionary Implementation*, *The Review of Economic Studies* **72** (jul, 2005) 885–915.

[36] V. Bonifaci, T. Harks, and G. Shaefer, *The Impact of Stackelberg Routing in General Networks*, *Mathematics of Operations Research* **32** (2010), no. 2 1–17.

[37] R. Arnott, A. de Palma, and R. Lindsey, *Does providing information to drivers reduce traffic congestion?*, *Transportation Research Part A: General* **25** (1991), no. 5 309–318.

[38] H.-L. Chen, T. Roughgarden, and G. Valiant, *Designing Networks with Good Equilibria*, in *Proc. 19th ACM SIAM Symp. on Discrete Algorithms*, pp. 854—-863, 2008.

[39] 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, 2003.

[40] 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), pp. 277–285, 2004.

[41] G. Karakostas and S. Kolliopoulos, *Edge pricing of multicommodity networks for heterogeneous selfish users*, in *Proc. 45th IEEE Symp. on Foundations of Computer Science*, (Rome, Italy), pp. 268–276, 2004.

[42] M. J. Beckmann, C. B. McGuire, and C. B. Winsten, *Studies in the Economics of Transportation*, 1955.

[43] W. Sandholm, *Evolutionary Implementation and Congestion Pricing*, *The Review of Economic Studies* **69** (2002), no. 3 667–689.

[44] C. Sun, *A Time Variant Log-Linear Learning Approach to the SET K-COVER Problem in Wireless Sensor Networks*, 2017.

[45] S. Rahili, J. Lu, W. Ren, and U. Al-Saggaf, *Distributed Coverage Control of Mobile Sensor Networks in Unknown Environment Using Game Theory: Algorithms and Experiments*, 2017.

[46] P. Li and H. Duan, *A potential game approach to multiple UAV cooperative search and surveillance*, *Aerospace Science and Technology* **68** (2017) 403–415.

[47] M. S. Ali, P. Coucheney, and M. Coupechoux, *Load Balancing in Heterogeneous Networks Based on Distributed Learning in Near-Potential Games*, *IEEE Transactions on Wireless Communications* **15** (2016), no. 7 5046–5059.

[48] K. Cohen, A. Nedic, and R. Srikant, *Distributed Learning Algorithms for Spectrum Sharing in Spatial Random Access Wireless Networks*, *IEEE Transactions on Automatic Control* **62** (2017), no. 6 2854–2869, [arXiv:1507.0566].

[49] Y. Lim and J. Shamma, *Robustness of stochastic stability in game theoretic learning*, in *American Control Conference (ACC)*, pp. 6160–6165, 2013.

[50] J. R. Marden and A. Wierman, *Overcoming the limitations of utility design for multiagent systems*, *IEEE Transactions on Automatic Control* **58** (2013), no. 6 1402–1415.

[51] O. Candogan, A. Ozdaglar, and P. A. Parrilo, *Dynamics in near-potential games*, *Games and Economic Behavior* **82** (2013) 66–90, [arXiv:1107.4386].

[52] N. Li and J. R. Marden, *Designing games for distributed optimization*, *IEEE Journal on Selected Topics in Signal Processing* **7** (2013), no. 2 230–242.

[53] P. J. Mackie, S. Jara-Díaz, and A. S. Fowkes, *The value of travel time savings in evaluation*, *Transportation Research Part E: Logistics and Transportation Review* **37** (2001), no. 2-3 91–106, [1011.1669].

[54] A. Mas-Colell, *On a Theorem of Schmeidler*, *Mathematical Economics* **13** (1984) 201–206.

[55] P.-a. Chen, B. De Keijzer, D. Kempe, and G. Shaefer, *Altruism and Its Impact on the Price of Anarchy*, *ACM Trans. Economics and Computation* **2** (2014), no. 4.

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

[57] T. Roughgarden, *How Bad Is Selfish Routing?*, *J. ACM* **49** (2002), no. 2 236–259.

[58] T. Roughgarden, *The price of anarchy is independent of the network topology*, *J. Computer and System Sciences* **67** (sep, 2003) 341–364.

[59] G. Perakis and G. Roels, *The Price of Anarchy in Supply Chains: Quantifying the Efficiency of Price-Only Contracts*, *Management Science* **53** (aug, 2007) 1249–1268.

[60] R. P. Leme and É. Tardos, *Pure and Bayes-Nash price of anarchy for generalized second price auction*, in *Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 735–744, 2010.

[61] V. Syrgkanis and E. Tardos, *Composable and Efficient Mechanisms*, in *Symposium on Theory of Computing*, pp. 211–220, 2013. arXiv:1211.1325.

[62] B. Yolken and N. Bambos, *Game based capacity allocation for utility computing environments*, *Telecommunication Systems* **47** (may, 2010) 165–181.

[63] R. Hoeksma and M. Uetz, *The price of anarchy for minsum related machine scheduling*, in *Approximation and Online Algorithms*, pp. 261–273. Springer, 2011.

[64] G. Piliouras, E. Nikolova, and J. S. Shamma, *Risk sensitivity of price of anarchy under uncertainty*, in *Proceedings of the 14th ACM Conference on Electronic Commerce*, vol. 9, (New York, New York, USA), pp. 715–732, 2013.

[65] T. Roughgarden and F. Schoppmann, *Local smoothness and the price of anarchy in splittable congestion games*, *J. of Economic Theory* **1** (apr, 2014) 1–26.

[66] T. Lianeas, E. Nikolova, and N. E. Stier-Moses, *Asymptotically tight bounds for inefficiency in risk-averse selfish routing*, in *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2016-Janua, pp. 338–344, 2016. arXiv:1510.0206.

[67] P. Kleer and G. Schäfer, *Path deviations outperform approximate stability in heterogeneous congestion games*, in *International Symposium on Algorithmic Game Theory*, pp. 212–224, 2017. arXiv:1707.0127.

[68] T. Roughgarden, *Stackelberg Scheduling Strategies*, *SIAM Journal on Computing* **33** (2004), no. 2 332–350.

[69] V. Bonifaci, M. Salek, and G. Schäfer, *Efficiency of restricted tolls in non-atomic network routing games*, in *Algorithmic Game Theory*, vol. 6982 LNCS, pp. 302–313, 2011. 10.1007/978-3-642-24829-0_27.

[70] T. Jelinek, M. Klaas, and G. Schäfer, *Computing Optimal Tolls with Arc Restrictions and Heterogeneous Players*, *LIPIcs - Leibniz International Proceedings in Informatics* **25** (2014) 444.

[71] U. Bhaskar, K. Ligett, L. Schulman, and C. Swamy, *Achieving Target Equilibria in Network Routing Games without Knowing the Latency Functions*, in *Proc. IEEE Symp. Foundations of Computer Science*, (Philadelphia, PA, USA), pp. 31–40, 2014. arXiv:1408.1429.

[72] R. Colini-Baldeschi, M. Klimm, and M. Scarsini, *Demand-Independent Tolls*, arXiv:1708.0273.

[73] D. A. Lazar, S. Coogan, and R. Pedarsani, *Capacity Modeling and Routing for Traffic Networks with Mixed Autonomy*, *2017 56th IEEE Conference on Decision and Control (CDC)* (2017) 5678–5683.

[74] P. N. Brown and J. R. Marden, *The Robustness of Marginal-Cost Taxes in Affine Congestion Games*, *Transactions on Automatic Control* **62** (2017), no. 8 3999–4004.

[75] P. N. Brown and J. R. Marden, *Optimal Mechanisms for Robust Coordination in Congestion Games*, *IEEE Transactions on Automatic Control* **PP** (2018), no. 1.

[76] I. Milchtaich, *Topological Conditions for Uniqueness of Equilibrium in Networks*, *Mathematics of Operations Research* **30** (2005), no. 1 225–244.

[77] P. N. Brown and J. R. Marden, *Avoiding Perverse Incentives in Affine Congestion Games*, in *55th IEEE Conf. on Decision and Control*, (Las Vegas, NV, USA), pp. 7010–7015, 2016.

[78] P. N. Brown and J. R. Marden, *Fundamental Limits of Locally-Computed Incentives in Network Routing*, in *American Control Conf.*, (Seattle, WA, USA), pp. 5263–5268, 2017.

[79] P. N. Brown and J. R. Marden, *The Benefit of Perversity in Distributed Network Routing*, in *56th IEEE Conf. on Decision and Control*, (Melbourne, Australia), pp. 6229–6234, 2017.

[80] R. Meir and D. Parkes, *Playing the Wrong Game: Smoothness Bounds for Congestion Games with Behavioral Biases*, *Performance Evaluation Review* **43** (2015), no. 3 67–70, [arXiv:1411.1751].

[81] P. N. Brown and J. R. Marden, *A Study on Price-Discrimination for Robust Social Coordination*, in *American Control Conf.*, (Boston, MA, USA), pp. 1699–1704, 2016.

[82] H. Varian, *Monopoly Behavior*, in *Intermediate Microeconomics, a Modern Approach, 8th ed.*, pp. 462–484. 2010.

[83] R. B. Wilson, *Nonlinear Pricing.* Oxford University Press, 1993.

[84] W. J. Baumol, *Ramsey Pricing*, *The New Palgrave* **4** (1987) 49–51.

[85] K. Tsakalozos, H. Kllapi, E. Sitaridi, M. Roussopoulos, D. Paparas, and A. Delis, *Flexible use of cloud resources through profit maximization and price discrimination*, in *Proc. International Conference on Data Engineering*, pp. 75–86, 2011.

[86] H. Shen and T. Basar, *Optimal Nonlinear Pricing for a Monopolistic Network Service Provider with Complete and Incomplete Information*, *IEEE Journal on Selected Areas in Communications* **25** (2007), no. 6 1216–1223.

[87] D. Bergemann and K. H. Schlag, *Pricing without Priors*, *Journal of the European Economic Association* **6** (2008), no. 2-3 560–569.

[88] H. P. Borowski, J. R. Marden, and E. W. Frew, *Fast convergence in semi-anonymous potential games*, in *Proceedings of the IEEE Conference on Decision and Control*, pp. 2418–2423, 2013. arXiv:1511.0591.

[89] D. H. Wolpert, K. Tumer, and J. Frank, *Using collective intelligence to route Internet traffic*, *Proceedings of the 1998 conference on Advances in neural information processing systems II* (1999) 952–958, [9905004].

[90] B. Pradelski and H. P. Young, *Learning Efficient Nash Equilibria in Distributed Systems*, *Games and Economic Behavior* **75** (2012), no. 2 882–897.

[91] L. E. Blume, *The Statistical Mechanics of Strategic Interaction*, *Games and Economic Behavior* **5** (1993), no. 3 387–424.

[92] L. E. Blume, *The Statistical Mechanics of Best-Response Strategy Revision*, *Games and Economic Behavior* **11** (1995), no. 2 111–145.

[93] H. P. Young, *The Evolution of Conventions*, *Econometrica* **61** (1993), no. 1 57–84.

[94] C. Alós-Ferrer and N. Netzer, *The logit-response dynamics*, *Games and Economic Behavior* **68** (2010), no. 2 413–427.

[95] M. Kandori, G. J. Mailath, and R. Rob, *Learning, Mutation, and Long Run Equilibria in Games*, *Econometria* **61** (1993), no. 1 29–56.

[96] J. R. Marden, G. Arslan, and J. S. Shamma, *Joint strategy fictitious play with inertia for potential games*, *IEEE Transactions on Automatic Control* **54** (2009), no. 2 208–220.

[97] M.-F. Balcan, A. Blum, and Y. Mansour, *The Price of Uncertainty*, *ACM Transactions on Economics and Computation* **1** (2013), no. 3 1–29.

[98] N. I. Al-Najjar and R. Smorodinsky, *Pivotal Players and the Characterization of Influence*, *Journal of Economic Theory* **92** (2000), no. 2 318–342.

[99] Y. Azrieli and E. Shmaya, *Lipschitz Games*, *Mathematics of Operations Research* **38** (2012), no. 2 350–357, [arXiv:1107.1520].

[100] D. Monderer and L. S. Shapley, *Potential Games*, *Games and Economic Behavior* **14** (1996), no. 1 124–143.

[101] J. R. Marden and J. S. Shamma, *Revisiting Log-Linear Learning: Asynchrony, Completeness and a Payoff-based Implementation*, *Games and Economic Behavior* **75** (2012), no. 4 788–808.

[102] T. Tatarenko, *Log-Linear Learning: Convergence in Discrete and Continuous Strategy Potential Games*, in *53rd IEEE Conference on Decision and Control*, pp. 426–432, 2014.

[103] R. D. McKelvey and T. R. Palfrey, *Quantal response equilibria for normal form games*, *Games and Economic Behavior* **10** (1995), no. 1 6–38.

[104] P. N. Brown, H. P. Borowski, and J. R. Marden, *Security Against Impersonation Attacks in Distributed Systems*, *IEEE Transactions on Control of Network Systems* **PP** (2018), no. 1 [arxiv.org/abs/1710.08500].

[105] R. W. Cooper, *Coordination Games.* Cambridge University Press, Cambridge, UK, 1999.

[106] E. Ullmann-Margalit, *The Emergence of Norms.* Oxford University Press, 1977.

[107] M. Kearns, M. L. Littman, and S. Singh, *Graphical Models for Game Theory*, *Proceedings of the 17th conference on Uncertainty in artificial intelligence* (2001), no. Owen 253–260, [arXiv:1301.2281].

[108] H. P. Young, *Strategic Learning and its Limits*, vol. 9780199269. 2011.

[109] A. Montanari and A. Saberi, *The spread of innovations in social networks.*, *Proceedings of the National Academy of Sciences of the United States of America* **107** (2010), no. 47 20196–20201.

[110] D. Shah and J. Shin, *Dynamics in Congestion Games*, *Proc. ACM SIGMETRICS'10* **38** (2010), no. 1 107–118.

[111] D. Foster and P. Young, *Stochastic evolutionary game dynamics*, *Theoretical Population Biology* **38** (1990), no. 2 219–232, [1011.1669].

[112] H. P. Borowski and J. R. Marden, *Understanding the Influence of Adversaries in Distributed Systems*, in *IEEE Conference on Decision and Control (CDC)*, pp. 2301–2306, 2015.