Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Efficiency Analysis of Cournot Competition in Service Industries with Congestion

Georgia Perakis, Wei Sun

To cite this article:
Georgia Perakis, Wei Sun (2014) Efficiency Analysis of Cournot Competition in Service Industries with Congestion. Management Science 60(11):2684-2700. http://dx.doi.org/10.1287/mnsc.2014.1943

MANAGEMENT SCIENCE
Vol. 60, No. 11, November 2014, pp. 2684–2700
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.1943

Efficiency Analysis of Cournot Competition in Service Industries with Congestion

Georgia Perakis
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, georgiap@mit.edu

Wei Sun
IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, sunw@us.ibm.com

We consider Cournot competition in the presence of congestion effects. Our model consists of several service providers with differentiated services, each competing for users who are sensitive to both price and congestion. We distinguish two types of congestion effects, depending on whether spillover costs exist, that is, where one service provider's congestion cost increases with the other providers' output level. We quantify the efficiency of an unregulated oligopoly with respect to the optimal social welfare with tight upper and lower bounds. We show that, when there is no spillover, the welfare loss in an unregulated oligopoly is limited to 25% of the social optimum, even in the presence of highly convex costs. On the other hand, when spillover cost is present, there does not exist a constant lower bound on the efficiency of an unregulated oligopoly, even with affine cost. We show that the efficiency depends on the relative magnitude between the marginal spillover cost and the marginal benefit to consumers.

Keywords: efficiency analysis; price of anarchy; congestion; convex costs; service industries
History: Received October 9, 2011; accepted February 13, 2014, by Christian Terwiesch, operations management. Published online in *Articles in Advance* August 1, 2014.

1. Introduction

With rapid proliferation of content-rich multimedia devices such as smartphones and tablets, wireless congestion has increasingly become a common problem in many urban areas. Users in an affected network may experience spotty service, dropped calls, and sluggish data speeds. In the United States, the Federal Communication Committee (FCC) is responsible for allocating spectrum bands. In one of its recent studies (Federal Communications Commission 2010), it predicted that the demand for wireless bandwidth would grow between 25 and 50 times the current levels in the subsequent five years and surpass the available spectrum by as early as 2014. Given the finite bandwidth, one may wonder what kind of measure the FCC should take to control congestion and ensure an efficient usage of this limited resource.

Another area routinely plagued by congestion is airports. According to the Bureau of Transportation Statistics (2014), nearly 20% of domestic flights arrived late in 2013. The Joint Economic Committee (2008) reported that flight delays cost passengers, airlines, and the U.S. economy more than $40 billion every year. Current landing fees across the United States typically depend only on aircraft weight and do not vary with traffic. Given that the existing system is unlikely to promote an efficient use of the scarce runway space, it prompts for other mechanisms to alleviate the widespread problem of airport congestion across the country.

One common feature in the two examples mentioned above is oligopolistic competition with congestion effects. In the example of wireless communication, in the fourth quarter of 2013, Verizon, AT&T, Sprint, Nextel, and T-Mobile together controlled nearly 93% of the U.S. market (Statista 2014). Although airports vary in sizes, a small number of airlines usually dominate an airport in terms of flight share. For both industries, the Cournot model, also known as quantity competition, has received both theoretical and empirical support as a good modeling choice. From a theoretical perspective, Kreps and Scheinkman (1983) show that Cournot models best approximate the long-run results of two-stage competition with capacity choice followed by price setting: airlines compete by first setting schedules and later setting prices to fill seats, and wireless service providers first purchase bandwidth and then determine prices for subscriptions. Empirical works, including Weisman (1990), Brander and Zhang (1990, 1993), Oum et al. (1993), Parker and Roller (1997), and Faulhaber and Hogendorn (2000), also support Cournot models reflecting actual competition in these two industries.

We would like to point out a subtle difference associated with congestion effects in the two examples. As wireless service providers use different frequency bands to reduce signal interference, congestion

experienced with one carrier does not affect others. The congestion effect in this case is *fully self-contained*—i.e., the cost associated with one service provider only depends on his output level, such as the amount of bandwidth consumed. On the other hand, when one airline schedules an additional flight in a congested airport, it creates additional delays for every flight, which includes landing and taking off. Thus, for the airport example, besides self-contained cost, there also exists *spillover cost*, where an increase in one's output level also increases others' delay.

In this paper, we study both settings, depending on the presence of spillover cost. We evaluate the efficiency of an unregulated oligopoly by comparing its welfare with the social optimum, which can be achieved by implementing congestion pricing. Our model consists of users, service providers, and a facility manager: users are sensitive to both price and congestion, service providers compete for users with differentiated services by adjusting the output level, and the facility manager imposes an admission-level pricing scheme on service providers, with the goal of improving social welfare. Relating this model to the two examples, the FCC could play the role of facility manager and determine a unit price for the bandwidth allocated to each wireless carrier, which in turn would determine how many users to enroll. Similarly, the Federal Aviation Administration would be responsible for imposing a landing fee, and each airline would then determine its flight frequency.

1.1. Contributions

Our study contributes to the extant literature on operations management and congestion in the following areas:

- *Incentive alignment through congestion pricing:* We use an admission-level pricing scheme as a mechanism to coordinate the actions of profit-maximizing oligopolists with the goal of maximizing social welfare. We show that, when costs are fully self-contained, the facility manager must provide subsidies because service providers tend to underproduce compared with the social optimum. Moreover, the facility manager does not need to explicitly manage congestion because service providers have taken the effect into account when they optimize their output level. When spillover costs are present, service providers could still be entitled to subsidies when their services offer more benefits to users relative to the congestion cost. The facility manager only needs to issue a penalty when the congestion cost that a service provider imposes onto others outweighs the consumer surplus contributed by his services.
- *Efficiency analysis to quantify welfare loss:* We compare the total welfare in an unregulated oligopoly, where service providers have free access to the facility, with the social optimal welfare to assess how much efficiency is lost because of the lack of coordination. We show that with self-contained costs, the maximum efficiency loss in an unregulated oligopoly is limited to 25% of the optimal welfare, even in the presence of nonlinear costs. On the other hand, when spillover cost is present, there does not exist a constant lower bound on the efficiency of an unregulated oligopoly, even with affine cost. In particular, we show that when the marginal spillover cost associated with enrolling an additional user exceeds the increase in consumer surplus, the efficiency loss in an unregulated oligopoly could be significant.
- *Identification of key performance indicators and justification of coordination:* In the absence of spillover cost, we show that efficiency of the unregulated oligopoly improves as competition among service providers increases. The result implies the limited role that coordination could play from a societal viewpoint. With spillover cost, we show that the efficiency depends on the external cost-to-benefit ratio, which measures the severity of congestion by comparing the marginal spillover cost and the marginal consumer surplus. All things being equal, a higher external cost-to-benefit ratio indicates a larger welfare loss in the unregulated setting. It implies greater benefits that coordination could bring, thus making congestion pricing more desirable.
- *Tight analytical bounds with novel proof techniques:* Our modeling choice is made in the context of service industries that compete with quantities and experience congestion. We present conditions and classes of cost functions for which the analytical bounds in the efficiency analysis are tight. We develop the analysis by utilizing tools such as Jacobian similarity, copositivity, and spectral theorem from matrix analysis. We believe the methodology proposed in this work could potentially be used in other settings that incorporate asymmetric competition and nonlinear costs.

1.2. Related Literature

Motivated by congestion management in transportation and communication networks, there has been a huge body of literature that analyzes traffic in a congested network (e.g., Hamdouch et al. 2007, Hayrapetyan et al. 2007, Maillé and Stier-Moses 2009). Acemoglu and Ozdaglar (2007) and Ozdaglar (2008) study competition among profit-maximizing oligopolists who set prices on the links, and congestion cost on each link only depends on its traffic volume. In the operations management literature, several more recent papers have addressed the issue of congestion in service industries (see, for example, Cachon and Harker 2002; Allon and Federgruen 2007, 2008; Johari et al. 2010), where each firm is only sensitive to congestion caused by its users. A common method to describe users' behavior in congestion models is by the Wardrop equilibrium

(e.g., Johari et al. 2010, Acemoglu and Ozdaglar 2007), also known as unimodal traffic equilibrium in the transportation literature (Wardrop 1952). It implies that with identical or perfectly substitutable services, all users will choose the service provider who offers the lowest price. Our model differentiates itself from prior work by incorporating new features, such as asymmetric oligopolists with differentiated services. We consider the general Wardrop equilibrium, an extension introduced by Dafermos (1982), that describes a multimodal traffic network equilibrium. The equilibrium conditions imply that several prices could exist when service providers offer imperfect substitutes.

Besides the traditional network setting and applications with only self-contained costs, congestion pricing has also been proposed to address airport delays where spillover costs exist. Daniel (1995, 2001) demonstrates the potential benefits of congestion pricing for an airport via a simulation model based on stochastic queuing theory. Brueckner (2002) argues that airport congestion pricing is different from road tolls because each road user is small relative to total traffic, in contrast to the observation that airlines have market power. The author shows, among other things, that airlines internalize some congestion costs based on a symmetric duopoly and generalizes the result to a network setting in Brueckner (2005). In this work, we model congestion pricing as a "coordinating contract" between the airlines and a facility manager, with the goal of quantifying the benefits in terms of the social welfare that such a scheme can offer.

Our work measures the performance of an unregulated setting with respect to a centralized system, so it is closely related to a stream of literature on the *price of anarchy*, popularized by Koutsoupias and Papadimitriou (1999). It compares the performance of the worst Nash outcome with respect to the centralized solution. The concept has been used in transportation networks (Roughgarden and Tardos 2002; Correa et al. 2004, 2007; Roughgarden 2005; Perakis 2007), network pricing (Acemoglu and Ozdaglar 2007, Johari et al. 2010), single-tier oligopolistic pricing games (Farahat and Perakis 2009, 2010; Kluberg and Perakis 2012), and supply chain games (Perakis and Roels 2007, Martínez-de-Albéniz and Simchi-Levi 2009, Perakis and Sun 2012).

A common feature across Farahat and Perakis (2009, 2010), Kluberg and Perakis (2012) (which will be referred to as KP), Perakis and Sun (2012), and this paper is that part of analysis utilizes tools from matrix analysis to study an asymmetric game with multiple players. All prior work aforementioned (with the exception of KP) study Bertrand competition (also known as price competition) as opposed to Cournot competition. Unfortunately, the results and insights, as well as the proof techniques, generally do not transfer from the price to the quantity competition setting (see Farahat and Perakis 2011 for a discussion on how the two settings differ). Between KP and this work, there are several distinctions. In terms of the model, KP focuses on a single-tier oligopoly, whereas our model also considers users who are both price and delay sensitive. Second, KP uses constant marginal cost; we consider nonlinear costs. In addition, we also consider the setting with spillover cost. Third, KP only focuses on the worst-case analysis with a lower bound on the performance, whereas we also determine the best case with parametric bounds that indicate the key drivers in the unregulated oligopoly.

The rest of this paper is organized as follows. In §2, we introduce our model and assumptions used in the paper. Section 3 presents the efficiency analysis for the setting with fully self-contained costs; §4 focuses on the setting where spillover costs also exist. We evaluate the tightness of the bounds in §5 and conclude the paper in §6.

2. Problem Formulation

We consider a facility with n differentiated services, each offered by a provider. We denote q_i as the output level chosen by service provider $i = 1, \ldots, n$. Given an output level $\mathbf{q} = (q_1, \ldots, q_n)$, we denote the marginal utility derived from consumption as $\mathbf{u}(\mathbf{q}) = (u_1(\mathbf{q}), \ldots, u_n(\mathbf{q}))$. One can interpret $u_i(\mathbf{q})$ as the additional utility of service i obtained by an infinitesimal user as when the facility maintains a service level $\mathbf{q}$. As is common in the pricing literature (see Vives 1999), we consider the marginal utility function as a affine function of output level:

$$\mathbf{u}(\mathbf{q}) = \bar{\mathbf{p}} - \mathbf{B}\mathbf{q} = \begin{bmatrix} \bar{p}_1 \\ \vdots \\ \bar{p}_n \end{bmatrix} - \begin{bmatrix} \beta_{11} & \cdots & \beta_{1n} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \cdots & \beta_{nn} \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix},$$

where $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_n)$ represents the maximum prices that a user is willing to pay for the services. Different $\bar{p}_i$ captures the quality differences perceived by consumers, which could be affected by factors such as brand recognition, word-of-mouth effects, and prior experience with the product, among others.

Assumption 1. *Matrix* $\mathbf{B}$ *is a symmetric and positive definite matrix. In addition,* $\beta_{ii} > 0$ *for all* i *and* $\beta_{ij} \geq 0$ *for all* $j \neq i$.

The assumption implies that $\partial u_i(\mathbf{q})/\partial q_i < 0$—i.e., each service faces a downward sloping demand resulted from users' diminishing return from consumption. Meanwhile, $\partial u_i(\mathbf{q})/\partial q_j \leq 0$ suggests that services i and j are "strategic substitutes" (Bulow et al. 1985). The symmetry of the matrix is a natural consequence of maximizing a quasilinear utility function of a representative consumer.

In a congested facility, both users and service providers are affected by congestion. For the case of airport congestion, airlines have to pay extra for crew, fuel, and maintenance costs while delayed travelers and their employers lose productivity, business opportunities, and leisure activities. Let $l_i^p(\mathbf{q})$ and $l_i^u(\mathbf{q})$ denote the congestion cost per service incurred by the service provider i and his users, respectively. Let $l_i(\mathbf{q}) = l_i^p(\mathbf{q}) + l_i^u(\mathbf{q})$ represent the aggregate congestion cost per i's service. When service provider i enrolls q_i users, the total congestion cost associated with i and his users is $l_i(\mathbf{q})q_i$. Given $\mathbf{l}(\mathbf{q}) = (l_1(\mathbf{q}), \ldots, l_n(\mathbf{q}))$, the total congestion cost in the facility is given by $\mathbf{q}^\top \mathbf{l}(\mathbf{q})$.

Assumption 2. *For every i, the cost function $l_i(\mathbf{q})$ is convex, component-wise nondecreasing, and continuously differentiable with respect to q_j for all $j = 1, \ldots, n$.*

Denote the Jacobian matrix of $\mathbf{l}(\mathbf{q})$ as

$$\mathbf{R} = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{bmatrix},$$

where $r_{ij} = \partial l_i / \partial q_j$. Except for affine cost functions, the Jacobian matrix $\mathbf{R}$ depends on the value of $\mathbf{q}$. We will use superscripts to differentiate the matrices evaluated at different values of $\mathbf{q}$.

Next, we formally define the two types of congestion cost studied in this work.

Definition 1. *Self-contained cost* is denoted by $\partial l_i / \partial q_i \geq 0$ for all $i \in \{1, \ldots, n\}$. *Spillover cost* is denoted by $\partial l_i / \partial q_j \geq 0$ for all $j \neq i$.

Denote the diagonal and off-diagonal elements of the Jacobian matrix $\mathbf{R}$ as $\mathbf{\Gamma_R}$ and $\mathbf{R}_{\text{off}}$, respectively. We distinguish two settings, depending on the existence of spillover cost. When there is no spillover cost, the cost associated with service i depends only on i's output level. It implies that the Jacobian matrix on cost is simply a diagonal matrix; i.e., $\mathbf{R} = \mathbf{\Gamma_R}$. In the setting with spillover costs, an increase in i's output level can lead to an increase in j's cost. The spillover cost is captured by $\mathbf{R}_{\text{off}}$.

Assumption 3. *The maximum reservation price must satisfy $\bar{\mathbf{p}} - \mathbf{l}(\mathbf{0}) > \mathbf{0}$.*

The assumption states that the maximum profit per service must be positive. If this assumption is violated, it implies that no user is willing to pay for the service, and the corresponding "inactive" service provider could be removed from the equilibrium.

For the analysis with spillover cost, we use the following assumption.

Assumption 4. *The Jacobian matrix of the cost function $\mathbf{l}(\mathbf{q})$, $\mathbf{R}$, is symmetric. Denote the Jacobian matrix of $(l_1'(\mathbf{q}), \ldots, l_n'(\mathbf{q}))$ as $\mathbf{H}$, where $l_i'(\mathbf{q}) = \partial l_i / \partial q_i$ is positive quasi-definite (i.e., $\mathbf{H} + \mathbf{H}^\top$ is positive definite).*

Assumption 4 imposes restrictions on the Jacobian and Hessian matrices of the cost function $\mathbf{l}(\mathbf{q})$. Note that when service providers have the same congestion cost per service $l_i(\mathbf{q})$, the assumption is in fact unnecessary. For example, consider a $M/M/1$ facility that is shared by the service providers, where the cost per unit time is c and the maximum capacity is μ. The average delay cost in such a system is given by $l_i(\mathbf{q}) = c/(\mu - \sum_i q_i)$ for all i. Then one can show that both $\mathbf{R}$ and $\mathbf{H}$ are rank-1 positive semidefinite matrices that satisfy Assumption 4.

This assumption simplifies the derivation in the analysis with spillover cost as $\mathbf{R} = \mathbf{R}^\top$, and the condition on the Hessian matrix ensures the uniqueness of the solution in the unregulated setting. It is possible to relax it to some extent, but it involves more tedious derivations. As a result, we impose this assumption to enhance the transparency of the model.

2.1. Three-Tier Model

Our model consists of three key players: users, service providers, and a facility manager. Users select service providers by taking into account the price and congestion. Every service provider is a profit maximizer who competes with differentiated services by adjusting his output level. The facility manager imposes an access fee on service providers, with the goal of improving social welfare. The three-stage problem is solved with backward induction. We now present the problem faced by each player and characterize the respective outcome.

2.1.1. User Behavior. A user's total disutility associated with using service i is the sum of the price he pays, p_i, and the congestion cost he experiences, $l_i^u(\mathbf{q})$. The expression $p_i + l_i^u(\mathbf{q})$ is known as the *full price* or *effective price* in the literature (Johari et al. 2010, Acemoglu and Ozdaglar 2007).

We assume that each user is "small" compared with the total traffic volume in the sense that, when he switches from a service provider to another, there is no considerable change in the congestion cost. To model users' behavior of choosing differentiated services, we use a multimodal traffic network equilibrium analogy with elastic demands. The resulting equilibrium is also known as the general Wardrop equilibrium in the transportation literature (Beckmann et al. 1956, Dafermos 1982): a vector of output level $\mathbf{q}$ is a *general Wardrop equilibrium* (GWE), if for every service provider i,

$$p_i + l_i^u(\mathbf{q}) = u_i(\mathbf{q}), \quad \text{if } q_i > 0;$$
$$p_i + l_i^u(\mathbf{q}) \geq u_i(\mathbf{q}), \quad \text{if } q_i = 0.$$

The equilibrium conditions state that, for every active provider whose service level is positive, his full price must be equal to the marginal utility function obtained in the equilibrium.

Remark 1. In a setting with a single type of service or symmetric service providers, $u_i(\mathbf{q}) = u(\mathbf{q})$ for all i. GWE implies that a *single* full price prevails in an equilibrium. Intuitively, with perfectly substitutable services, when one provider charges a higher full price, all his users would switch to other providers. With differentiated services, different full prices could coexist in a market as service providers leverage product differentiation.

Without loss of generality, we restrict our attention to these active service providers with $q_i > 0$, whose fare prices follow

$$p_i = u_i(\mathbf{q}) - l_i^u(\mathbf{q}). \tag{1}$$

When there is no congestion, the market-clearing price of an active service is simply its marginal utility function; i.e., $p_i = u_i(\mathbf{q})$. The presence of congestion directly lowers users' willingness to pay and service providers' profitability.

2.1.2. Service Provider's Profit Maximization Problem. Service provider i's profit function is defined as $\pi_i(q_i, \mathbf{q}_{-i}) = q_i(p_i(q_i, \mathbf{q}_{-i}) - t_i - l_i^p(q_i, \mathbf{q}_{-i}))$, where $\mathbf{q}_{-i}$ are the service levels set by i's competitors and t_i is the access fee per service imposed by the facility manager. Following Equation (1), which describes the users' behavior, we obtain the profit function

$$\begin{aligned}\pi_i(q_i, \mathbf{q}_{-i}) &= q_i(u_i(q_i, \mathbf{q}_{-i}) - l_i^u(q_i, \mathbf{q}_{-i}) - t_i - l_i^p(q_i, \mathbf{q}_{-i})) \\ &= q_i(u_i(q_i, \mathbf{q}_{-i}) - t_i - l_i(q_i, \mathbf{q}_{-i})).\end{aligned} \tag{2}$$

The dynamics between the facility manager and the service providers are modeled as a Stackelberg game. The facility manager announces the access fee per service, $\mathbf{t}$, and the service providers then determine their appropriate output level, $\mathbf{q}(\mathbf{t})$. We assume service providers behave according to a *subgame perfect equilibrium*. That is, for a fixed access fee t_i, service provider i determines his service level to maximize his profit given the service level set by his competitors.

2.1.3. Facility Manager's Welfare Maximization Problem. The goal of the facility manager is to maximize the total social welfare (W), obtained by aggregating consumer surplus (CS), producer surplus (PS), and revenue collected from access fees (TR); i.e., $W = CS + PS + TR$.

Consumer surplus is defined as the difference between the total utility derived from consuming $\mathbf{q}$ units of services and the total cost incurred by users. The total utility for an affine marginal utility function is given by $\int_{\mathbf{q}} \mathbf{u}(\mathbf{x})\,\mathbf{dx} = \mathbf{q}^\top(\bar{\mathbf{p}} - \frac{1}{2}\mathbf{B}\mathbf{q})$. The total cost is equivalent to the full price that users perceive, $\mathbf{q}^\top\mathbf{u}(\mathbf{q})$. Thus, consumer surplus can be written as

$$\begin{aligned}CS &= \mathbf{q}^\top(\bar{\mathbf{p}} - \tfrac{1}{2}\mathbf{B}\mathbf{q}) - \mathbf{q}^\top\mathbf{u}(\mathbf{q}) \\ &= \mathbf{q}^\top(\bar{\mathbf{p}} - \tfrac{1}{2}\mathbf{B}\mathbf{q}) - \mathbf{q}^\top(\bar{\mathbf{p}} - \mathbf{B}\mathbf{q}) = \tfrac{1}{2}\mathbf{q}^\top\mathbf{B}\mathbf{q}.\end{aligned}$$

Producer surplus is the total profit generated by all service providers,

$$PS = \sum_i \pi_i(\mathbf{q}) = \mathbf{q}^\top(\mathbf{u}(\mathbf{q}) - \mathbf{l}(\mathbf{q}) - \mathbf{t}) = \mathbf{q}^\top(\bar{\mathbf{p}} - \mathbf{B}\mathbf{q} - \mathbf{l}(\mathbf{q}) - \mathbf{t}).$$

Revenue collected from congestion pricing is captured by $TR = \mathbf{q}^\top\mathbf{t}$. Combining all three terms, the total welfare is given by the following:

$$W(\mathbf{q}) = \mathbf{q}^\top(\bar{\mathbf{p}} - \tfrac{1}{2}\mathbf{B}\mathbf{q} - \mathbf{l}(\mathbf{q})). \tag{3}$$

Under Assumptions 1–4, the regulator's problem (3) is a strictly concave optimization problem with respect to the output level $\mathbf{q}$, where $\mathbf{q}^* = \arg\max_{\mathbf{q}} W(\mathbf{q})$. To achieve the maximum welfare $W(\mathbf{q}^*)$ in the three-level model, the facility manager can use the access fees so that the desirable service level $\mathbf{q}^*$ is achieved; i.e., $\mathbf{q}(\mathbf{t}^*) = \mathbf{q}^*$. The access fee could be viewed as a coordinating contract that aligns the profit-maximizing objective of service providers to one that maximizes the social welfare.

Proposition 1. *The access fee per service i is given by $t_i(\mathbf{q}) = -\beta_{ii}q_i + \sum_{j\neq i} r_{ji}q_j$, where $r_{ji} = \partial l_j/\partial q_i$.*

The access fee consists of two components of opposite signs: a subsidy that is aimed to correct underproduction and a penalty that targets overproduction by service providers who ignore the spillover cost imposed on others. In the setting without spillover cost, $t_i(\mathbf{q}) = -\beta_{ii}q_i$, implying that the facility manager has to provide appropriate subsidies to induce the socially optimal output level. In the setting with spillover cost, $\sum_{j\neq i} r_{ji}q_j$ captures the marginal negative externalities i has imposed onto others. It is important to note that the access fee does not manage self-contained cost because service providers have already taken it into consideration when they determine their output level.

Proposition 1 provides a way of using the access fee to achieve the optimal social welfare. One natural question arises: What good does it do? Clearly, any attempt to implement such a scheme on an industry-wide scale is certain to face institutional, political, and financial challenges. An answer to the question could help policy makers gauge the need for regulation. In particular, if the benefit is proven to be substantial for some instances, it may provide some evidences to support its implementation. In the next subsection, we formulate this question mathematically and introduce some proxies used in the analysis as the key performance indicators.

2.2. Performance Indicators

Efficiency analysis compares the social welfare achieved in an unregulated setting with the social optimum. It describes how efficient the unregulated setting is from the societal point of view. It also helps to pinpoint

key factors that drive efficiency in the unregulated setting. This becomes particularly important when the "optimum" might be infeasible to attain in certain applications.

In our context, service providers use the facility for free in the unregulated setting; i.e., $\mathbf{t} = \mathbf{0}$. In a socially optimal setting, $\mathbf{t}^*$ is implemented such that the coordinated output level is maximizing total welfare; i.e., $\mathbf{q}(\mathbf{t}^*) = \mathbf{q}^*$. Denote $\mathbf{q}^{\mathrm{N}}$ and $\mathbf{q}^*$ as the output level in an unregulated oligopoly and a social optimum, respectively. Let $W(\mathbf{q}^{\mathrm{N}})$ and $W(\mathbf{q}^*)$ be the corresponding total welfare attained in these two settings. The quantity of interest for efficiency analysis is $W(\mathbf{q}^{\mathrm{N}})/W(\mathbf{q}^*)$.

With nonlinear costs, it is generally hard to obtain closed-form solutions in the two settings. Nonetheless, one can derive optimality conditions that can help quantify the total welfare, as shown in Proposition 2. Because the Jacobian matrix of the cost function depends on the output level, we will use $\mathbf{R}^*$ and $\mathbf{R}^{\mathrm{N}}$ to distinguish the matrices evaluated at $\mathbf{q}^*$ and $\mathbf{q}^{\mathrm{N}}$, respectively.

Proposition 2. *Under Assumptions* 1–4, *there exists unique solutions in the social optimal and the unregulated settings. They are given by the following*:

$$W(\mathbf{q}^*) = (\mathbf{q}^*)^\top(\tfrac{1}{2}\mathbf{B} + \mathbf{R}^*)\mathbf{q}^*, \quad \textit{and}$$
$$W(\mathbf{q}^{\mathrm{N}}) = (\mathbf{q}^{\mathrm{N}})^\top(\tfrac{1}{2}\mathbf{B} + \boldsymbol{\Gamma}_{\mathbf{B}} + \boldsymbol{\Gamma}_{\mathbf{R}}^{\mathrm{N}})\mathbf{q}^{\mathrm{N}}.$$

The main focus of this work is on quantifying the efficiency of service industries for two settings depending on the presence of spillover cost. For each setting, we establish a lower bound on $W(\mathbf{q}^{\mathrm{N}})/W(\mathbf{q}^*)$, which gives the worst performance guarantee, as well as an upper bound, which sheds some insights on how to improve performance.

To establish these bounds, one of the machineries from matrix analysis that we utilize is the concept of a Jacobian similarity property.

Definition 2. In a *Jacobian similarity property*, a positive semidefinite matrix $\mathbf{F}(\mathbf{q})$ satisfies the property if there exists a constant $\kappa \geq 1$ such that for all $\mathbf{w}$, $\mathbf{q}$, and $\mathbf{q}'$,

$$\kappa \mathbf{w}^\top \mathbf{F}(\mathbf{q})\mathbf{w} \geq \mathbf{w}^\top \mathbf{F}(\mathbf{q}')\mathbf{w} \geq \frac{1}{\kappa}\mathbf{w}^\top \mathbf{F}(\mathbf{q})\mathbf{w}.$$

If $\mathbf{F}(\mathbf{q})$ is the Jacobian matrix of an affine function such that $\mathbf{F}(\mathbf{q})$ is independent of $\mathbf{q}$, the only κ that satisfies the condition is 1. When matrix $\mathbf{F}(\mathbf{q})$ is positive definite for all $\mathbf{q}$, it is easy to derive a loose bound on this constant based on its definition: $\kappa = \max_{\mathbf{q}}((\max_i \lambda_i\{\mathbf{F}(\mathbf{q})\})/(\min_i \lambda_i\{\mathbf{F}(\mathbf{q})\}))$, which is the maximum conditional number of the Jacobian matrix. We refer the reader to Perakis (2007) for more information on this concept. Later in this paper, we will show that it is possible to obtain bounds on κ without searching through the entire space of $\mathbf{q}$ and show how this term is related to nonlinearity of the cost structure.

To interpret the bounds in efficiency analysis, we will introduce the following two parameters, which measure the intensity of competition among service providers and the severity of spillover costs, respectively.

Definition 3. In a *competition index adjusted with self-contained cost*, given a price sensitivity matrix $\mathbf{B}$ and self-contained cost $\boldsymbol{\Gamma}_{\mathbf{R}}$, the competition index for service provider i is defined as $\gamma_i = \sum_{j\neq i}\beta_{ij}/\beta_{ii} + 2r_{ii}/\beta_{ii}$ for all i. Let $\bar{\gamma} = \max_i \gamma_i|_{\mathbf{q}=\mathbf{q}^*}$.

The notion that $\sum_{j\neq i}\beta_{ij}/\beta_{ii}$ is used to measure the intensity of competition can be found in Sun (2006), Farahat and Perakis (2009, 2010). If every service provider in the market changes his output level by one unit, β_{ii} reflects the amount of price change, which is solely contributed by i's own output change, while $\sum_{j\neq i}\beta_{ij}$ measures the price change contributed by i's competitors. A high value of $\sum_{j\neq i}\beta_{ij}/\beta_{ii}$ suggests that i's price is more susceptible to his competitors' output change than his own change, implying that service provider i faces a high level of competition. When $\sum_{j\neq i}\beta_{ij} = 0$ for all i, it implies that $\beta_{ij} = 0$ for all $j \neq i$. That is, each service provider acts as a monopolist and does not face any competition.

When there is self-contained cost (i.e., $r_{ii} > 0$), the competition index also contains the term r_{ii}/β_{ii}. It compares the marginal decrease in the revenue per service stemming from one's self-contained cost increase to the decrease stemming from diminishing returns of the demand. Thus, when r_{ii}/β_{ii} is large, it implies that the cost increase is steep.

Thus, γ_i measures the level of competition faced by service provider i, taking into account the self-contained cost, i.e., comparing the aggregate price impact from i's competitors and the self-contained congestion to the price change solely contributed by i's output change. With asymmetric service providers, γ_i differs across i. With nonlinear cost, γ_i also depends on the output level $\mathbf{q}$. We will use $\bar{\gamma} = \max_i \gamma_i|_{\mathbf{q}=\mathbf{q}^*}$ to approximate the competition intensity under the optimum output level.

Definition 4. To obtain the *external cost-to-benefit ratio*, given a price sensitivity matrix $\mathbf{B}$ and congestion cost $\mathbf{R}$, define $\rho_i = \sum_{j\neq i} r_{ji}/\beta_{ii}$ for all i. Let $\bar{\rho} = \max_i \rho_i|_{\mathbf{q}=\mathbf{q}^*}$.

This quantity is nonzero only when spillover cost is present. The numerator $\sum_{j\neq i} r_{ji}$ captures the marginal spillover cost created by service provider i, and the denominator reflects the additional consumer surplus when service provider i increases his output. Thus, the term ρ_i could be interpreted as the *external cost* (spillover congestion cost) versus the *external benefit* (additional consumer surplus) that service provider i

brings to the society. The maximum net externality in an oligopoly is indicated by $\bar{\rho}$; when $\bar{\rho} \le 1$, every service provider is contributing more welfare to the society than the spillover cost, so the net externality remains positive. However, when $\bar{\rho} > 1$, there exists some service provider whose spillover cost outweighs the external benefit that he brings to the society, implying negative net externality.

With nonlinear costs, the external cost-to-benefit ratio depends on $\mathbf{q}$. Similar to the competition index, this parameter is evaluated at the social optimum. Because the welfare maximization problem is a strictly concave optimization problem, $\mathbf{q}^*$ can be easily computed. We will begin the analysis for the setting without spillover cost in §3. The setting with spillover cost will be examined in §4.

3. Efficiency Analysis in the Absence of Spillover Cost

Motivated by modern technology-based (e.g., telecommunications, computing) service providers with fully self-contained costs, in this section, we focus on the setting when there is no spillover cost. The first result compares the output level in the unregulated setting with the optimum.

Proposition 3. *In the absence of spillover cost,* $\frac{1}{2}\mathbf{q}^* \le \mathbf{q}^N \le \mathbf{q}^*$.

The proposition states that the output level in the unregulated setting is always below the socially optimum level. This result seemingly contradicts the observation on wireless congestion discussed in the beginning of the paper. One plausible argument is that Proposition 3 refers to the outcome of a market that is operating in a static equilibrium, whereas the booming sales of smartphones and tablets imply a market that is far from being in a steady state. Nevertheless, later in this section, we will discuss how efficiency analysis reveals the service providers' attitude toward self-contained costs and draws connections with recent developments in the wireless communications industry.

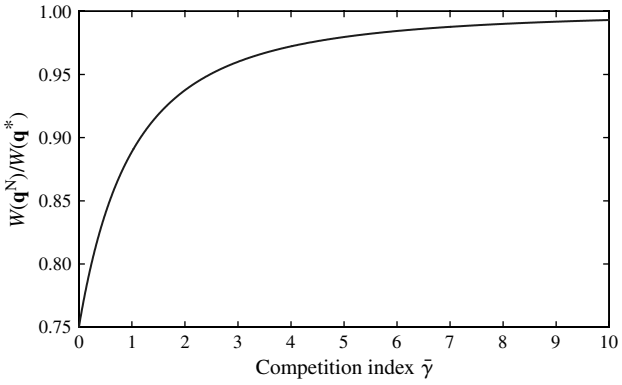
3.1. Efficiency Analysis

We present our first key result, which quantifies the efficiency analysis of the unregulated oligopoly with a constant lower bound and a parametric upper bound. The proof for the bounds is shown in §§3.2 and 3.3, respectively.

Theorem 1. *In the absence of spillover cost, under Assumptions* 1–3, *total social welfare in an unregulated oligopoly with nonlinear convex costs is bounded by*

$$\frac{3}{4} \le \frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} \le \kappa\left(1 - \frac{1}{(2+\bar{\gamma})^2}\right),$$

Figure 1 **For Symmetric Service Providers with Constant Marginal Cost, $W(\mathbf{q}^N)/W(\mathbf{q}^*) = 1 - 1/(2+\bar{\gamma})^2$ with Respect to Competition Index $\bar{\gamma}$ (i.e., When Cost Is Independent of Output Level q)**



where κ *is the Jacobian similarity factor. The lower bound is tight when service providers are noncompeting and the cost* $\mathbf{l}(\mathbf{q})$ *is independent of* $\mathbf{q}$. *The upper bound is tight when service providers are symmetric and* $\mathbf{l}(\mathbf{q})$ *is either independent of* $\mathbf{q}$ *or an affine function of* $\mathbf{q}$ (*i.e.,* $\kappa = 1$ *for both cases*).

Theorem 1 states that when costs are fully self-contained, the unregulated setting achieves at least 75% of the optimal social welfare. The worst case for efficiency loss occurs when $\mathbf{l}(\mathbf{q})$ is independent of $\mathbf{q}$ ($\kappa = 1$) and the service providers are not competing ($\bar{\gamma} = 0$). For this particular case, the upper and lower bounds in Theorem 1 coincide.

Figure 1 illustrates the upper bound with respect to the competition index when the cost $\mathbf{l}(\mathbf{q})$ is independent of $\mathbf{q}$. Note that the bound coincides with the exact value of $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ in the case with symmetric service providers. It shows that the efficiency of an unregulated oligopoly increases as the intensity of competition in the facility increases. When the service providers are independent with monopolistic power (i.e., $\bar{\gamma} = 0$), the efficiency loss is at its maximum of 25%. As the competition among the service providers intensifies, the efficiency gap between the unregulated setting and the social optimum diminishes. For example, when $\bar{\gamma} = 1$, the gap is 11.1%; as $\bar{\gamma}$ increases to 2, the gap reduces to 6.25%. It implies that in industries with a fair amount of competition, the unregulated setting could be efficient.

To explain this behavior, recall that in Proposition 3 we showed that incentive misalignment leads to underproduction in the unregulated setting. For a given cost structure, Theorem 1 shows that the extent of misalignment in incentive is directly related to the market power of service providers. When service providers have monopolistic power (i.e., no competition), the output level is at its lowest level with respect to the optimum, resulting in the lowest efficiency. As competition increases, it reduces the market power of

individual service providers and subsequently shrinks the efficiency gap. Take perfect competition as an extreme example, where $\bar{\gamma} \to \infty$; the upper bound in Theorem 1 shows that $W(\mathbf{q}^N) \to W(\mathbf{q}^*)$. That is, the total welfare obtained under an unregulated oligopoly converges to that of a fully coordinated setting.

The constant lower bound in Theorem 1 might be surprising at first, especially compared with other works on congestible games. For example, Roughgarden and Tardos (2002) and Roughgarden (2005) study the selfish behavior of noncooperative network users. In their setting, the authors show that performance degrades with nonlinearity of the latency functions. In contrast, we have shown that the performance degradation caused by noncooperative service providers is bounded, even with highly nonlinear costs. This is because when the costs are fully self-contained, the service providers have fully taken the cost impact into consideration while determining the equilibrium output. Although our setting is restricted to a competing market in an equilibrium state, there is well-documented evidence suggesting that service providers take full responsibility of self-contained cost. A recent study (Research and Markets 2013) states that the four largest carriers in the United States are going to spend $10 billion on infrastructure in 2013 as they compete to roll out newer and faster networks to meet users' rising demand for wireless connectivity. The study forecasts that the total expenditure on network infrastructure in the United States will reach $37.5 billion by 2017.

We would like to point out that in order to compute the upper bound in Theorem 1, the knowledge of the Jacobian similarity factor κ is required. In the proof of the upper bound (in §3.3), we have given a precise definition on κ (see Equation (8)). However, despite its tightness, the definition has a rather complex form and does not offer much insight. We simplify its expression by giving it an upper bound, as shown below.

Proposition 4. *Without spillover cost, the Jacobian similarity factor is bounded by* $\kappa \le \max_i (l_i'(q_i^*)/l_i'(q_i^N))^2 \le \max_i (l_i'(q_i^*)/l_i'(q_i^*/2))^2$, *where* $l_i'(\cdot) = \partial l_i/\partial q_i$.

- *For monomial cost,* $l_i(q_i) = c_i q_i^k$, *where* c_i *is the constant cost coefficient and* k *is the degree of nonlinearity, the Jacobian similarity factor is bounded by* $\kappa \le 2^{2(k-1)}$.
- *For the M/M/1 model with delay cost,* $l_i(q_i) = c_i/(\mu_i - q_i)$, *where* c_i *is congestion cost per unit time,* μ_i *is i's service rate and* q_i *is the user arrival rate. Then,* $\kappa \le \max_i ((2\mu_i - q_i^*)/(2(\mu_i - q_i^*)))^2$.

The Jacobian similarity factor κ is bounded by the maximum ratio of the marginal cost obtained in the optimal and the unregulated settings. Clearly, when $\mathbf{l}(\mathbf{q})$ is affine in $\mathbf{q}$, because its marginal cost is the same in the two settings, Proposition 4 shows that $\kappa = 1$. We have also presented an upper bound for two families of cost functions—namely, the monomial and the $M/M/1$ delay costs. To do so, we need to have a lower bound on $\mathbf{q}^N$, and we have used $\mathbf{q}^N \ge 1/2\mathbf{q}^*$ from Proposition 3. In general, one can obtain a better bound on κ by providing a tighter lower bound for $\mathbf{q}^N$.

Theorem 1 offers us some comfort in knowing that when costs are fully self-contained, the worst efficiency loss is bounded and the actual number could be considerably smaller in reality when competition exists. In the following two subsections, we will present the proof for Theorem 1. We begin with the analysis for the constant lower bound and follow with the parametric upper bound.

3.2. Proof of Theorem 1 for the Lower Bound

The proof heavily utilizes some tools from matrix analysis. This allows us to analyze multiple asymmetric players and nonlinear cost. To enhance clarity, a proof outline is given as below. It consists of three key steps:

Step 1. We make use of the convexity of the cost function (Lemma 4 in the appendix) and the optimality conditions in Proposition 2 to derive an upper bound on $W(\mathbf{q}^*)$ in terms of $\mathbf{q}^N$.

Step 2. We simplify the lower bound for $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ to a composite matrix using eigendecomposition and properties of similar matrices.

Step 3. We establish a constant lower bound on $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ by showing the composite matrix is copositive.[1] We establish copositivity by showing that the composite matrix is the product of two nonnegative matrices. Thus, it is also a nonnegative matrix that is copositive by definition.

Step 1 of the proof is labeled as Lemma 1, which also holds with spillover costs.

Lemma 1. *The optimal social welfare is bounded from above by the following*:

$$W(\mathbf{q}^*) \le \tfrac{1}{2}(\mathbf{q}^N)^\top (\mathbf{B} + \boldsymbol{\Gamma}_\mathbf{B} + \boldsymbol{\Gamma}_\mathbf{R}^N + \mathbf{R}^N)(\mathbf{B} + 2\mathbf{R}^N)^{-1} \cdot (\mathbf{B} + \boldsymbol{\Gamma}_\mathbf{B} + \boldsymbol{\Gamma}_\mathbf{R}^N + \mathbf{R}^N)\mathbf{q}^N.$$

Proof of Lemma 1. Denote $\boldsymbol{\Omega} = \mathbf{B} + \boldsymbol{\Gamma}_\mathbf{B} + \boldsymbol{\Gamma}_\mathbf{R}^N + \mathbf{R}^N$ and $\boldsymbol{\Sigma} = \mathbf{B} + \mathbf{R}^* + \mathbf{R}^N$. By Lemma 4, we have already shown that $\boldsymbol{\Omega}\mathbf{q}^N \ge \boldsymbol{\Sigma}\mathbf{q}^*$. Note that both $\boldsymbol{\Omega}\mathbf{q}^N$ and $\boldsymbol{\Sigma}\mathbf{q}^*$ are two nonnegative vectors. By Proposition 2,

$$W(\mathbf{q}^*) = (\mathbf{q}^*)^\top \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\tfrac{1}{2}\mathbf{B} + \mathbf{R}^*)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\mathbf{q}^* \le (\mathbf{q}^N)^\top \boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}(\tfrac{1}{2}\mathbf{B} + \mathbf{R}^*)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\mathbf{q}^N,$$

where we reach the inequality by replacing $\boldsymbol{\Sigma}\mathbf{q}^*$ by $\boldsymbol{\Omega}\mathbf{q}^N$. We can expand this expression further:

$$W(\mathbf{q}^*) \le \tfrac{1}{2}(\mathbf{q}^N)^\top \boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}(\mathbf{B} + 2\mathbf{R}^*)\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}\mathbf{q}^N = \tfrac{1}{2}(\mathbf{q}^N)^\top \boldsymbol{\Omega}(\mathbf{B} + 2\mathbf{R}^N)^{-0.5}$$

[1] A matrix $\mathbf{A}$ is copositive if, for any positive vector $\mathbf{x}$, $\mathbf{x}^\top\mathbf{A}\mathbf{x} \ge 0$. Clearly, every positive semidefinite matrix is copositive, but the converse is false. We refer the reader to Horn and Johnson (1985) for more information.

$$\cdot \underbrace{(\mathbf{B}+2\mathbf{R}^{\mathrm{N}})^{0.5}\boldsymbol{\Sigma}^{-1}(\mathbf{B}+2\mathbf{R}^{*})\boldsymbol{\Sigma}^{-1}(\mathbf{B}+2\mathbf{R}^{\mathrm{N}})^{0.5}}_{\boldsymbol{\Delta}}$$
$$\cdot (\mathbf{B}+2\mathbf{R}^{\mathrm{N}})^{-0.5}\boldsymbol{\Omega}\mathbf{q}^{\mathrm{N}}.$$

By the definition of the eigenvalue, this expression is bounded from above by the maximum eigenvalue of composite matrix Δ as follows:

$$W(\mathbf{q}^{*}) \le \tfrac{1}{2}\lambda_{\max}\{\boldsymbol{\Delta}\}(\mathbf{q}^{\mathrm{N}})^{\top}\boldsymbol{\Omega}(\mathbf{B}+2\mathbf{R}^{\mathrm{N}})^{-1}\boldsymbol{\Omega}\mathbf{q}^{\mathrm{N}}. \quad (4)$$

Now let us focus on this composite matrix $\boldsymbol{\Delta}$. By the property of similar matrices, $\lambda_{\max}\{\Delta\} = \lambda_{\max}\cdot\{\boldsymbol{\Sigma}^{-1}(\mathbf{B}+2\mathbf{R}^{*})\boldsymbol{\Sigma}^{-1}(\mathbf{B}+2\mathbf{R}^{\mathrm{N}})\}$. Under Assumptions 1, 2, and 4, $\lambda_{\max}\{\boldsymbol{\Delta}\} \ge 0$ since $\boldsymbol{\Delta}$ is positive semidefinite. Expanding this matrix,

$$\begin{aligned}
\lambda_{\max}\{\Delta\} &= \lambda_{\max}\{(\mathbf{B}+\mathbf{R}^{*}+\mathbf{R}^{\mathrm{N}})^{-1}(\mathbf{B}+2\mathbf{R}^{*})\\
&\qquad \cdot (\mathbf{B}+\mathbf{R}^{*}+\mathbf{R}^{\mathrm{N}})^{-1}(\mathbf{B}+2\mathbf{R}^{\mathrm{N}})\}\\
&= \lambda_{\max}\{(\mathbf{I}+(\mathbf{B}+\mathbf{R}^{*}+\mathbf{R}^{\mathrm{N}})^{-1}(\mathbf{R}^{*}-\mathbf{R}^{\mathrm{N}}))\\
&\qquad \cdot (\mathbf{I}+(\mathbf{B}+\mathbf{R}^{*}+\mathbf{R}^{\mathrm{N}})^{-1}(\mathbf{R}^{\mathrm{N}}-\mathbf{R}^{*}))\}\\
&= \lambda_{\max}\{\mathbf{I}-((\mathbf{B}+\mathbf{R}^{*}+\mathbf{R}^{\mathrm{N}})^{-1}(\mathbf{R}^{*}-\mathbf{R}^{\mathrm{N}}))^{2}\}\\
&= 1-\lambda_{\min}\{((\mathbf{B}+\mathbf{R}^{*}+\mathbf{R}^{\mathrm{N}})^{-1}(\mathbf{R}^{*}-\mathbf{R}^{\mathrm{N}}))^{2}\}.
\end{aligned}$$

It is clear that $\lambda_{\min}\{((\mathbf{B}+\mathbf{R}^{*}+\mathbf{R}^{\mathrm{N}})^{-1}(\mathbf{R}^{*}-\mathbf{R}^{\mathrm{N}}))^{2}\} \ge 0$ because it is also a positive semidefinite matrix. Thus, $\lambda_{\max}\{\Delta\} \le 1$. From Equation (4), we can conclude that $W(\mathbf{q}^{*}) \le \frac{1}{2}(\mathbf{q}^{\mathrm{N}})^{\top}\boldsymbol{\Omega}(\mathbf{B}+2\mathbf{R}^{\mathrm{N}})^{-1}\boldsymbol{\Omega}\mathbf{q}^{\mathrm{N}}$.

We then proceed to Step 2. With Proposition 2 and Lemma 1, $W(\mathbf{q}^{\mathrm{N}})/W(\mathbf{q}^{*})$ can be lower bounded as follows:

$$\begin{aligned}
\frac{W(\mathbf{q}^{\mathrm{N}})}{W(\mathbf{q}^{*})} &\ge ((\mathbf{q}^{\mathrm{N}})^{\top}(\mathbf{B}+2\boldsymbol{\Gamma}_{\mathbf{B}}+2\boldsymbol{\Gamma}_{\mathbf{R}}^{\mathrm{N}})\mathbf{q}^{\mathrm{N}})\\
&\qquad \cdot((\mathbf{q}^{\mathrm{N}})^{\top}(\mathbf{B}+\boldsymbol{\Gamma}_{\mathbf{B}}+\boldsymbol{\Gamma}_{\mathbf{R}}^{\mathrm{N}}+\mathbf{R}^{\mathrm{N}})\\
&\qquad \cdot(\mathbf{B}+2\mathbf{R}^{\mathrm{N}})^{-1}(\mathbf{B}+\boldsymbol{\Gamma}_{\mathbf{B}}+\boldsymbol{\Gamma}_{\mathbf{R}}^{\mathrm{N}}+\mathbf{R}^{\mathrm{N}})\mathbf{q}^{\mathrm{N}})^{-1}. \quad (5)
\end{aligned}$$

In the absence of spillover, $\mathbf{R}^{\mathrm{N}} = \boldsymbol{\Gamma}_{\mathbf{R}}^{\mathrm{N}}$. Because all quantities are the Nash equilibrium quantities in this proof, we drop the superscript on matrices for ease of notation:

$$\begin{aligned}
&\frac{W(\mathbf{q}^{\mathrm{N}})}{W(\mathbf{q}^{*})}\\
&\quad \ge \frac{(\mathbf{q}^{\mathrm{N}})^{\top}(\mathbf{B}+2\boldsymbol{\Gamma}_{\mathbf{R}}+2\boldsymbol{\Gamma}_{\mathbf{B}})\mathbf{q}^{\mathrm{N}}}{(\mathbf{q}^{\mathrm{N}})^{\top}(\mathbf{B}+2\boldsymbol{\Gamma}_{\mathbf{R}}+\boldsymbol{\Gamma}_{\mathbf{B}})(\mathbf{B}+2\boldsymbol{\Gamma}_{\mathbf{R}})^{-1}(\mathbf{B}+2\boldsymbol{\Gamma}_{\mathbf{R}}+\boldsymbol{\Gamma}_{\mathbf{B}})\mathbf{q}^{\mathrm{N}}}.
\end{aligned}$$

Denote $\mathbf{G} = \boldsymbol{\Gamma}_{\mathbf{B}}^{-0.5}(\mathbf{B}+2\boldsymbol{\Gamma}_{\mathbf{R}})\boldsymbol{\Gamma}_{\mathbf{B}}^{-0.5} = \boldsymbol{\Gamma}_{\mathbf{B}}^{-0.5}\tilde{\mathbf{B}}\boldsymbol{\Gamma}_{\mathbf{B}}^{-0.5}$. This is a symmetric, nonnegative positive definite matrix. We will first rewrite the right-hand side of the inequality above in terms of matrix $\mathbf{G}$ and identity matrix $\mathbf{I}$:

$$\begin{aligned}
&\frac{(\mathbf{q}^{\mathrm{N}})^{\top}(\tilde{\mathbf{B}}+2\boldsymbol{\Gamma}_{\mathbf{B}})\mathbf{q}^{\mathrm{N}}}{(\mathbf{q}^{\mathrm{N}})^{\top}(\tilde{\mathbf{B}}+\boldsymbol{\Gamma}_{\mathbf{B}})\tilde{\mathbf{B}}^{-1}(\tilde{\mathbf{B}}+\boldsymbol{\Gamma}_{\mathbf{B}})\mathbf{q}^{\mathrm{N}}}\\
&= \frac{(\mathbf{q}^{\mathrm{N}})^{\top}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\boldsymbol{\Gamma}_{\mathbf{B}}^{-0.5}(\tilde{\mathbf{B}}+2\boldsymbol{\Gamma}_{\mathbf{B}})\boldsymbol{\Gamma}_{\mathbf{B}}^{-0.5}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{q}^{\mathrm{N}}}{(\mathbf{q}^{\mathrm{N}})^{\top}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\boldsymbol{\Gamma}_{\mathbf{B}}^{-0.5}(\tilde{\mathbf{B}}+\boldsymbol{\Gamma}_{\mathbf{B}})\tilde{\mathbf{B}}^{-1}(\tilde{\mathbf{B}}+\boldsymbol{\Gamma}_{\mathbf{B}})\boldsymbol{\Gamma}_{\mathbf{B}}^{-0.5}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{q}^{\mathrm{N}}}\\
&= \frac{(\mathbf{q}^{\mathrm{N}})^{\top}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}(\mathbf{G}+2\mathbf{I})\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{q}^{\mathrm{N}}}{(\mathbf{q}^{\mathrm{N}})^{\top}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}(\mathbf{G}+\mathbf{I})\mathbf{G}^{-1}(\mathbf{G}+\mathbf{I})\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{q}^{\mathrm{N}}}\\
&= \frac{(\mathbf{q}^{\mathrm{N}})^{\top}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{G}^{-0.5}\mathbf{G}^{0.5}(\mathbf{G}+2\mathbf{I})\mathbf{G}^{0.5}\mathbf{G}^{-0.5}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{q}^{\mathrm{N}}}{(\mathbf{q}^{\mathrm{N}})^{\top}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{G}^{-0.5}\mathbf{G}^{0.5}(\mathbf{G}+\mathbf{I})\mathbf{G}^{-1}(\mathbf{G}+\mathbf{I})\mathbf{G}^{0.5}\mathbf{G}^{-0.5}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{q}^{\mathrm{N}}}\\
&= \frac{\mathbf{w}^{\top}\mathbf{G}^{0.5}(\mathbf{G}+2\mathbf{I})\mathbf{G}^{0.5}\mathbf{w}}{\mathbf{w}^{\top}\mathbf{G}^{0.5}(\mathbf{G}+\mathbf{I})\mathbf{G}^{-1}(\mathbf{G}+\mathbf{I})\mathbf{G}^{0.5}w},\\
&\qquad \text{where } \mathbf{w}=\mathbf{G}^{-0.5}\boldsymbol{\Gamma}_{\mathbf{B}}^{0.5}\mathbf{q}^{\mathrm{N}}.
\end{aligned}$$

In Step 3, we show that this ratio has a constant lower bound of 3/4, which implies that

$$\begin{aligned}
&4\mathbf{w}^{\top}\mathbf{G}^{0.5}(\mathbf{G}+2\mathbf{I})\mathbf{G}^{0.5}\mathbf{w}-3\mathbf{w}^{\top}\mathbf{G}^{0.5}(\mathbf{G}+\mathbf{I})\\
&\quad \cdot\mathbf{G}^{-1}(\mathbf{G}+\mathbf{I})\mathbf{G}^{0.5}\mathbf{w}\ge 0, \quad \text{or equivalently,}\\
&\mathbf{w}^{\top}(4\mathbf{G}^{0.5}(\mathbf{G}+2\mathbf{I})\mathbf{G}^{0.5}-3\mathbf{G}^{0.5}(\mathbf{G}+\mathbf{I})\\
&\quad \cdot\mathbf{G}^{-1}(\mathbf{G}+\mathbf{I})\mathbf{G}^{0.5})\mathbf{w}\ge 0.
\end{aligned} \quad (6)$$

To establish this statement, we will show that the composite matrix in Equation (6) is in fact copositive. We express it as follows:

$$\begin{aligned}
&4\mathbf{G}^{0.5}(\mathbf{G}+2\mathbf{I})\mathbf{G}^{0.5}-3\mathbf{G}^{0.5}(\mathbf{G}+\mathbf{I})\mathbf{G}^{-1}(\mathbf{G}+\mathbf{I})\mathbf{G}^{0.5}\\
&\quad =4(\mathbf{G}^{2}+2\mathbf{G})-3(\mathbf{G}+\mathbf{I})^{2}\\
&\quad =4\mathbf{G}^{2}+8\mathbf{G}-3\mathbf{G}^{2}-6\mathbf{G}-3\mathbf{I}\\
&\quad =\mathbf{G}^{2}+2\mathbf{G}-3\mathbf{I}\\
&\quad =(\mathbf{G}-\mathbf{I})(\mathbf{G}+3\mathbf{I}).
\end{aligned} \quad (7)$$

Given that matrix $\mathbf{G}$ is nonnegative, the second term, which is a sum with an identity matrix, is clearly nonnegative. Now consider the first term, $\mathbf{G}-\mathbf{I}$. The off-diagonal elements $\beta_{ij}/(\sqrt{\beta_{ii}\beta_{ij}})$ are nonnegative under Assumption 2, where β_{ij} is the (i,j)th element of matrix $\mathbf{B}$. Its diagonal elements are given by $(\beta_{ii}+2r_{ii})/\beta_{ii} = 1+2r_{ii}/\beta_{ii} \ge 1$, where r_{ii} is the ith diagonal element of the Jacobian matrix $\mathbf{R}$. Therefore, $\mathbf{G}-\mathbf{I}$ must also be nonnegative. We have shown that the composite matrix in Equation (6) could be expressed as the product of two nonnegative matrices. Therefore, this composite matrix must also be nonnegative and, therefore, copositive.

Finally, to show that the bound is tight, note that with noncompeting service providers, $\mathbf{B} = \boldsymbol{\Gamma}_{\mathbf{B}}$. Moreover, when $\mathbf{l}(\mathbf{q})$ is independent of $\mathbf{q}$, its Jacobian matrix $\mathbf{R}$ is $\mathbf{0}$. Using the equilibrium and optimality conditions shown in the proof for Lemma 2, it is easy to show that the service level in the unregulated setting is exactly half of the optimal level; i.e., $\mathbf{q}^{\mathrm{N}} = \frac{1}{2}\mathbf{q}^{*}$. Substituting this condition into the welfare objective function in Equation (3) shows that the ratio is exactly 3/4.

3.3. Proof of Theorem 1 for the Upper Bound

Instead of a constant bound, as shown in the previous section, the upper bound on the efficiency is a parametric function of the competition index, $\bar{\gamma}$, evaluated at

the optimal output level $\mathbf{q}^*$. There are three key steps in the proof, which are outlined as follows. The first two steps (denoted as Lemmas 2 and 3) continue to hold with spillover costs.

Step 1. We make use of Proposition 2 and the convexity of the cost (Lemma 4) to provide an upper bound on $W(\mathbf{q}^N)$. Next, we express $W(\mathbf{q}^N)$ in terms of the optimal output level $\mathbf{q}^*$ by using the Jacobian similarity property.

Step 2. We simplify the upper bound of $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ to a composite matrix using eigendecomposition and the Rayleigh–Ritz theorem.

Step 3. We characterize an upper bound on $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ in terms of the maximum eigenvalue of a composite matrix. We show that the bound can be simplified by replacing the eigenvalue with the competition index using the Gershgorin disc theorem, which bounds the spectrum of any square matrix.

Step 1 proceeds as follows.

Lemma 2. *The welfare in the unregulated setting is bounded from above by the following*: $W(\mathbf{q}^N) \le \kappa(\mathbf{q}^*)^\top(\mathbf{B}+2\mathbf{R}^*)(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^{-1}(\frac{1}{2}\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*)(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^{-1}(\mathbf{B}+2\mathbf{R}^*)\mathbf{q}^*$, *where* $\kappa \ge 1$ *is the Jacobian similarity factor.*

Proof of Lemma 2. From Proposition 2, we obtain that

$$\begin{aligned} W(\mathbf{q}^N) &= (\mathbf{q}^N)^\top(\tfrac{1}{2}\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^N)\mathbf{q}^N \\ &= (\mathbf{q}^N)^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1}(\tfrac{1}{2}\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^N)\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}\mathbf{q}^N, \end{aligned}$$

where $\boldsymbol{\Psi} = \mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^N+\mathbf{R}^*$. Making use of Lemma 4, which shows that $\boldsymbol{\Psi}\mathbf{q}^N \le (\mathbf{B}+2\mathbf{R}^*)\mathbf{q}^*$, it follows that

$$\begin{aligned} W(\mathbf{q}^N) &\le (\mathbf{q}^*)^\top(\mathbf{B}+2\mathbf{R}^*)(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^N+\mathbf{R}^N)^{-1} \\ &\quad\cdot(\tfrac{1}{2}\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^N)(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^N+\mathbf{R}^N)^{-1} \\ &\quad\cdot(\mathbf{B}+2\mathbf{R}^*)\mathbf{q}^* \\ &\le \kappa(\mathbf{q}^*)^\top(\mathbf{B}+2\mathbf{R}^*)(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^{-1} \\ &\quad\cdot(\tfrac{1}{2}\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*)(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^{-1} \\ &\quad\cdot(\mathbf{B}+2\mathbf{R}^*)\mathbf{q}^*, \end{aligned}$$

where we obtain the last inequality by using the Jacobian similarity property, and κ is defined as the maximum eigenvalue of a positive definite matrix,

$$\begin{aligned} \kappa \le \lambda_{\max}\big\{&(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^N+\mathbf{R}^N)^{-2}(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^2 \\ &\cdot(\tfrac{1}{2}\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^N)(\tfrac{1}{2}\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*)^{-1}\big\}. \end{aligned} \tag{8}$$

Note that a more intuitive (but less tight) upper bound for κ is given in Proposition 4.

We then proceed to Step 2.

Lemma 3. $W(\mathbf{q}^N)/W(\mathbf{q}^*) \le \kappa(1-\lambda_{\min}\{((\mathbf{I}-\boldsymbol{\Xi})(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1})^2\})$, *where* $\kappa \ge 1$ *is the Jacobian similarity factor,* $\mathbf{G} = \boldsymbol{\Gamma}_\mathbf{B}^{-0.5}(\mathbf{B}+2\boldsymbol{\Gamma}_\mathbf{R}^*)\boldsymbol{\Gamma}_\mathbf{B}^{-0.5}$ *and* $\boldsymbol{\Xi} = \boldsymbol{\Gamma}_\mathbf{B}^{-0.5}\mathbf{R}_{\text{off}}^*\boldsymbol{\Gamma}_\mathbf{B}^{-0.5}$.

Proof of Lemma 3. By combining Lemma 2 and $W(\mathbf{q}^*)$ from Proposition 2, we establish an upper bound on $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ as follows:

$$\begin{aligned} &\frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} \\ &\le \kappa\frac{(\mathbf{q}^*)^\top(\mathbf{B}+2\mathbf{R}^*)(\frac{1}{2}\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*)(\mathbf{B}+2\mathbf{R}^*)\mathbf{q}^*}{(\mathbf{q}^*)^\top(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)(\frac{1}{2}\mathbf{B}+\mathbf{R}^*)(\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{B}+\boldsymbol{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)\mathbf{q}^*}. \end{aligned}$$

Because all quantities are in the socially optimum setting, we will skip the superscript on matrices. Let $\mathbf{G} = \boldsymbol{\Gamma}_\mathbf{B}^{-0.5}(\mathbf{B}+2\boldsymbol{\Gamma}_\mathbf{R})\boldsymbol{\Gamma}_\mathbf{B}^{-0.5}$, and let $\boldsymbol{\Xi} = \boldsymbol{\Gamma}_\mathbf{B}^{-0.5}\mathbf{R}_{\text{off}}\boldsymbol{\Gamma}_\mathbf{B}^{-0.5}$; the expression becomes

$$\begin{aligned} &\frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} \\ &\le \kappa\frac{(\mathbf{q}^*)^\top\boldsymbol{\Gamma}_\mathbf{B}^{0.5}(\mathbf{G}+2\mathbf{I})\boldsymbol{\Gamma}_\mathbf{B}^{0.5}\mathbf{q}^*}{(\mathbf{q}^*)^\top\boldsymbol{\Gamma}_\mathbf{B}^{0.5}(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})(\mathbf{G}+2\boldsymbol{\Xi})^{-1}(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})\boldsymbol{\Gamma}_\mathbf{B}^{0.5}\mathbf{q}^*}. \end{aligned}$$

Using the Rayleigh–Ritz theorem, the upper bound can be simplified as follows:

$$\begin{aligned} \frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} &\le \kappa\lambda_{\max}\{(\mathbf{G}+2\mathbf{I})(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1}(\mathbf{G}+2\boldsymbol{\Xi}) \\ &\qquad\cdot(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1}\} \\ &= \kappa\lambda_{\max}\{(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi}+\mathbf{I}-\boldsymbol{\Xi})(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1} \\ &\qquad\cdot(\mathbf{G}+\boldsymbol{\Xi}+\mathbf{I}-\mathbf{I}+\boldsymbol{\Xi})(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1}\} \\ &= \kappa\lambda_{\max}\{(\mathbf{I}+(\mathbf{I}-\boldsymbol{\Xi})(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1})(\mathbf{I}-(\mathbf{I}-\boldsymbol{\Xi}) \\ &\qquad\cdot(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1})\} \\ &= \kappa\lambda_{\max}\{\mathbf{I}-((\mathbf{I}-\boldsymbol{\Xi})(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1})^2\} \\ &= \kappa(1-\lambda_{\min}\{((\mathbf{I}-\boldsymbol{\Xi})(\mathbf{G}+\mathbf{I}+\boldsymbol{\Xi})^{-1})^2\}). \end{aligned}$$

In Step 3, we note that in the absence of spillover cost, $\mathbf{R}_{\text{off}} = \mathbf{0}$ or $\boldsymbol{\Xi} = \mathbf{0}$, and the upper bound in Lemma 3 can be simplified into

$$\begin{aligned} \frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} &\le \kappa(1-\lambda_{\min}\{(\mathbf{G}+\mathbf{I})^{-2}\}) \\ &= \kappa\left(1-\frac{1}{\lambda_{\max}(\mathbf{G}+\mathbf{I})^2}\right) \\ &\le \kappa\left(1-\frac{1}{(1+\lambda_{\max}\{\mathbf{G}\})^2}\right). \end{aligned} \tag{9}$$

Note that $\mathbf{G} = \boldsymbol{\Gamma}_\mathbf{B}^{-0.5}(\mathbf{B}+2\boldsymbol{\Gamma}_\mathbf{R})\boldsymbol{\Gamma}_\mathbf{B}^{-0.5}$ is a symmetric, positive definite matrix. By the property of similar matrices, this matrix is similar to $\boldsymbol{\Gamma}_\mathbf{B}^{-1}(\mathbf{B}+2\boldsymbol{\Gamma}_\mathbf{R})$. Thus, they share the same set of eigenvalues. Using the Gershgorin disc theorem, we can add an upper bound to the maximum eigenvalue of a matrix; i.e.,

$$\begin{aligned} \lambda_{\max}\{\mathbf{G}\} = \lambda_{\max}\{\boldsymbol{\Gamma}_\mathbf{B}^{-1}(\mathbf{B}+2\boldsymbol{\Gamma}_\mathbf{R})\} &\le 1+\max_i\frac{\sum_{j\ne i}\beta_{ij}+2r_{ii}}{\beta_{ii}} \\ &\le 1+\max_i\gamma_i^* = 1+\bar{\gamma}. \end{aligned}$$

Substituting these two inequalities into Equation (9), we obtain the desired bound.

To show the tightness result, note that when the cost is affine, $\kappa = 1$. The inequalities in Lemmas 2 and 3 become equalities. Moreover, with symmetric service providers, $\gamma_i = \bar{\gamma}$ for all i, one can show that the maximum eigenvalue of matrix $\mathbf{G}$ is exactly equal to $1 + \bar{\gamma}$.

4. Efficiency Analysis in the Presence of Spillover Cost

In the previous section, we saw that, in the absence of spillover cost, the efficiency loss of an unregulated setting is always bounded by 25% of the optimum welfare, irrespective of the nonlinearity of cost functions. In this section, we present the efficiency analysis with spillover costs. This analysis reveals that the efficiency depends heavily on $\bar{\rho}$ (the maximum external cost-to-benefit ratio evaluated at the socially optimal output level), and the loss in efficiency could be much more severe for this setting.

Theorem 2. *With spillover congestion cost, under Assumptions* 1–4, *the efficiency of an unregulated oligopoly depends on* $\bar{\rho}$.

(a) *When* $\bar{\rho} \le 1$,

$$\frac{3}{4} \le \frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} \le \kappa\left(1 - \left(\frac{1-\bar{\rho}}{\bar{\rho}+2+\bar{\gamma}}\right)^2\right),$$

where $\kappa \ge 1$ *is the Jacobian similarity factor. The lower bound is tight when service providers are noncompeting and the cost* $\mathbf{l}(\mathbf{q})$ *is independent of* $\mathbf{q}$. *The upper bound is tight when service providers are symmetric and* $\mathbf{l}(\mathbf{q})$ *is either independent of* $\mathbf{q}$ *or an affine function of* $\mathbf{q}$.

(b) *When* $\bar{\rho} \ge 1$,

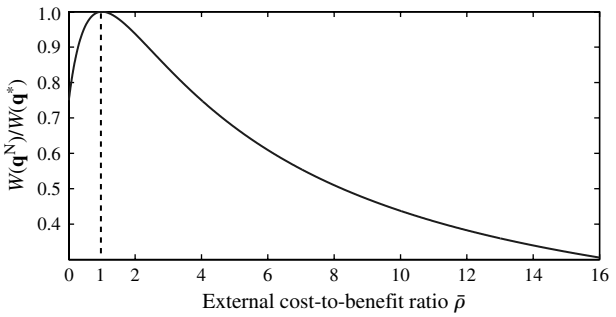
$$\frac{1}{\kappa'}\left(1 - \left(\frac{\bar{\rho}-1}{\bar{\rho}+1+\max(1-\bar{\gamma},0)}\right)^2\right) \le \frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} \le 1,$$

where $\kappa' \ge 1$ *is the Jacobian similarity factor. The lower bound is asymptotically tight when the cost* $\mathbf{l}(\mathbf{q})$ *is an affine function of* $\mathbf{q}$ *and* $\bar{\gamma} \to 0$. *The upper bound is tight with symmetric service providers and the cost-to-benefit ratio* $\bar{\rho} = 1$.

Remark 2. When there is no spillover cost (i.e., $\bar{\rho} = 0$), Theorem 2, part (a) coincides with Theorem 1.

Theorem 2 states that the maximum benefit-to-cost ratio $\bar{\rho}$ has a critical value of 1 as it has an opposite impact on the efficiency of an unregulated oligopoly depending on whether $\bar{\rho} \ge 1$. To illustrate this effect, Figure 2 depicts $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ against $\bar{\rho}$ for noncompeting service providers with affine cost $\kappa = \kappa' = 1$. We set $\bar{\gamma}$ to 0 so as to isolate the impact of competition and self-contained cost from $\bar{\rho}$. Note that in this case, the upper bound in Theorem 2, part (a) and the lower bound in Theorem 2, part (b) coincide. As shown in Figure 2, efficiency increases with $\bar{\rho}$ when $\bar{\rho} \le 1$ and decreases after $\bar{\rho}$ exceeds 1. It is important to note that, even with affine cost, efficiency loss can be significant with a large $\bar{\rho}$.

Figure 2 **For Symmetric Noncompeting Service Providers with Affine Cost, $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ with Respect to Spillover Cost-to-Benefit Ratio $\bar{\rho}$**



When an additional user is enrolled, two types of externalities are imposed on the society. The positive externality comes from additional consumer surplus through acquiring the service, whereas the negative externality is due to the existence of spillover cost. By definition, $\bar{\rho}$ is the ratio between the marginal spillover cost to the marginal consumer surplus. When $\bar{\rho} \le 1$, the marginal consumer surplus outweighs the marginal spillover cost, implying that enrolling an additional user leads to a net positive welfare to the society. As a result, the efficiency of a regulated oligopoly improves with this quantity. When $\bar{\rho}$ reaches the critical value of 1, the marginal spillover is completely offset by consumer surplus, and $W(\mathbf{q}^N) = W(\mathbf{q}^*)$. When $\bar{\rho} > 1$, with every additional enrollment, the increase in spillover cost outweighs the welfare gain, resulting in a net negative welfare change. Therefore, as $\bar{\rho}$ continues to increase, the gap between the unregulated setting and the social optimum widens.

We can also use $\bar{\rho}$ to draw a connection to the optimal access fee and the output level in the unregulated oligopoly. We have seen in Proposition 1 that when $\beta_{ii}q_i > \sum_{j\neq i} r_{ji}q_j$ or $\sum_{j\neq i} r_{ji}q_j/(\beta_{ii}q_i) \le 1$, the access fee is a subsidy that aims to promote higher output level. One can show that with asymmetric service providers, $\bar{\rho} \le 1$ indicates that $\mathbf{q}^N \le \mathbf{q}^*$. On the other hand, when $\beta_{ii}q_i < \sum_{j\neq i} r_{ji}q_j$, the access fee is a penalty to reduce the output level, implying the existence of congestion in the unregulated oligopoly. This helps to explain the similarity in the efficiency result with only self-contained costs (Theorem 1) and the setting with spillover costs with $\bar{\rho} \le 1$ (Theorem 2, part (a)). In both cases, the unregulated oligopoly is underproducing compared with its optimal counterpart. Competition among service providers reduces the oligopolistic power and consequently promotes its efficiency.

With nonlinear spillover costs, we have two Jacobian similarity factors, i.e., κ in the upper bound for $\bar{\rho} \le 1$ and κ' in the lower bound for $\bar{\rho} \ge 1$. Their precise definitions are given in Equations (8) and (18), respectively. The following result provides a more intuitive upper bound that shows their dependence on the cost structure.

Proposition 5. (a) *With spillover costs, the Jacobian similarity factors are bounded by* $\kappa \le \lambda_{\max}\{(\mathbf{\Gamma}_{\mathbf{R}}^{\mathrm{N}} + \mathbf{R}^{\mathrm{N}})^{-2} \cdot (\mathbf{\Gamma}_{\mathbf{R}}^{*} + \mathbf{R}^{*})^{2}\}$, *and* $\kappa' \le \lambda_{\max}\{(\mathbf{\Gamma}_{\mathbf{R}}^{\mathrm{N}} + \mathbf{R}^{\mathrm{N}})^{2}(\mathbf{\Gamma}_{\mathbf{R}}^{*} + \mathbf{R}^{*})^{-2}\}$. *When service providers are symmetric with* $u_i(\mathbf{q}) = \bar{p} - \beta_i q_i - \sum_j \beta_j q_j$ *and* $r_{ij} = r$ *for all* $i, j \in \{1, \ldots, n\}$, *these bounds simplify to* $\kappa \le (l'(\mathbf{q}^*)/l'(\mathbf{q}^{\mathrm{N}}))^2$ *and* $\kappa' \le (l'(\mathbf{q}^{\mathrm{N}})/l'(\mathbf{q}^*))^2$.

(b) *Denote as* $\hat{q} = \bar{p}/(2\beta_i + (n-1)\beta_j)$ *the output level in the unregulated oligopoly without congestion cost.*

- *For monomial costs,* $l_i(\mathbf{q}) = c(\sum_j q_j)^k$, *where* $c > 0$ *is the constant cost coefficient and* k *is the degree of nonlinearity, the Jacobian similarity factor is bounded by* $\kappa \le 2^{2(k-1)}$ *and* $\kappa' \le (\hat{q}/q^*)^{k-1}$.
- *For the M/M/1 model with delay cost,* $l_i(\mathbf{q}) = c/(\mu - \sum_j q_j)$, *where* c *is congestion cost per unit time and* μ *is the service rate. Then,* $\kappa \le \max_i((2\mu - nq^*)/(2(\mu - nq^*)))^2$ *and* $\kappa' \le ((\mu - nq^*)/(\mu - n\hat{q}))^2$.

The method to find an upper bound for κ without knowing $\mathbf{q}^{\mathrm{N}}$ is very similar to the setting with only self-contained cost. We use a lower bound $\mathbf{q}^{\mathrm{N}} \ge \frac{1}{2}\mathbf{q}^*$ because $\bar{\rho} \le 1$ indicates that $\mathbf{q}^{\mathrm{N}} \le \mathbf{q}^*$. To determine κ', as $\mathbf{q}^{\mathrm{N}} \ge \mathbf{q}^*$, we need to provide an upper bound on $\mathbf{q}^{\mathrm{N}}$, which is denoted as $\hat{\mathbf{q}}$. One such lower bound, which we have used in Proposition 5, is to let $\hat{\mathbf{q}}$ be the output level in the absence of congestion costs.

In the proof of Theorem 2, we express the Jacobian matrix $\mathbf{R}$ as $\mathbf{\Gamma}_{\mathbf{R}}$ and $\mathbf{R}_{\mathrm{off}}$ (representing the self-contained and the spillover costs) and show that some steps for Theorem 1 continue to hold. Because the main techniques are similar to what is shown in §§3.2 and 3.3, we relegate the proof of Theorem 2 to the appendix.

5. Simulation Experiments

So far, we have developed several bounds to evaluate the efficiency of an unregulated oligopoly. We show that the bounds are tight for special cases, such as noncompeting service providers or symmetric service providers with affine costs. This section addresses a natural follow-up question: How "good" are our bounds for more general cases?

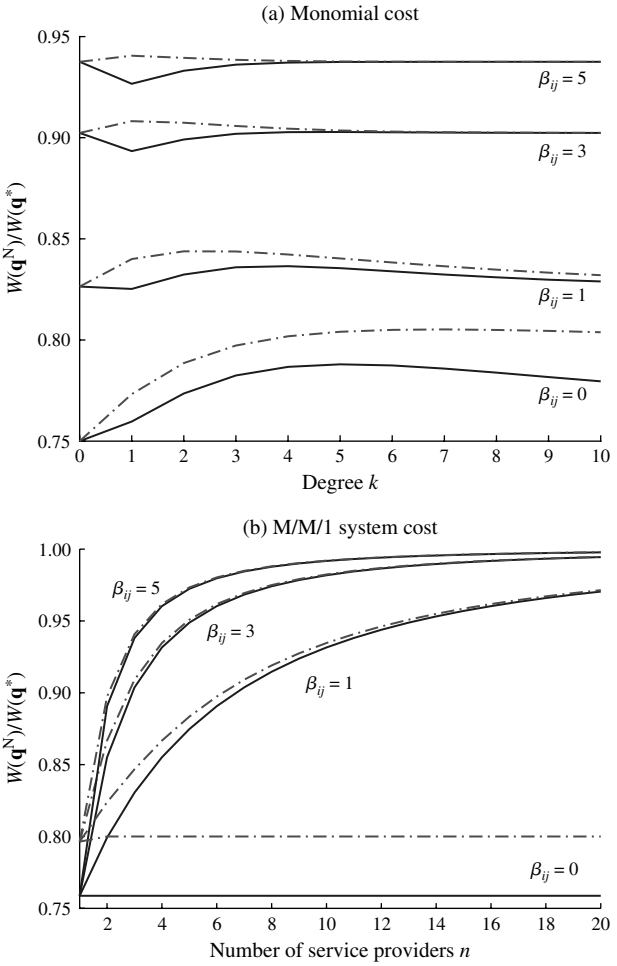
Two factors affect the tightness of our bounds—nonlinearity of the cost functions and asymmetry among service providers. To pinpoint the individual impact on the bound performance, we first focus on symmetric service providers and evaluate the impact of cost nonlinearity in §5.1. Next, in §5.2, we focus on studying the impact of asymmetry among service providers when the costs are affine functions. Because the performance on the parametric bounds is similar across different settings, we will illustrate the upper bound result from Theorem 1 in §5.1 and that from Theorem 2 in §5.2.

5.1. Effect of Nonlinearity

We conduct two experiments with different nonlinear cost functions for symmetric service providers to evaluate the impact of nonlinearity of the costs on the tightness of the bounds. In both experiments, for the marginal utility function $p_i(q) = \bar{p}_i - \beta_{ii}q_i + \sum_{j \ne i} \beta_{ij} q_j$, we let $\bar{p}_i = 20$, $\beta_{ii} = 10$ for all i. We allow β_{ij} to take values from $[0, 1, 3, 5]$ for all $j \ne i$. Increasing the β_{ij}s while keeping the β_{ii}s fixed implies that the intensity of competition is increasing. Note that $\beta_{ij} = 0$ represents noncompeting service providers.

In the first experiment, we consider a monomial cost function, i.e., $l_i(q_i) = c_i q_i^k$, where $c_i = 1$ is the constant cost coefficient and the power k denotes the degree of nonlinearity. Figure 3, panel (a) shows the actual

Figure 3 Simulation Experiments Illustrating the Performance of the Bound with Respect to Nonlinearity



Note. The solid and dashed lines represent the actual efficiency ratio and the bound, respectively.

efficiency ratio $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ for a setting with $n=5$ service providers and the upper bound developed in Theorem 1 against the nonlinearity factor k, depicted by the solid and dashed lines, respectively.

In the second experiment, we consider the cost function from a standard $M/M/1$ system, $l_i(q_i) = c_i/(\mu_i - q_i)$, where $c_i = 1$ denotes the congestion cost per unit time and μ_i denotes i's service rate and $\mu_i = 10$ for all i. Besides increasing β_{ij}s, as we have done in the first experiment, we also vary the number of service providers from 1 to 20. The exact efficiency ratio and the upper bound are shown in Figure 3, panel (b).

In both experiments, we observe that the worst case for the efficiency loss occurs with noncompeting service providers. The exact efficiency ratio $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ improves as competition increases, either by increasing the β_{ij}s or by increasing the number of service providers (except the case of noncompeting service providers). We also observe from Figure 3, panel (b) that the efficiency increase from having an additional service provider is more pronounced when there is a fair amount of competition in the existing market. For example, with $\beta_{ij} = 5$, when the market expands from a monopoly ($n=1$) to a duopoly ($n=2$), the increase in $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ is approximately 9%.

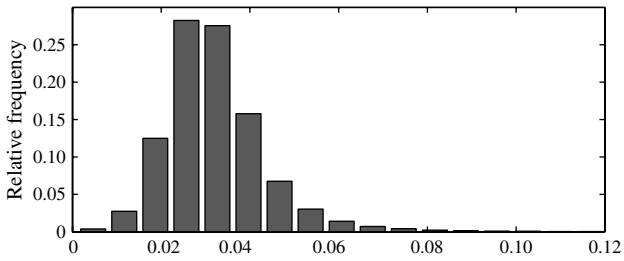
The impact of varying the degree of nonlinearity k in the first experiment is less conclusive. As we see in Figure 3, panel (a), efficiency first decreases, then increases with k. However, with competing service providers, the general trend seems to suggest that the effect of cost nonlinearity on efficiency is limited because the variations in $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ across k are rather small.

The bounds in both experiments appear to follow the exact ratio closely. The error between the bound and the exact ratio is at its largest with noncompeting service providers. With competing service providers, the bound becomes considerably tighter as competition intensity increases. Figure 3, panel (a) shows that with competing service providers, when $k \geq 5$, the bound matches the exact ratio for $\beta_{ij} \geq 3$. In Figure 3, panel (b), with noncompeting service providers, the error between the bound and the exact efficiency ratio is about 4%. When $\beta_{ij} \geq 3$, the differences are reduced to less than 1%.

5.2. Effect of Asymmetry

In this section, we focus on a setting with asymmetric service providers and affine spillover costs so as to evaluate the performance of the upper bound in Theorem 2 (when $\bar{\rho} \leq 1$). Figure 4 reports the result of a simulation experiment with 500,000 instances. For each instance, a number corresponding to the number of service providers is drawn from a discrete uniform distribution on the interval $[2, 25]$. Next, we generate a random positive vector $\bar{\mathbf{p}}$, a positive definite matrix with nonnegative elements $\mathbf{B}$, and a positive semidefinite matrix $\mathbf{R}$ of the corresponding dimension to represent the asymmetric service providers. (Because costs are affine, $\mathbf{R}$ is independent of $\mathbf{q}$.)

Figure 4 **A Simulation Experiment to Illustrate the Strength of the Upper Bound in Theorem 2 When $\bar{\rho} \leq 1$**



Notes. The x axis represents the differences between the exact value of $W(\mathbf{q}^N)/W(\mathbf{q}^*)$ and the upper bound grouped in bins. The y axis represents the relative frequency of the instances within the bin size.

We first solve the problem in the unregulated and the optimal settings and compute the exact value for $W(\mathbf{q}^N)/W(\mathbf{q}^*)$. Next, we determine the corresponding upper bound in terms of $\bar{\gamma}$ and $\bar{\rho}$. We summarize the differences between the exact quantity and the lower bound in the histogram as shown in Figure 4. The experiment suggests that the upper bound provides a fairly accurate estimate of the exact quantity. For most of the 500,000 instances, the differences between the two quantities are within 0.05.

6. Conclusions

In this work, we have considered a setting where several service providers compete for users who are sensitive to both price and congestion by providing multiple differentiated services. We have shown that, in the absence of spillover cost, the unregulated setting could be efficient and the maximum welfare loss capped at 25%, even with highly nonlinear convex costs. Competition among service providers promotes efficiency in an unregulated setting. With spillover cost, the efficiency of the unregulated setting highly depends on the relative magnitude of the marginal spillover cost and the marginal consumer surplus associated with enrolling an additional user. When the marginal benefit outweighs the marginal spillover cost, enrolling more users improves the efficiency of the unregulated setting. On the other hand, when the marginal spillover cost exceeds the marginal benefit, the loss of efficiency can be potentially severe because it increases with the marginal spillover costs.

Acknowledgments

The authors thank the associate editor and the referees for their insightful comments and suggestions. They helped the authors improve both the content and exposition of this work. The authors also acknowledge the support from the National

Science Foundation [Awards CMMI-0758061, EFRI-0735905, and CMMI-1162034].

Appendix. Proofs for the Key Results

A. Proof of Proposition 1

One way to determine the optimal access fee $\mathbf{t}^*$ that induces the welfare-maximizing output level $\mathbf{q}^* = \mathbf{q}(\mathbf{t}^*)$ is to use backward induction: substitute the service providers' best response function $\mathbf{q}(\mathbf{t})$ into the welfare objective in Equation (3) and maximize it with respect to $\mathbf{t}$. For a given access fee $\mathbf{t}$, the best response function $\mathbf{q}(\mathbf{t})$ must satisfy the equilibrium condition for all i:

$$\bar{p}_i - t_i - l_i(\mathbf{q}(\mathbf{t})) - \sum_j \beta_{ij} q_j(\mathbf{t}) - \beta_{ii} q_i(\mathbf{t}) - q_i(\mathbf{t}) \frac{\partial l_i(\mathbf{q}(\mathbf{t}))}{\partial q_i} = 0. \quad (10)$$

The optimal service level, $\mathbf{q}^*$ or $\mathbf{q}(\mathbf{t}^*)$, must satisfy the following optimality condition with respect to Equation (3):

$$\bar{p}_i - l_i(\mathbf{q}) - \sum_j \beta_{ij} q_j - \sum_j \frac{q_j^* \partial l_j(\mathbf{q})}{\partial q_i} = 0. \quad (11)$$

Since $\mathbf{q}(\mathbf{t})$ must also satisfy the condition above, we obtain the desired result by equating Equation (11) with (10) and solve for t_i.

B. Proof of Proposition 2

Under Assumptions 1–4, the regulator's problem (3) is a strictly concave optimization problem with respect to the output level $\mathbf{q}$. The optimality condition for the problem in Equation (3) can be rewritten in matrix form, $\nabla W(\mathbf{q}) = \bar{\mathbf{p}} - \mathbf{B}\mathbf{q} - \mathbf{l}(\mathbf{q}) - \mathbf{R}\mathbf{q} = \mathbf{0}$, where matrix $\mathbf{R}$ denotes the Jacobian matrix of function $\mathbf{l}(\mathbf{q})$. It is important to note that the matrix $\mathbf{R}$ depends on the output level $\mathbf{q}$. Since $\nabla W(\mathbf{q}^*) = 0$, we get

$$\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^*) = (\mathbf{B} + \mathbf{R}^*)\mathbf{q}^*, \quad \text{or} \quad (12)$$

$$\mathbf{q}^* = (\mathbf{B} + \mathbf{R}^*)^{-1}(\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^*)). \quad (13)$$

Substituting Equation (13) into the welfare objective (Equation (3)), we obtain the following:

$$\begin{aligned} W(\mathbf{q}^*) &= (\mathbf{q}^*)^\top (\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^*) - \tfrac{1}{2}\mathbf{B}(\mathbf{B} + \mathbf{R}^*)^{-1}(\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^*))) \\ &= (\mathbf{q}^*)^\top (\tfrac{1}{2}\mathbf{B} + \mathbf{R}^*)(\mathbf{B} + \mathbf{R}^*)^{-1}(\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^*)) \\ &= (\mathbf{q}^*)^\top (\tfrac{1}{2}\mathbf{B} + \mathbf{R}^*)\mathbf{q}^*. \end{aligned}$$

In the unregulated setting, there is no access fee; i.e., $\mathbf{t} = 0$. To show the existence and uniqueness of the equilibrium, under Assumptions 1–3, the strategy space for each service provider is compact and convex. The payoff function is continuous and concave with respect to his own strategy. Therefore, the existence of the equilibrium is guaranteed (Debreu 1952). To show uniqueness, the Hessian matrix of the payoff function can be expressed as $-(\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{R} + \mathbf{H}\,\mathrm{diag}(\mathbf{q}))$, where $\mathbf{H}$ is the Jacobian matrix for $(l_i'(\mathbf{q}), \ldots, l_n'(\mathbf{q}))$ with $l_i' = \partial l_i / \partial q_i$, and $\mathrm{diag}(\mathbf{q})$ is a nonnegative matrix with $\mathbf{q}$ on its diagonal. Thus, under Assumptions 1, 2, and 4 (note that Assumption 4 is only needed for the spillover cost case), the uniqueness of the equilibrium follows from the observation that the Hessian matrix is negative quasidefinite (Rosen 1965).

The equilibrium condition for all service providers in Equation (2) can be written in the matrix form: $\bar{\mathbf{p}} - \mathbf{B}\mathbf{q} - \mathbf{l}(\mathbf{q}) - \mathbf{\Gamma_{Bq}} - \mathbf{\Gamma_{Rq}}$, where $\mathbf{\Gamma_B}$ and $\mathbf{\Gamma_R}$ represent the diagonal matrix of $\mathbf{B}$ and $\mathbf{R}$, respectively. Since $\mathbf{q}^N$ satisfies the equilibrium condition, we obtain

$$\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^N) = (\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N})\mathbf{q}^N, \quad \text{or} \quad (14)$$

$$\mathbf{q}^N = (\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N})^{-1}(\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^N)). \quad (15)$$

Substituting Equation (15) into the welfare objective gives the welfare achieved in the unregulated setting:

$$\begin{aligned} W(\mathbf{q}^N) &= (\mathbf{q}^N)^\top (\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^N) - \tfrac{1}{2}\mathbf{B}(\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N})^{-1}(\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^N))) \\ &= (\mathbf{q}^N)^\top (\tfrac{1}{2}\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N})(\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N})^{-1}(\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^N)) \\ &= (\mathbf{q}^N)^\top (\tfrac{1}{2}\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N})\mathbf{q}^N. \end{aligned}$$

C. Proof of Proposition 3

When costs are fully self-contained, $\mathbf{R} = \mathbf{\Gamma_R}$. Comparing the optimality condition (Equation (12)) and the equilibrium condition for the unregulated setting (Equation (14)), it is easy to see that $\mathbf{q}^N \le \mathbf{q}^*$ as service providers take all the costs into account.

To show the lower bound on $\mathbf{q}^N$, consider a special case with a monopolist whose cost is independent of output level, i.e., $\mathbf{B} = \beta$, $\mathbf{l}(\mathbf{q}) = l$, and $\mathbf{R} = 0$. Then, the unregulated problem is reduced to a one-player concave optimization problem. It is straightforward to show that $q^N = \frac{1}{2}(\bar{p} - l)/\beta$ while the optimum is $q^* = (\bar{p} - l)/\beta$. We will show that $\mathbf{q}^N/\mathbf{q}^* = 1/2$ is the worst case for all convex costs in the proof for Theorem 1 when we prove the bounds on efficiency analysis.

D. Lemma 4 and Its Proof

Lemma 4. *When each component of the cost function* $\mathbf{l}(\mathbf{q}) = (l_1(\mathbf{q}), \ldots, l_n(\mathbf{q}))$ *is a convex function with respect to vector* $\mathbf{q}$,
(a) $(\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N} + \mathbf{R^N})\mathbf{q}^N \ge (\mathbf{B} + \mathbf{R}^* + \mathbf{R^N})\mathbf{q}^*$, *and*
(b) $(\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N} + \mathbf{R}^*)\mathbf{q}^N \le (\mathbf{B} + 2\mathbf{R}^*)\mathbf{q}^*$.

Remark 3. When the cost function $\mathbf{l}(\mathbf{q})$ is an affine function of $\mathbf{q}$, then the Jacobian matrix $\mathbf{R}$ is independent from the output level $\mathbf{q}$. The inequalities in Lemma 4 become equalities.

Proof. To show Lemma 4, part (a), by convexity on the cost function $\mathbf{l}(\mathbf{q}) = (l_1(\mathbf{q}), \ldots, l_n(\mathbf{q}))$, we have

$$\begin{aligned} &\mathbf{l}(\mathbf{q}^*) - \mathbf{l}(\mathbf{q}^N) \ge \mathbf{R}^N(\mathbf{q}^* - \mathbf{q}^N) \\ &\quad \Rightarrow \ \mathbf{l}(\mathbf{q}^*) - \mathbf{R}^N\mathbf{q}^* \ge \mathbf{l}(\mathbf{q}^N) - \mathbf{R}^N\mathbf{q}^N \\ &\quad \Rightarrow \ -\mathbf{l}(\mathbf{q}^*) + \mathbf{R}^N\mathbf{q}^* \le -\mathbf{l}(\mathbf{q}^N) + \mathbf{R}^N\mathbf{q}^N. \end{aligned}$$

Adding a positive vector $\bar{\mathbf{p}}$ to both sides maintains the inequality; i.e.,

$$\bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^*) + \mathbf{R}^N\mathbf{q}^* \le \bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^N) + \mathbf{R}^N\mathbf{q}^N.$$

After substituting the optimality conditions derived in Equations (12) and (14), we obtain $(\mathbf{B} + \mathbf{\Gamma_B} + \mathbf{\Gamma_R^N} + \mathbf{R^N})\mathbf{q}^N \ge (\mathbf{B} + \mathbf{R}^* + \mathbf{R^N})\mathbf{q}^*$.

To establish Lemma 4, part (b), with a similar argument,

$$\begin{aligned} &\mathbf{l}(\mathbf{q}^N) - \mathbf{l}(\mathbf{q}^*) \ge \mathbf{R}^*(\mathbf{q}^N - \mathbf{q}^*) \\ &\quad \Rightarrow \ \mathbf{l}(\mathbf{q}^N) - \mathbf{R}^*\mathbf{q}^N \ge \mathbf{l}(\mathbf{q}^*) - \mathbf{R}^*\mathbf{q}^* \\ &\quad \Rightarrow \ \bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^N) + \mathbf{R}^*\mathbf{q}^N \le \bar{\mathbf{p}} - \mathbf{l}(\mathbf{q}^*) + \mathbf{R}^*\mathbf{q}^*. \end{aligned}$$

We obtain the desired result by substituting the optimality conditions in Equations (12) and (14).

E. Proof for Proposition 4

Without spillover cost, replace $\mathbf{R}$ with $\mathbf{\Gamma_R}$ in the definition for κ shown in Equation (8). Then we obtain

$$\begin{aligned}\kappa &\le \lambda_{\max}\{(\mathbf{B}+\mathbf{\Gamma_B}+2\mathbf{\Gamma_R^N})^{-2}(\mathbf{B}+\mathbf{\Gamma_B}+2\mathbf{\Gamma_R^*})^2(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N})\\ &\qquad\cdot(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*})^{-1}\}\\ &\le \lambda_{\max}\{(\mathbf{B}+\mathbf{\Gamma_B}+2\mathbf{\Gamma_R^N})^{-2}(\mathbf{B}+\mathbf{\Gamma_B}+2\mathbf{\Gamma_R^*})^2\}\\ &\quad\cdot\lambda_{\max}\{(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N})(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*})^{-1}\}. \qquad (16)\end{aligned}$$

Upper bounded by the maximum eigenvalue of a composite matrix, κ is further bounded by the product of two maximum eigenvalues.

We have shown that $\mathbf{q}^N \le \mathbf{q}^*$ in Proposition 3; therefore, by monotonicity of cost functions, we obtain $\mathbf{\Gamma_R^N} \le \mathbf{\Gamma_R^*}$. As a result, the second maximum eigenvalue in Equation (16) must be less than 1. Thus, $\kappa \le \lambda_{\max}\{(\mathbf{B}+\mathbf{\Gamma_B}+2\mathbf{\Gamma_R^N})^{-2}(\mathbf{B}+\mathbf{\Gamma_B}+2\mathbf{\Gamma_R^*})^2\}$. Since $\mathbf{B}+\mathbf{\Gamma_B}$ is positive definite and we know that the first maximum eigenvalue in Equation (16) is greater than 1, we can further bound the eigenvalue by $\kappa \le \lambda_{\max}\{((\mathbf{\Gamma_R^N})^{-1}\mathbf{\Gamma_R^*})^2\}$. Since $\mathbf{\Gamma_R}$ is a diagonal matrix, we obtain $\kappa \le \max_i(l_i'(q_i^*)/l_i'(q_i^N))^2$.

Since costs are monotone, if we can lower bound $\mathbf{q}^N$ with a quantity $\tilde{\mathbf{q}}$, we then have a lower bound on $l_i'(q_i^N)$ and show that $\kappa \le \max_i(l_i'(q_i^*)/l_i'(\tilde{q}_i))^2$. From Proposition 3, we set $\tilde{\mathbf{q}} = \mathbf{q}^*/2$ and obtain the desired result.

With a monomial cost, the marginal cost is given by $l_i'(q_i) = c_i k q_i^{k-1}$. Thus, κ can be bounded by $1 \le \kappa \le \max_i((c_i k q_i^*)/(c_i k \tilde{q}_i))^{2(k-1)} = \max_i(q_i^*/\tilde{q}_i)^{2(k-1)} = 2^{2(k-1)}$, where k is the degree of nonlinearity. With affine cost where the nonlinearity degree $k=1$, the Jacobian similarity factor $\kappa = 2^0 = 1$.

With congestion cost from a $M/M/1$ system, the marginal cost is given by $l_i'(q_i) = c_i/(\mu_i - q_i)^2$. Thus, the Jacobian similarity factor is bounded by $\kappa \le \max_i((\mu_i - \tilde{q}_i)/(\mu_i - q_i^*))^2 = \max_i((2\mu_i - q_i^*)/(2(\mu_i - q_i^*)))^2$. In general, one can find a tighter bound on the Jacobian similarity factor κ by providing a tighter lower bound on $\mathbf{q}^N$.

F. Proof for the Lower Bound in Theorem 2

The analysis with spillover costs depends on the matrix $\mathbf{\Xi} = \mathbf{\Gamma_B^{-0.5}}\mathbf{R_{off}}\mathbf{\Gamma_B^{-0.5}}$. It is a symmetric matrix and is similar to $\mathbf{\Gamma_B^{-1}}\mathbf{R_{off}}$, which shares the same set of eigenvalues. Note that matrix $\mathbf{\Xi}$ is not positive definite because it has all zeros on its diagonal. Since $\mathbf{\Xi}$ is a nonnegative matrix, by the Perron–Frobenius theorem, we have $|\lambda\{\mathbf{\Xi}\}| \le \lambda_{\max}\{\mathbf{\Xi}\}$, where $\lambda_{\max}\{\mathbf{\Xi}\}$ is real and nonnegative. In other words, the maximum eigenvalue of $\mathbf{\Xi}$ indicates its spectral radius. By the Gershgorin disc theorem, this eigenvalue can be bounded as follows:

$$\lambda_{\max}\{\mathbf{\Xi}\} = \lambda_{\max}\{\mathbf{\Gamma_B^{-1}}\mathbf{R_{off}}\} \le \max_i \frac{\sum_{j\ne i} r_{ji}}{\beta_{ii}} = \max_i \rho_i = \bar{\rho}. \qquad (17)$$

That is, the maximum cost-to-benefit ratio, $\bar{\rho}$, upper bounds the spectral radius of matrix $\mathbf{\Xi}$—in particular, when $\bar{\rho} \le 1$ (i.e., $\lambda_{\max}\{\mathbf{\Xi}\} \le 1$). It implies that the matrix $\mathbf{I}-\mathbf{\Xi}$ is positive semidefinite, where $\mathbf{I}$ is the identity matrix. However, one cannot make a similar claim on $\mathbf{\Xi}-\mathbf{I}$ when $\bar{\rho} \ge 1$ because $\mathbf{\Xi}$ can have negative eigenvalues.

F.1. Lower Bound for $\bar{\rho} \le 1$

The bulk of the analysis to derive the constant lower bound is very similar to the proof of Theorem 1. In fact, the first two steps are nearly identical as we express the Jacobian matrix $\mathbf{R}$ to be its diagonal and off-diagonal components $\mathbf{\Gamma_R}$ and $\mathbf{R_{off}}$ and follow through the steps shown in §3.2. We do need to modify the last step. Since $\mathbf{I}-\mathbf{\Xi}$ is positive definite, the expression for the composite matrix could be simplified to $(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})-2(\mathbf{I}-\mathbf{\Xi}) = \mathbf{G}-\mathbf{I}+\mathbf{\Xi}$. Because this matrix is nonnegative, it is copositive, and we have obtained the desired result.

F.2. Lower Bound for $\bar{\rho} \ge 1$

Let $\mathbf{\Sigma} = \mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N}+\mathbf{R}^N$, and let $\mathbf{\Phi} = \mathbf{B}+\mathbf{R}^*+\mathbf{R}^N$. From Lemma 4, we get $\mathbf{\Sigma}\mathbf{q}^N \ge \mathbf{\Phi}\mathbf{q}^*$. Note that the vectors on both sides of the inequality are nonnegative. By Proposition 2, we obtain the following:

$$\begin{aligned}W(\mathbf{q}^N) &= (\mathbf{q}^N)^\top(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N})\mathbf{q}^N\\ &= (\mathbf{q}^N)^\top\mathbf{\Sigma}\mathbf{\Sigma}^{-1}(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N})\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{q}^N.\end{aligned}$$

Replacing $\mathbf{\Sigma}\mathbf{q}^N$ with $\mathbf{\Phi}\mathbf{q}^*$, we obtain a lower bound on $W(\mathbf{q}^N)$; i.e.,

$$\begin{aligned}W(\mathbf{q}^N) &\ge (\mathbf{q}^*)^\top\mathbf{\Phi}\mathbf{\Sigma}^{-1}(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N})\mathbf{\Sigma}^{-1}\mathbf{\Phi}\mathbf{q}^*\\ &= (\mathbf{q}^*)^\top(\mathbf{B}+\mathbf{R}^*+\mathbf{R}^N)(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N}+\mathbf{R}^N)^{-1}\\ &\quad\cdot(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N})(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N}+\mathbf{R}^N)^{-1}\\ &\quad\cdot(\mathbf{B}+\mathbf{R}^*+\mathbf{R}^N)\mathbf{q}^*.\end{aligned}$$

By making use of the Jacobian similarity properties on matrices $\mathbf{\Gamma_R^N}$ and $\mathbf{R}^N$, there exists $\kappa' \ge 1$ such that

$$\begin{aligned}W(\mathbf{q}^N) &\ge \frac{1}{\kappa'}(\mathbf{q}^*)^\top(\mathbf{B}+\mathbf{R}^*+\mathbf{R}^*)(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*}+\mathbf{R}^*)^{-1}\\ &\quad\cdot(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*})(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*}+\mathbf{R}^*)^{-1}(\mathbf{B}+\mathbf{R}^*+\mathbf{R}^*)\mathbf{q}^*\\ &= \frac{1}{\kappa'}(\mathbf{q}^*)^\top(\mathbf{B}+2\mathbf{R}^*)(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*}+\mathbf{R}^*)^{-1}\\ &\quad\cdot(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*})(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*}+\mathbf{R}^*)^{-1}(\mathbf{B}+2\mathbf{R}^*)\mathbf{q}^*.\end{aligned}$$

Note that by using the definition of the minimum eigenvalue of a positive semidefinite matrix, one way to bound $1/\kappa'$ is shown as follows:

$$\begin{aligned}\frac{1}{\kappa'} &\ge \lambda_{\min}\{(\mathbf{B}+\mathbf{R}^*+\mathbf{R}^N)^2(\mathbf{B}+2\mathbf{R}^*)^{-2}(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N}+\mathbf{R}^N)^{-2}\\ &\qquad\cdot(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*}+\mathbf{R}^*)^2(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N})(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*})^{-1}\},\end{aligned}$$

or equivalently,

$$\begin{aligned}\kappa' &\le \lambda_{\max}\{(\mathbf{B}+\mathbf{R}^*+\mathbf{R}^N)^{-2}(\mathbf{B}+2\mathbf{R}^*)^2(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N}+\mathbf{R}^N)^2\\ &\qquad\cdot(\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*}+\mathbf{R}^*)^{-2}(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^N})^{-1}\\ &\qquad\cdot(\tfrac{1}{2}\mathbf{B}+\mathbf{\Gamma_B}+\mathbf{\Gamma_R^*})\}. \qquad (18)\end{aligned}$$

Combining the result with $W(\mathbf{q}^*)$, we can see that it is all in the optimal $\mathbf{q}^*$ space. Therefore, we will drop the superscript in this proof. Letting $\tilde{\mathbf{B}} = \mathbf{B}+2\mathbf{\Gamma_R}$,

$$\begin{aligned}\frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} &\ge \frac{1}{\kappa'}((\mathbf{q}^*)^\top(\tilde{\mathbf{B}}+2\mathbf{R_{off}})(\tilde{\mathbf{B}}+\mathbf{\Gamma_B}+\mathbf{R_{off}})^{-1}(\tilde{\mathbf{B}}+2\mathbf{\Gamma_B})\\ &\qquad\cdot(\tilde{\mathbf{B}}+\mathbf{\Gamma_B}+\mathbf{R_{off}})^{-1}(\tilde{\mathbf{B}}+2\mathbf{R_{off}})\mathbf{q}^*)\\ &\qquad\cdot((\mathbf{q}^*)^\top(\tilde{\mathbf{B}}+2\mathbf{R_{off}})\mathbf{q}^*)^{-1}.\end{aligned}$$

Let $\mathbf{G} = \mathbf{\Gamma}_\mathbf{B}^{-0.5}\tilde{\mathbf{B}}\mathbf{\Gamma}_\mathbf{B}^{-0.5}$, and let $\mathbf{\Xi} = \mathbf{\Gamma}_\mathbf{B}^{-0.5}\mathbf{R}_{\text{off}}\mathbf{\Gamma}_\mathbf{B}^{-0.5}$; the expression becomes

$$\frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} \geq \frac{1}{\kappa'}((\mathbf{q}^*)^\top\mathbf{G}^{0.5}(\mathbf{G}+2\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1}(\mathbf{G}+2\mathbf{I}) \cdot (\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1}(\mathbf{G}+2\mathbf{\Xi})\mathbf{G}^{0.5}\mathbf{q}^*) \cdot ((\mathbf{q}^*)^\top\mathbf{G}^{0.5}(\mathbf{G}+2\mathbf{\Xi})\mathbf{G}^{0.5}\mathbf{q}^*)^{-1}.$$

Using the Rayleigh–Ritz theorem, a lower bound is given by the minimum eigenvalue of the following composite matrix:

$$\frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} \geq \frac{1}{\kappa'}\lambda_{\min}\{(\mathbf{G}+2\mathbf{I})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1}(\mathbf{G}+2\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1}\} = \frac{1}{\kappa'}(1-\lambda_{\max}\{((\mathbf{I}-\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1})^2\}).$$

The goal is to find an upper bound on $\lambda_{\max}\{((\mathbf{I}-\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1})^2\}$, which depends on the eigenvalues of $\mathbf{G}$ and $\mathbf{\Xi}$. Since matrix $\mathbf{G}$ is a nonnegative and positive definite matrix,

$$\lambda_{\max}\{((\mathbf{I}-\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1})^2\} \leq \max_{i\in\{1,\ldots,n\}}\left(\frac{1-\lambda_i\{\mathbf{\Xi}\}}{\lambda_{\min}\{\mathbf{G}\}+\lambda_i\{\mathbf{\Xi}\}+1}\right)^2.$$

Consider a function $f(x) = (1-x)^2/(y+x+1)^2$. When $x > 1$, $f(x)$ increases in x. Thus, when $\lambda_{\max}\{\mathbf{\Xi}\} > 1$,

$$\lambda_{\max}\{((\mathbf{I}-\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1})^2\} \leq \left(\frac{\lambda_{\max}\{\mathbf{\Xi}\}-1}{\lambda_{\min}\{\mathbf{G}\}+\lambda_{\max}\{\mathbf{\Xi}\}+1}\right)^2. \quad (19)$$

We have obtained an upper bound on $\lambda_{\max}\{\mathbf{\Xi}\}$ in Equation (17). To determine the minimum eigenvalue of $\mathbf{G}$, we see that $\lambda_{\min}\{\mathbf{G}\} = \lambda_{\min}\{\mathbf{\Gamma}_\mathbf{B}^{-1}\mathbf{B}+2\mathbf{\Gamma}_\mathbf{B}^{-1}\mathbf{\Gamma}_\mathbf{R}\} \geq \lambda_{\min}\{\mathbf{\Gamma}_\mathbf{B}^{-1}\mathbf{B}\}$ since $\mathbf{\Gamma}_\mathbf{B}^{-1}\mathbf{\Gamma}_\mathbf{R}$ is a nonnegative diagonal matrix. By the Gershgorin disc theorem, $\lambda_{\min}\{\mathbf{\Gamma}_\mathbf{B}^{-1}\mathbf{B}\} \geq 1 - \max_i((\sum_{j\neq i}\beta_{ij})/\beta_{ii}) \geq 1 - \max_i((\sum_{j\neq i}\beta_{ij}+2r_{ii})/\beta_{ii}) = 1-\bar{\gamma}$. Meanwhile, since $\mathbf{G}$ is a positive definite matrix, $\lambda_{\min}\{\mathbf{G}\} > 0$. We conclude that $\lambda_{\min}\{\mathbf{G}\} \geq \max(1-\bar{\gamma}, 0)$. Substituting the bounds we have on $\lambda_{\min}\{\mathbf{G}\}$ and $\lambda_{\max}\{\mathbf{\Xi}\}$ into Equation (19), we obtain the desired lower bound for $\bar{\rho} > 1$.

G. Proof for the Upper Bound in Theorem 2

The proof has three main steps and the two steps are identical to what has been shown in the proof for the upper bound in Theorem 1. We only need to modify the last step as follows. In Lemma 3, we have established an upper bound on the comparison of welfare achieved in the two settings in the terms of eigenvalues of a composite matrix; i.e.,

$$\frac{W(\mathbf{q}^N)}{W(\mathbf{q}^*)} \leq \kappa(1-\lambda_{\min}\{((\mathbf{I}-\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1})^2\}).$$

Note that matrix $\mathbf{G}$ is positive definite and the minimum eigenvalue of the composite matrix is bounded below by

$$\lambda_{\min}\{((\mathbf{I}-\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1})^2\} \geq \min_{\lambda\{\mathbf{\Xi}\}}\left(\frac{\lambda\{\mathbf{\Xi}\}-1}{\lambda_{\max}\{\mathbf{G}\}+1+\lambda\{\mathbf{\Xi}\}}\right)^2.$$

Consider a function $g(x,y) = ((x-1)(x+y+1))^2$ with $x \in [\underline{x}, \bar{x}]$. If $\bar{x} \leq 1$, the function decreases in x. Thus, for a fixed y, the minimum of the function is achieved at $\bar{x}$. Therefore, we obtain the following: When $\lambda_{\max}\{\mathbf{\Xi}\} \leq 1$,

$$\lambda_{\min}\{((\mathbf{I}-\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1})^2\} \geq \left(\frac{\lambda_{\max}\{\mathbf{\Xi}\}-1}{\lambda_{\max}\{\mathbf{G}\}+1+\lambda_{\max}\{\mathbf{\Xi}\}}\right)^2.$$

Since $\lambda_{\max}\{\mathbf{\Xi}\} \leq \bar{\rho}$ and $\lambda_{\max}\{\mathbf{G}\} \leq 1+\bar{\gamma}$, we can further lower bound the eigenvalue of the composite matrix by

$$\lambda_{\min}\{((\mathbf{I}-\mathbf{\Xi})(\mathbf{G}+\mathbf{I}+\mathbf{\Xi})^{-1})^2\} \geq \left(\frac{\bar{\rho}-1}{\bar{\gamma}+2+\bar{\rho}}\right)^2.$$

To obtain the tightness result, note that with symmetric service providers, $\lambda_{\max}\{\mathbf{\Xi}\} = \bar{\rho}$, and all the steps become equalities.

H. Proof for Proposition 5

With spillover cost, when $\bar{\rho} \leq 1$, $\mathbf{q}^N \leq \mathbf{q}^*$ and $\mathbf{R}^N \leq \mathbf{R}^*$ are implied. Using the same argument in Lemma 4, an upper bound on κ is given by $\lambda_{\max}\{(\mathbf{\Gamma}_\mathbf{R}^N+\mathbf{R}^N)^{-2}(\mathbf{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^2\}$. When service providers are symmetric with costs $r_{ij} = r$ for all i and j, $\mathbf{R}$ is a rank-1 matrix with two distinct eigenvalues, nr and 0. Then it is straightforward to show that the bound on κ reduces to $(l'(q^*)/l'(q^N))^2$. Thus, one can determine κ by substituting a lower bound on $\mathbf{q}^N$ such as $\mathbf{q}^N \geq 1/2\mathbf{q}^*$.

Another Jacobian similarity factor, κ', is defined in Equation (18). Note that an upper bound on κ' can be written as the product of the maximum eigenvalue of three matrices, $(\mathbf{B}+\mathbf{R}^*+\mathbf{R}^N)^{-2}(\mathbf{B}+2\mathbf{R}^*)^2$, $(\mathbf{B}+\mathbf{\Gamma}_\mathbf{B}+\mathbf{\Gamma}_\mathbf{R}^N+\mathbf{R}^N)^2(\mathbf{B}+\mathbf{\Gamma}_\mathbf{B}+\mathbf{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^{-2}$, and $(\frac{1}{2}\mathbf{B}+\mathbf{\Gamma}_\mathbf{B}+\mathbf{\Gamma}_\mathbf{R}^N)^{-1}(\frac{1}{2}\mathbf{B}+\mathbf{\Gamma}_\mathbf{B}+\mathbf{\Gamma}_\mathbf{R}^*)$. Clearly, when $\mathbf{R}^* = \mathbf{R}^N$, they are all reduced to an identity matrix with eigenvalues equal to 1. When $\bar{\rho} > 1$, the unregulated setting is overproducing; thus, $\mathbf{R}^N > \mathbf{R}^*$. Therefore, the maximum eigenvalue of the first and the third matrices is less than 1, and we can bound κ' by $\lambda_{\max}\{(\mathbf{B}+\mathbf{\Gamma}_\mathbf{B}+\mathbf{\Gamma}_\mathbf{R}^N+\mathbf{R}^N)^2(\mathbf{B}+\mathbf{\Gamma}_\mathbf{B}+\mathbf{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^{-2}\} \leq \lambda_{\max}\{(\mathbf{\Gamma}_\mathbf{R}^N+\mathbf{R}^N)^2(\mathbf{\Gamma}_\mathbf{R}^*+\mathbf{R}^*)^{-2}\}$. Using the same argument as above, when service providers are symmetric, the bound on κ' simplifies to $(l'(q^N)/l'(q^*))^2$. It can be further bounded by replacing q^N with an upper bound $\hat{q}$. One such example is to compute the output level when costs are zero, i.e., $\hat{\mathbf{q}} = (\mathbf{B}+\mathbf{\Gamma}_\mathbf{B})^{-1}\bar{\mathbf{p}}$, which is equivalent to $\hat{q} = \bar{p}/(2\beta_i+(n-1)\beta_j)$ for the symmetric case.

References

Acemoglu D, Ozdaglar A (2007) Competition and efficiency in congested markets. *Math. Oper. Res.* 32(1):1–31.

Allon G, Federgruen A (2007) Competition in service industries. *Oper. Res.* 55(1):37–55.

Allon G, Federgruen A (2008) Service competition with general queueing facilities. *Oper. Res.* 56(4):827–849.

Beckmann MJ, McGuire CB, Winsten CB (1956) *Studies in the Economics of Transportation* (Yale University Press, New Haven, CT).

Brander JA, Zhang A (1990) Market conduct in the airline industry: An empirical investigation. *RAND J. Econom.* 21(4):567–583.

Brander JA, Zhang A (1993) Dynamic oligopoly behaviour in the airline industry. *Internat. J. Indust. Organ.* 11(3):407–435.

Brueckner JK (2002) Airport congestion when carriers have market power. *Amer. Econom. Rev.* 92(5):1357–1375.

Brueckner JK (2005) Internalization of airport congestion: A network analysis. *Internat. J. Indust. Organ.* 23(7–8):599–614.

Bulow JI, Geanakoplos JD, Klemperer PD (1985) Multimarket oligopoly: Strategic substitutes and complements. *J. Political Econom.* 93(3):488–511.

Bureau of Transportation Statistics (2004) On-time performance—Flight delays at a glance. Accessed July 15, 2014, http://www.transtats.bts.gov/HOMEDRILLCHART.ASP.

Cachon GP, Harker PT (2002) Competition and outsourcing with scale economies. *Management Sci.* 48(10):1314–1333.

Correa JR, Schulz AS, Stier-Moses NE (2004) Selfish routing in capacitated networks. *Math. Oper. Res.* 29(4):961–976.

Correa JR, Schulz AS, Stier-Moses NE (2007) Fast, fair, and efficient flows in networks. *Oper. Res.* 55(2):215–225.

Dafermos S (1982) The general multimodal network equilibrium problem with elastic demand. *Networks* 12(1):57–72.

Daniel JI (1995) Congestion pricing and capacity of large hub airports: A bottleneck model with stochastic queues. *Econometrica* 63(2):327–370.

Daniel JI (2001) Distributional consequences of airport congestion pricing. *J. Urban Econom.* 50(2):230–258.

Debreu G (1952) A social equilibrium existence theorem. *Proc. Natl. Acad. Sci. USA* 38(10):886–893.

Farahat A, Perakis G (2009) Profit loss in differentiated oligopolies. *Oper. Res. Lett.* 37(1):43–46.

Farahat A, Perakis G (2010) A nonnegative extension of the affine demand function and equilibrium analysis for multiproduct price competition. *Oper. Res. Lett.* 38(4):280–286.

Farahat A, Perakis G (2011) Comparison of Bertrand and Cournot profits in oligopolies with differentiated products. *Oper. Res.* 59(2):507–513.

Faulhaber GR, Hogendorn C (2000) The market structure of broadband telecommunications. *J. Indust. Econom.* 48(3):305–329.

Federal Communications Commission (2010) Mobile broadband: The benefits of additional spectrum. Report, FCC, Washington, DC. http://www.fcc.gov/document/mobile-broad-band-benefits-additional-spectrum.

Hamdouch Y, Florian M, Hearn DW, Lawphongpanich S (2007) Congestion pricing for multi-modal transportation systems. *Transportation Res. Part B* 41(3):275–291.

Hayrapetyan A, Tardos É, Wexler T (2007) A network pricing game for selfish traffic. *Distributed Comput.* 19(4):255–266.

Horn RA, Johnson CR (1985) *Matrix Analysis* (Cambridge University Press, Cambridge, UK).

Johari R, Weintraub GY, Van Roy B (2010) Investment and market structure in industries with congestion. *Oper. Res.* 58(5):1303–1317.

Joint Economic Committee (2008) Your flight has been delayed again: Flight delays cost passengers, airlines, and the U.S. economy billions. Report, Joint Economic Committee Majority Staff, Washington, DC. http://www.jec.senate.gov/public/?a=Files.Serve&File_id=47e8d8a7-661d-4e6b-ae72-0f1831dd1207.

Kluberg J, Perakis G (2012) Generalized quantity competition for multiple products and loss of efficiency. *Oper. Res.* 60(2):335–350.

Koutsoupias E, Papadimitriou C (1999) Worst-case equilibria. *Proc. 16th Sympos. Theoret. Aspects Comput. Sci.* (Springer-Verlag, Berlin), 404–413.

Kreps DM, Scheinkman JA (1983) Quantity precommitment and Bertrand competition yield Cournot outcomes. *Bell J. Econom.* 14(2):326–337.

Maillé P, Stier-Moses NE (2009) Eliciting coordination with rebates. *Transportation Sci.* 43(4):473–492.

Martínez-de-Albéniz V, Simchi-Levi D (2009) Competition in the supply option market. *Oper. Res.* 57(5):1082–1097.

Oum TH, Zhang A, Zhang Y (1993) Inter-firm rivalry and firm-specific price elasticities in deregulated airline markets. *J. Transport Econom. Policy* 27(2):171–192.

Ozdaglar A (2008) Price competition with elastic traffic. *Networks* 52(3):141–155.

Parker PM, Roller L-H (1997) Collusive conduct in duopolies: Multimarket contact and cross-ownership in the mobile telephone industry. *RAND J. Econom.* 28(2):304–322.

Perakis G (2007) The "price of anarchy" under nonlinear and asymmetric costs. *Math. Oper. Res.* 32(3):614–628.

Perakis G, Roels G (2007) The price of anarchy in supply chains: Quantifying the efficiency of price-only contracts. *Management Sci.* 53(8):1249–1268.

Perakis G, Sun W (2012) Price of anarchy for supply chains with partial positive externalities. *Oper. Res. Lett.* 40(2):78–83.

Research and Markets (2013) U.S. mobile backhaul 2013: Broadband wireless realized. Report, Visant Strategies, Kings Park, NY. http://www.researchandmarkets.com/research/nhq8sm/us_mobile.

Rosen JB (1965) Existence and uniqueness of equilibrium points for concave *n*-person games. *Econometrica* 33(3):520–534.

Roughgarden T (2005) *Selfish Routing and the Price of Anarchy* (MIT Press, Cambridge, MA).

Roughgarden T, Tardos E (2002) How bad is selfish routing? *J. ACM* 49(2):236–259.

Statista (2014) Market share of wireless subscriptions held by carriers in the U.S. from 1st quarter 2011 to 1st quarter 2014. Accessed July 15, 2014, http://www.statista.com/statistics/199359/market-share-of-wireless-carriers-in-the-us-by-subscriptions/.

Sun W (2006) Price of anarchy in a Bertrand oligopoly. Master's thesis, Massachusetts Institute of Technology, Cambridge.

Vives X (1999) *Oligopoly Pricing: Old Ideas and New Tools* (MIT Press, Cambridge, MA).

Wardrop JG (1952) Road paper. Some theoretical aspects of road traffic research. *ICE Proc.: Engrg. Divisions* 1(3):325–362.

Weisman E (1990) *Trade in Services and Imperfect Competition: Application to International Aviation* (Kluwer Academic Publishers, Dordrecht, The Netherlands).