The Economics of the Cloud: Price Competition and Congestion

Jonatha Anselmi
Basque Center for Applied Mathematics, jonatha.anselmi@gmail.com

Danilo Ardagna
Dip. di Elettronica e Informazione Politecnico di Milan, ardagna@elet.polimi.it

John C.S. Lui
Computer Science & Engineering, Chinese University of Hong Kong, cslui@cse.cuhk.edu.hk

Adam Wierman
Department of Computing and Mathematical Sciences, California Institute of Technology, adamw@caltech.edu

Yunjian Xu
Engineering Systems and Design, Singapore University of Technology and Design, xuyunjian@gmail.com

Zichao Yang
Computer Science & Engineering, Chinese University of Hong Kong, yangtze2301@gmail.com

This paper proposes a model to study the interaction of price competition and congestion in the cloud computing marketplace. Specifically, we propose a three-tier market model that captures a marketplace with users purchasing services from Software-as-a-Service (SaaS) providers, which in turn purchase computing resources from either Provider-as-a-Service (PaaS) or Infrastructure-as-a-Service (IaaS) providers. Within each level, we define and characterize market equilibria. Further, we use these characterizations to understand the relative profitability of SaaSs and PaaSs/IaaSs, and to understand the impact of price competition on the user experienced performance, i.e., the 'price of anarchy' of the cloud marketplace. Our results highlight that both of these depend fundamentally on the degree to which congestion results from shared or dedicated resources in the cloud.

Key words: cloud computing; pricing; network economics

1. Introduction

The cloud computing marketplace has evolved into a highly complex economic system made up of a variety of services, which are typically classified into three categories:

(i) In *Infrastructure-as-a-Service (IaaS)*, cloud providers rent out the use of (physical or virtual) servers, storage, networks, etc. To deploy applications users must install and maintain operating systems, software, etc. Examples include Amazon EC2, Google Cloud, and Rackspace Cloud.

(ii) In *Platform-as-a-Service (PaaS)*, cloud providers deliver a computing platform on which users can develop, deploy and run their application. Examples include Google App Engine, Amazon Elastic MapReduce, and Microsoft Azure.

(iii) In *Software-as-a-Service (SaaS)*, cloud providers deliver a specific application (service) for users. There are a huge variety of SaaS solutions these days, such as email services, calendars, music services, etc. Examples include services such as Dropbox, Gmail and Google Docs.

Naturally, each type of cloud service (IaaS, PaaS, SaaS) uses different pricing and contracting structures, which yields a complicated economic marketplace. For example, Amazon computing services are billed on an hourly basis, while some other Amazon services (e.g., queue or datastore) are billed according to the data transfer in and out (Amazon (2014a,b)). Google App engine pricing is applied on a per application or user per month basis and more complex billing rules are applied if monthly quotas are exceeded (Google (2014)).

Further, adding to the complexity of the cloud marketplace is the fact that a particular SaaS is likely running on top of either a PaaS or IaaS. Thus, there is a multi-tier economic interaction between the PaaS or IaaS and the SaaS, and then between the SaaS and the user. This multi-tier interaction was illustrated prominently by the recent crashes of IaaS provider Amazon EC2, which in turn brought down dozens of prominent SaaS providers (Cloudtimes (2012), NetworkWorld (2012)).

As a result of the complicated economic marketplace within the cloud, the performance delivered by SaaS providers to consumers depends on both the resource allocation design of the service itself (as traditionally considered) and the strategic incentives resulting from the multi-tiered economic interactions. Importantly, it is impossible to separate these two components in this context. For example, users are both price-sensitive and performance-sensitive when choosing a SaaS; however the bulk of the performance component for a SaaS comes from the back-end Iaas/PaaS. Further, the IaaS/PaaS does not charge the consumer, it charges the SaaS. Additionally, there is competition among SaaS providers for consumers and among IaaS/PaaS providers for SaaS providers, which yields a competitive marketplace that in turn determines the resource allocation of infrastructure to users, and thus the performance experienced by users.

Contributions of this paper

This paper aims to introduce and analyze a stylized model capturing the multi-tiered interaction between users and cloud providers in a manner that exposes the interplay of congestion, pricing, and performance issues.

To accomplish this, we introduce a novel three-tier model for the cloud computing marketplace. This model, illustrated in Figure 1, considers the strategic interaction between users and SaaS providers (the first and second tiers), in addition to the strategic interaction between SaaS providers and either IaaS or PaaS providers (the second and third tiers). Of course, within each tier there is also competition among users, SaaS providers, and IaaS or PaaS providers, respectively. To the best of our knowledge, this is the first paper that jointly considers the interactions and the equilibria arising from the full cloud computing stack (i.e., users, services, and infrastructures/platforms) – previous work has focused only on pairwise interactions, e.g., Acemoglu and Ozdaglar (2007), Anselmi et al. (2011), Ardagna et al. (2012).

The details of the model are provided in Section 2 but, briefly, the key features are: (i) users strategically determine which SaaS provider to use depending on a combination of performance and price; (ii) SaaS providers compete by strategically determining their price and the IaaS/PaaS provider they use in order to maximize profit, which depends on the number of users they attract; (iii) IaaS/PaaS providers compete by strategically determining their price to maximize their profit; (iv) the performance experienced by the users is affected by the congestion of the resources procured at the IaaS/PaaS chosen by the SaaS, and that this congestion is a result of the combination of congestion at *dedicated resources*, where congestion depends only on traffic from the SaaS, and *shared resources*, where congestion depends on the total traffic to the IaaS/PaaS.

The complex nature of the cloud marketplace means that the model introduced in this paper is necessarily complicated too. To highlight this, note that an analytic study of the model entails characterizing equilibria within each of the three tiers, in a context where decisions within one tier impact profits (and thus equilibria) at every other tier.

Due to the complexity of the model, we need to consider a limiting regime in order to be able to provide analytic results. Motivated by the huge, and growing, number of SaaS providers and the (comparatively) smaller number of IaaS/PaaS providers, the limiting regime we consider is one where the number of users and the number of SaaS providers are both large (see Section 4 for a formal statement). In this setting, we can attain an analytic characterization of the interacting markets which yield interesting qualitative insights.

More specifically, with our analysis we seek to provide insights into the following fundamental questions:

(i) How profitable are SaaS providers as compared to PaaS/IaaS providers? Does either have market power?

(ii) How good is user performance? Is the economic structure such that increased competition among cloud providers yields efficient resource allocation?

(iii) How does the degree to which cloud resources are shared/dedicated impact the answers to (i) and (ii)?

Our analysis highlights a number of important, novel qualitative insights with respect to these questions, and we discuss these in detail in Sections 5 and 6. For example, our results highlight that SaaSs extract profits only as a result of dedicated latency; while IaaS/PaaS providers extract profits from both shared and dedicated latencies. However, the profit of IaaS/PaaS providers reduces significantly as competition grows, and converges to zero in the limit, while services remain profitable even when there are a continuum of services. This highlights that SaaS providers maintain market power over IaaS/PaaS providers even when services are highly competitive, and that one should not expect the cloud marketplace to support a large number of IaaS/PaaS providers. This observation is similar to the relationship of content providers to ISPs in the internet (Musacchio et al. (2009), Economides and Tåg (2012)). However, because IaaS/PaaS providers can extract profits from both shared and dedicated latencies they remain reasonably profitable relative to services as long as competition is not extreme. This highlights that the cloud market structure seems not to be as susceptible as the internet to a lack of incentives for infrastructure investment. But, our analysis highlights an issue with the current market structure: the interaction of SaaS providers and IaaS/PaaS providers serves to protect inefficient IaaS/PaaSs. That is, even if one IaaS/PaaS provider is extremely inefficient compared to another, the inefficient provider still obtains significant profit. Given the suggestion from the results discussed above that the profitability of IaaS/PaaS providers will limit the market to a small level of competition, this "protection" of inefficient providers is a dangerous phenomenon.

Another danger that our analysis highlights is that the market structure studied here can yield significant performance loss for users, as compared with optimal resource allocation. Specifically, the price competition among services and providers yields inefficient resource allocation, i.e., the price of anarchy can be arbitrarily large. However, competition among PaaS/IaaS providers can result in significant improvements in user efficiency. In particular, as the number of providers (and thus competition) grows, in the limit we show that the price of anarchy cannot be higher than 2, when congestion costs are linear, and $k+1$ if congestion costs are polynomial with degree k. Since the price of anarchy of the two-tier model (users and SaaSs) converges to one in the limit as the number of services grows (Anselmi et al. (2011)), our result reveals that the addition of providers into the marketplace "undoes" the efficiency created by competition among services. Further, these results highlight that it is crucial to find ways to incentivize participation of IaaS/PaaS providers in the cloud marketplace, especially given the above observation that profitability of providers decreases quickly with increasing competition.

Relationship to prior work

There is a large literature that focuses on strategic behavior and pricing in cloud systems and, more generally, in the internet. This area of 'network economics' or 'network games' is full of increasingly rich models incorporating game theoretic tools into more traditional network models. For surveys providing an overview of the modeling and equilibrium concepts in typically used networking games, and additionally an overview of their applications in telecommunications and wireless networks, see van den Nouweland et al. (1996), Haviv (2001), Altman et al. (2006).

In the context of cloud systems specifically, an increasing variety of network games have been investigated and three main areas of attention in this literature are resource allocation (Teng and Magoules (2010), Hong et al. (2011)), load balancing (Altman et al. (2008), Chen et al. (2009), Anselmi et al. (2011), Anselmi and Gaujal (2011)), and pricing (Yolken and Bambos (2008), Ardagna et al. (2012), Acemoglu and Ozdaglar (2007), Feng et al. (2013)). It is this last line of work that is most related to the current paper. Within this pricing literature, the most related papers to our work are Acemoglu and Ozdaglar (2007), Yolken and Bambos (2008), Anselmi et al. (2011), Ardagna et al. (2012), Song et al. (2012), Feng et al. (2013); see also the references therein.

Each of these papers focuses on deriving the existence and efficiency (as measured by the price of anarchy) of pricing mechanisms in the cloud. For example, Ardagna et al. (2012) considers a two-tier model capturing the interaction between SaaSs and a single IaaS, and studies the existence and efficiency of equilibria allocations. Similarly, Acemoglu and Ozdaglar (2007), Anselmi et al. (2011), Feng et al. (2013) consider two-tier models capturing the interaction between users and SaaSs or between SaaSs and PaaSs/IaaSs, and study the existence and efficiency of equilibrium allocations.

Thus, the questions asked in these (and other) papers are similar to those in our work. However, the model considered in this paper is the first to capture the three-tier competing dynamics between users, SaaSs, and IaaSs/PaaSs simultaneously. Further, we model the distinction between congestion from shared and dedicated resources. Neither of these factors was studied in the previous work; and both lead to novel qualitative insights about the cloud marketplace (while simultaneously presenting significant technical challenges to overcome).

2. Modeling framework

We construct a model for studying the interactions among three parities in the cloud marketplace: users, service providers (services for short) and infrastructure providers (providers for short). In this section we define the three types of players in our model, but we discuss their strategic behavior only informally. A formal description of the strategic aspects of the model is deferred to Sections 3 and 4. Note that Figure 1 is helpful in understanding the structure of our modeling framework.

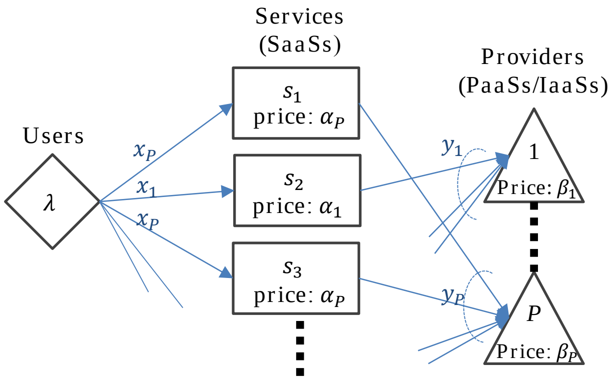

Figure 1 Overview of model structure and notation.

Providers. We consider P providers who sell resources to services, as done by Amazon EC2 and Google Cloud. The resources sold can represent virtual machines, in the case of an IaaS, or platforms provided for development, in the case of a PaaS. Each provider p charges a price β_p per unit of data flow for services that use its infrastructure. This charge-per-flow model is very common, e.g., it is used by Google App Engine. We let y_p denote the total flow of provider p and model the profit of provider p by

$$\text{Provider-Profit}(p) = \beta_p y_p, \tag{1}$$

where, due to the economics of scale in cloud computing, we have ignored the small marginal cost of supporting data flow.

Services. We consider $S \geq 2$ services interacting both with users and providers: they pay infrastructure providers for infrastructure and charge users for usage. Each service pays the provider that it has chosen to join for infrastructure and charge users for usage. We assume that each service s chooses only one (infrastructure) provider, denoted by f_s. So, $f : \{1, \dots, S\} \to \{1, \dots, P\}$ is the service to provider mapping. Further, each service s charges a unit price α_s to users. Let x_s denote the flow (users per time unit) of service s, which implies $y_p = \sum_{s:f_s=p} x_s$. Then, the profit of service s is

$$\text{Service-Profit}(s) = (\alpha_s - \beta_{f_s}) x_s. \tag{2}$$

Users. The customer base of cloud services is typically quite large, and therefore we assume a continuum of users having mass λ, as it is done in nonatomic congestion games. We model the total user flow to the services as inelastic. Therefore, $\lambda = \sum_s x_s$ is constant, where x is the flow vector of services.

When joining a service s, users pay α_s to s, as stated above, and incur a congestion cost. In the cloud, congestion is determined by the combination of both the amount of flow at the service chosen,

x_{f_s}, and the amount of flow using the provider chosen by the service y_{f_s}. Thus, we further break down the latency experienced into two types of congestion costs: 1) the *dedicated cost (latency)* from the service $\tilde{\ell}_{f_s}(x_{f_s})$ and 2) the *shared cost (latency)* from the provider $\hat{\ell}_{f_s}(y_{f_s})$. The dedicated cost represents congestion cost incurred at the service provider, e.g., due to the limited number of virtual machines held by the service. The shared cost represents the congestion at the infrastructure provider, e.g., the delay resulting from the network capacity shared by all services using the same infrastructure provider. We assume that $\tilde{\ell}_{f_s}(.)$ and $\hat{\ell}_{f_s}(.)$ are continuously-differentiable, strictly increasing and convex with $\tilde{\ell}_{f_s}(0) = 0$ and $\hat{\ell}_{f_s}(0) = 0$. Combining these latencies with the service price yields the "effective cost" that users seek to minimize. In particular, the effective cost of a user who chooses service s is

$$\text{User-Effective-Cost(s)} = \alpha_s + \tilde{\ell}_{f_s}(x_{f_s}) + \hat{\ell}_{f_s}(y_{f_s}). \tag{3}$$

In this paper, we sometimes focus on linear latency functions, i.e., latencies of the form

$$\tilde{\ell}_p(x) = \tilde{a}_p x, \qquad \hat{\ell}_p(y) = \hat{a}_p y, \qquad \forall p, \tag{4}$$

where the slopes $\{\tilde{a}_p\}_{p=1}^P$ and $\{\hat{a}_p\}_{p=1}^P$ are assumed to be positive.

2.1. Strategic interactions and time scale separation

Throughout this paper we interpret the three characters described above as players of a game. Informally, in this game each provider sets the price that maximizes its individual profit, each service sets the price and chooses the provider that maximizes its individual profit, and each user chooses to join the service that minimizes its individual effective cost.

Of course, in practice these strategic decisions are taken at different time scales. Because of this, it is reasonable to assume that *players acting at a slow time scale will see only the equilibrium behavior of players operating at a faster time scale.* To this end, we assume that the users act at the fastest time scale, responding to fixed prices of the services and a fixed mapping of the services to the providers. The next fastest time scale is pricing, with services responding optimally to the prices first set by providers. Finally, how services decide to distribute among the providers is modeled as the slowest time scale.

This ordering will be used in the next sections to define strategic equilibria and is motivated by the behavior observed in practice: users move quickly between cloud services depending on price, service and provider prices also change quickly (hourly or faster), while the migration of services across providers happens infrequently.

3. A model with atomic services

In this section we take a first step toward describing a tractable and reasonable model for the equilibria that result from strategic interactions among users, services and providers. We consider non-atomic users, but atomic services and providers. In this context, we define the equilibria concepts of interest and highlight the analytic difficulties of equilibria characterization. These difficulties motivate the consideration of a model with non-atomic services, which we then consider in the remainder of the paper.

3.1. User (Wardrop) equilibria

Given a fixed service to providers mapping f and fixed service and provider prices, we assume that users distribute to minimize their individual effective cost, defined by (3). Similarly to non-atomic congestion games, e.g., Roughgarden and Tardos (2002, 2004), this yields a Wardrop equilibrium, which states that all active services have the same and minimum effective cost. This is defined as follows.

Definition 1. *Let mapping f and service prices α be fixed. A vector $x^{UE}=x^{UE}(\alpha,f)\in[0,\lambda]^S$ is a* user equilibrium *if there exists some $\mu^{UE}\geq 0$ such that*

$$\begin{aligned}
&\tilde{\ell}_{f_s}(x_s^{UE})+\hat{\ell}_{f_s}(y_{f_s}^{UE})+\alpha_{f_s}=\mu^{UE}, && \forall s: x_s^{UE}>0,\\
&\tilde{\ell}_{f_s}(x_s^{UE})+\hat{\ell}_{f_s}(y_{f_s}^{UE})+\alpha_{f_s}\geq\mu^{UE}, && \forall s: x_s^{UE}=0,\\
&\sum\nolimits_{s:f_s=p} x_s^{UE}=y_p^{UE}, && \forall p,\\
&\sum\nolimits_{s} x_s^{UE}=\lambda.
\end{aligned}$$

The existence and uniqueness of a user equilibrium can be easily proven using that conditions in Definition 1 coincide with the optimality conditions of a strictly-convex optimization problem, as done in Dafermos and Sparrow (1969). This is summarized in the following proposition.

Proposition 1. *Let mapping f, service prices α and provider prices $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_P)$ be fixed. There exists a unique user equilibrium, which is given by the unique optimal solution of the following strictly convex optimization problem:*

$$\begin{aligned}
\min_{x\geq 0}\ &\sum_s\left[\int_0^{x_s}\tilde{\ell}_{f_s}(z)\,dz+\alpha_s x_s\right]+\sum_p\int_0^{y_p}\hat{\ell}_p(z)\,dz,\\
\text{s.t.}\ &\sum\nolimits_{s:f_s=p} x_s=y_p, \qquad p=1,\ldots,P,\\
&\sum\nolimits_s x_s=\lambda.
\end{aligned}\tag{5}$$

3.2. Service and provider equilibria

We now build on top of the user level competition described above to consider the price competition of services and providers. In particular, consider a fixed provider mapping f and define equilibrium price vectors for services and providers. This is a natural choice for time-scale separation because prices of cloud providers such as Amazon and Google fluctuate minute-to-minute, but services typically cannot switch between providers at such a fast time scale due to infrastructure setup differences.

In this context, we consider the equilibria of service and provider prices according to a Stackelberg model where providers first set their prices and then services observe these prices and determine the prices they charge to end users. Of course, the capability to act first confers a strategic advantage for providers over the case where all market participants must choose their moves simultaneously; however this ordering is natural given the realities of the cloud marketplace.

Given the above discussion, we can now formally define the service and provider equilibria.

Definition 2. *Let mapping f and provider prices $\boldsymbol{\beta}$ be fixed. A vector $\boldsymbol{\alpha}^{SE} = (\alpha_1^{SE}, \ldots, \alpha_S^{SE})$ is a* service equilibrium *(SE) if*

$$\alpha_s^{SE} \in \arg\max_{\alpha_s \geq 0} (\alpha_s - \beta_{f_s})\, x_s^{UE}(\alpha_s, \alpha_{-s}^{SE}, f), \qquad \forall s. \tag{6}$$

Definition 2 is similar to the definition of oligopolistic equilibrium used in Acemoglu and Ozdaglar (2007), Hayrapetyan et al. (2007). The essential difference stands in the structure of the latency function of each service s, which in our case depends on the vector $(x_{s'})_{s':f_{s'}=f_s}$ through y_{f_s} instead of just x_s. For this reason, existence and uniqueness of a service equilibrium remain difficult issues to study. Furthermore, service equilibria may not exist if the latency functions are highly convex because this model is a variant of the Bertrand-Edgeworth competition model, which does not admit equilibrium for highly-convex cost functions; see, e.g., Acemoglu and Ozdaglar (2007), Maskin and Tirole (1988). The possible non-existence of a service equilibrium makes it difficult to study the interaction among (strategic) providers who set prices to maximize their individual profit, and so it motivates considering a variation of the model for analysis.

As a result, in this paper our approach is to analyze an asymptotic scaling of the model as the number of services becomes large, i.e., increases to infinity. In this case, we obtain a limiting model with non-atomic services where we can establish the existence and uniqueness of a (non-atomic) service equilibrium (cf. Proposition 3), which enables us to formally define the price equilibrium among multiple providers. Note that this limiting non-atomic model (introduced formally in Section 4) can be proven to be the limiting regime of the atomic model introduced in this section. Specifically, we show in Appendix A that a symmetric service equilibrium (at which all services that choose the same provider charge the same price), if exists, must converge to the non-atomic service equilibrium (cf. its definition and characterization in Section 4.3), as the number of services increases to infinity.

4. A model with non-atomic services

The previous section develops the equilibria concepts necessary to characterize the strategic behavior of the users, services, and providers in our model. However, as commented above, there are significant analytic challenges in characterizing these equilibria that make the atomic model intractable to study.

But, a key observation about cloud markets in practice is that there are generally many more service providers than infrastructure providers. Thus, it is natural to consider a situation where there are many more services than providers. This motivates a change to the model, where services become non-atomic rather than atomic, i.e., considering a finite number of providers P but treating services and users as infinitesimals in a non-atomic model. Importantly, the non-atomic model that we define can be interpreted as the limit of the atomic model as the number of services grows proportionally with the mass of end users (see Appendix A).

In this section, we introduce the changes to the equilibria concepts that come when non-atomic services are considered. These changes are driven by properties of the atomic model. Importantly, the model becomes much more analytically tractable. In particular, as we show in Sections 5 and 6, it becomes possible to derive characterizations of the resulting equilibria that provide interesting insights about market power, profitability, and price of anarchy.

The remainder of this section is organized as follows. In Section 4.1, we first develop a model with non-atomic services that approximates the three-tier market model introduced in the previous section. Then, in Sections 4.2-4.4, we define equilibrium concepts that are based on the non-atomic service model introduced above; proceeding by backward induction to study existence and uniqueness in each case. However, these equilibria are clearly entangled. As a result, the definitions and initial analytic characterizations of the equilibria concepts are intermingled in this section so that the characterizations can aid in simplifying the presentation of the definitions that follow. All proofs are deferred to Appendix B for the ease of the reader.

4.1. A model for non-atomic services

We consider a non-atomic model involving a continuum of infinitesimally small (and homogeneous) services, indexed by $s \in [0, 1]$.

As before, let λ denote the total user flow. If the mapping $x_s : [0,1] \to [0,\infty)$ (from the index of a service to its flow) is Lebesgue measurable, then λ can be calculated through the following Lebesgue integral:

$$\lambda = \int_{[0,1]} x_s \mu(ds),$$

where μ is the Lebesgue measure defined on $[0,1]$.

Note that, because the latency cost of users depends only the provider chosen (not the service), all services that choose the same provider are essentially identical. Further, since all services that choose the same provider are faced with the same profit-maximization problem, it is reasonable to assume that they charge the same price to their users[1]. By some abuse of notation, for the rest of the paper we will write the price charged by service s as α_{f_s}, which depends only on the provider it chooses, f_s. Since all users are cost-minimizing, it follows that all services that choose the same provider attract the same amount of data flow, i.e., $x_s = x_{s'}$ if $f_s = f_{s'}$. So, for the rest of the paper, we use x_p to denote the flow of a service that chooses provider p. We can therefore rewrite the profit of a service that chooses provider p as (cf. Eq. (2))

$$\text{Service-Profit}(s) = (\alpha_p - \beta_p)x_p, \qquad \forall s : f_s = p.$$

Let g_p denote the fraction of services that choose provider p, and define a service distribution as a nonnegative P-dimensional vector $\mathbf{g} = (g_1, \ldots, g_P)$ such that $\sum_{p=1}^{P} g_p = 1$. Under the assumption that all services associated with a single provider charge their users the same price, we note that different service to provider mappings f that lead to a single service distribution $\mathbf{g}$ will result in the same service prices and user flow. That is, the service to provider mapping f can be fully "represented" by its corresponding service distribution $\mathbf{g}$, and therefore, we will use the latter in the rest of the paper. We have

$$y_p = g_p x_p, \quad \forall p; \qquad\qquad \lambda = \sum_p g_p x_p.$$

4.2. User (Wardrop) Equilibrium

Under given service prices $\boldsymbol{\alpha}$ and a service distribution $\mathbf{g}$, user equilibrium can be defined in a way analogous to Definition 1.

Definition 3. *Given the prices charged by services $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_P)$ and a service distribution $\mathbf{g} = (g_1, \ldots, g_P)$, a flow vector $\{x_p^{UE}\}_{p=1}^{P}$ is a* ***user equilibrium*** *if there exists some μ^{UE} such that*

$$\tilde{\ell}_p(x_p^{UE}) + \hat{\ell}_p(y_p^{UE}) + \alpha_p = \mu^{UE}, \qquad \forall p : x_p^{UE} > 0,$$

$$\tilde{\ell}_p(x_p^{UE}) + \hat{\ell}_p(y_p^{UE}) + \alpha_p \geq \mu^{UE}, \qquad \forall p : x_p^{UE} = 0,$$

$$g_p x_p^{UE} = y_p^{UE}, \qquad \forall p,$$

$$\sum_p y_p^{UE} = \lambda.$$

Further, we denote the set of user equilibria as $W(\boldsymbol{\alpha}, \mathbf{g})$.

Similarly to Proposition 1, we have the following characterization on a user equilibrium.

Proposition 2. *Given a service price vector $\boldsymbol{\alpha}$ and a service distribution $\mathbf{g}$, there exists a unique user equilibrium, which is the unique optimal solution of the following (strictly) convex optimization problem,*

$$\begin{aligned} \textit{minimize} \quad & \sum_p \left[\int_0^{x_p} \tilde{\ell}_p(z)dz + \alpha_p x_p \right] + \sum_p \int_0^{y_p} \hat{\ell}_p(z)dz \\ \textit{subject to} \quad & g_p x_p = y_p, \qquad \forall p, \\ & \textstyle\sum_p y_p = \lambda, \\ & x_p \geq 0, \qquad \forall p. \end{aligned} \tag{7}$$

4.3. Service and Provider Equilibria

Before moving to the provider equilibria, let us start with the service (price) equilibrium.

Definition 4. *Given a service distribution $\mathbf{g}$ and a provider price vector $\boldsymbol{\beta}$, a service price vector $\boldsymbol{\alpha}^{SE} = (\alpha_1^{SE}, \ldots, \alpha_P^{SE})$ is a* ***service (price) equilibrium****, if*

$$\alpha_p^{SE} \in \arg\max_{\alpha \geq 0} (\alpha - \beta_p) x(\alpha, \boldsymbol{\alpha}^{SE}), \qquad \forall p, \tag{8}$$

where

$$\begin{aligned} & x(\alpha, \boldsymbol{\alpha}^{SE}) = 0, && \text{if } \mu^{SE} - \hat{\ell}_p(y_p^{SE}) < \alpha, \\ & \tilde{\ell}_p(x(\alpha, \boldsymbol{\alpha}^{SE})) + \alpha = \mu^{SE} - \hat{\ell}_p(y_p^{SE}), && \text{otherwise.} \end{aligned} \tag{9}$$

Here, $y_p^{SE} = g_p x_p^{SE}$, where $(x_1^{SE}, \ldots, x_P^{SE})$ is the unique user equilibrium under the price vector $\boldsymbol{\alpha}^{SE}$ and the service distribution $\mathbf{g}$, and μ^{SE} is the user effective cost of an active service at the user equilibrium $(x_1^{SE}, \ldots, x_P^{SE})$ (cf. Definition 3).

Definition 4 is closely related to its atomic counterpart in Definition 2. The major difference between these two definitions is that, for an infinitesimally small service that chooses provider p, the user equilibrium $(x_1^{SE}, \ldots, x_P^{SE})$ and the corresponding effective cost level μ^{SE} depend only on the prices set by other services. It follows that the service is faced with the constraint in (9), and at a service equilibrium, the price α_p^{SE} maximizes its profit provided that the other services set their prices according to the equilibrium. We show in Appendix A that the equilibria defined above correspond to its counterpart among a finite number of (atomic) services (cf. Definition 2), when the number of services increases to infinity.

Further, we show in Proposition 3 that, under a given $\mathbf{g}$ and $\boldsymbol{\beta}$, all service equilibria yield a unique user equilibrium, i.e., result in the same user flow. We can therefore use $\mathbf{x}^{SE}(\mathbf{g},\boldsymbol{\beta})$ to denote the user equilibrium under a service equilibrium $\boldsymbol{\alpha}^{SE}$ induced by $\boldsymbol{\beta}$, and a service distribution $\mathbf{g}$.

We now move to the provider (price) equilibrium.

Definition 5. *Given a service distribution* $\mathbf{g}$*, a provider price vector* $\boldsymbol{\beta}^{PE}=(\beta_1^{PE},\dots,\beta_P^{PE})$ *is a* ***provider (price) equilibrium****, if*

$$\beta_p^{PE}\in\arg\max_{\beta_p\geq 0}\beta_p x_p^{SE}(\beta_p,\boldsymbol{\beta}_{-p}^{PE},\mathbf{g}),\qquad \forall p, \tag{10}$$

where $(x_1^{SE}(\beta_p,\boldsymbol{\beta}_{-p}^{PE},\mathbf{g}),\dots,x_P^{SE}(\beta_p,\boldsymbol{\beta}_{-p}^{PE},\mathbf{g}))$ *is the unique user equilibrium induced by provider price vector* $(\beta_p,\boldsymbol{\beta}_{-p}^{PE})$*, and the service distribution* $\mathbf{g}$*.*[2]

To interpret the above definition note that, given the prices set by other providers $\boldsymbol{\beta}_{-p}^{PE}$, the equilibrium price β_p^{PE} maximizes every provider p's profit at the user equilibrium induced by the price vector $(\beta_p,\boldsymbol{\beta}_{-p}^{PE})$ among all possible prices $\beta_p\geq 0$.

Given the definitions of service and provider equilibria, the first questions to address are those of existence and uniqueness. We address both these issues for service equilibria in the following proposition, and for provider equilibria in Proposition 4.

Proposition 3. *Given a service distribution* $\mathbf{g}$ *and a provider price vector* $\boldsymbol{\beta}$*, there exists a service equilibrium, and all service equilibria result in a unique user equilibrium* $\mathbf{x}^{SE}(\mathbf{g},\boldsymbol{\beta})$*. Further, the equilibrium price of a service who selects a provider* p *with* $x_p^{SE}(\mathbf{g},\boldsymbol{\beta})>0$ *is uniquely determined:*

$$\alpha_p^{SE}-\beta_p=x_p^{SE}(\mathbf{g},\boldsymbol{\beta})\tilde{\ell}_p'(x_p^{SE}(\mathbf{g},\boldsymbol{\beta})),\qquad \forall p: x_p^{SE}(\mathbf{g},\boldsymbol{\beta})>0. \tag{11}$$

In Appendix A, we show that a symmetric equilibrium among atomic services (cf. Definition 2) must converge to the non-atomic service equilibrium characterized in Proposition 3, as the number of services increases to infinity. In the case of linear latency functions, it is also possible to explicitly characterize the provider price vector $\boldsymbol{\beta}$ at equilibrium.

Proposition 4. *Suppose that latency functions are linear as in* (4)*. Given a service distribution* $\mathbf{g}$ *with at least two positive components*[3]*, there exists a unique provider equilibrium* $\boldsymbol{\beta}^{PE}$ *such that*

1. *For every* p*, we have* $x_p^{PE}>0$*, where* $\mathbf{x}^{PE}\triangleq\mathbf{x}^{SE}(\mathbf{g},\boldsymbol{\beta}^{PE})$ *is the unique user equilibrium resulting from* $\mathbf{g}$ *and* $\boldsymbol{\beta}^{PE}$ *(cf. Proposition 3 for its definition).*
2. *The provider price vector* $\boldsymbol{\beta}^{PE}$ *is characterized by*

$$\beta_p^{PE}=(2\tilde{a}_p+\hat{a}_p g_p)x_p^{PE}+\frac{g_p x_p^{PE}}{\sum_{p':p'\neq p}\frac{g_{p'}}{2\tilde{a}_{p'}+\hat{a}_{p'}g_{p'}}},\qquad \forall\, p. \tag{12}$$

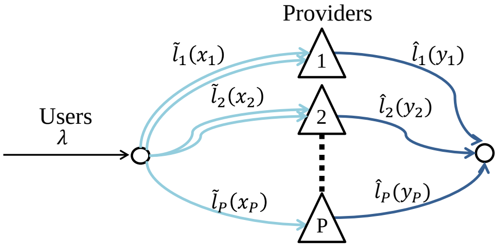

Figure 2 A congestion game that yields the same user flow (at a Wardrop equilibrium) as that resulting from a provider equilibrium of our model.

Before moving to the next section and introducing our main equilibrium concept, we would like to provide an alternative view of the three equilibrium concepts defined so far, which provides more intuition. In particular, Figure 2 shows an oligopolistic congestion game that mimics both user flow and (service and provider) profit resulting from a provider equilibrium defined above. In this congestion game, each user has to go through two serial links to reach the “destination”. An intermediate node represents a provider, and the p-th node attracts g_p fraction of services. The latency of each link is marked in Figure 2, which depends only on the total flow of the link.[4] Once a user chooses its service, its provider p is determined, and the user's cost is given by

$$\tilde{\ell}(x_p) + \hat{\ell}(y_p) + \gamma_p + \beta_p,$$

where $\gamma_p = \alpha_p - \beta_p$ can be regarded as the price charged by the light blue link the user chooses, and β_p is the price set by the dark blue link (the user's provider p). Actually, in both the congestion game presented in Figure 2 and our model, the profit of a service that chooses provider p is $\gamma_p x_p$, and provider p obtains a profit of $\beta_p y_p$.

Since the congestion game depicted in Figure 2 has the same payoff structure as our model (with a fixed service distribution $\mathbf{g}$), the equilibrium concepts defined in Definitions 4 and 5 essentially form a Stackelberg equilibrium of the congestion game where the P dark blue links choose their prices (simultaneously) at the first stage, and then at the second stage, all (non-atomic) light blue links set their prices.

4.4. Distribution Equilibrium

The last component to incorporate into the definition is the mapping of services to providers, i.e., the distribution equilibrium, which fully characterizes the strategic interaction among the three market participants.

Definition 6. *A triple,* $(\mathbf{g}, \boldsymbol{\alpha}, \boldsymbol{\beta})$*, is a* ***distribution equilibrium****, if: (i)* $\boldsymbol{\beta}$ *is a provider equilibrium under the service distribution* $\mathbf{g}$*, and* $\boldsymbol{\alpha}$ *is a service equilibrium under* $\mathbf{g}$ *and* $\boldsymbol{\beta}$*; and (ii) no* *service has an incentive to change its provider because all providers yield services the same profit, i.e.,*

$$\begin{aligned} &x_p^{SE}(\mathbf{g},\boldsymbol{\beta})(\alpha_p-\beta_p)=\xi\geq 0, && \forall p: g_p>0,\\ &x_p^{SE}(\mathbf{g},\boldsymbol{\beta})(\alpha_p-\beta_p)\leq \xi, && \forall p: g_p=0,\end{aligned}$$

where $\mathbf{x}^{SE}(\mathbf{g},\boldsymbol{\beta})$ *is the unique user equilibrium induced by* $\mathbf{g}$ *and* $\boldsymbol{\beta}$*.*[5]

Though compact, the above definition tends to be difficult to work with directly. However, in the case of linear latencies, the following conditions are easier to work with and are necessary and sufficient conditions for $(\mathbf{g},\boldsymbol{\beta})$ to be a distribution equilibrium, with $\mathbf{x}=(x_1,\ldots,x_P)$ being the user equilibrium resulting from $\mathbf{g}$ and $\boldsymbol{\beta}$ (i.e., $\mathbf{x}=\mathbf{x}^{SE}(\mathbf{g},\boldsymbol{\beta})$),

$$\begin{cases} \tilde{a}_{p'}x_{p'}^2=\tilde{a}_p x_p^2, \qquad \text{if } \; g_p g_{p'}>0, & (13)\\ \tilde{a}_p x_p^2\leq \tilde{a}_{p'}x_{p'}^2, \qquad \text{if } \; g_p=0, \;\; g_{p'}>0, & (14)\\ 2\tilde{a}_{p'}x_{p'}+\hat{a}_{p'}g_{p'}x_{p'}+\beta_{p'}=2\tilde{a}_p x_p+\hat{a}_p g_p x_p+\beta_p, \qquad \forall\, p,p', & (15)\\ \beta_p=(2\tilde{a}_p+\hat{a}_p g_p)x_p+\dfrac{g_p x_p}{\sum_{p'\neq p}\frac{g_{p'}}{2\tilde{a}_{p'}+\hat{a}_{p'}g_{p'}}}, & (16)\\ \sum_p g_p x_p=\lambda, & (17)\\ \sum_p g_p=1, & (18)\end{cases}$$

where (13) and (14) follow from the definition of a distribution equilibrium, and the service equilibrium prices characterized in Proposition 3. Eq. (15) states that at the user equilibrium $\mathbf{x}$, all users have the same effective cost; this is true because all providers have positive user flows (cf. Proposition 4). The equality in (16) is the provider equilibrium price[6] characterized in Proposition 4.

We can further massage the conditions above in order to highlight that the distribution equilibrium can be interpreted as a generalized Wardrop equilibrium. In particular, for a triple $(\mathbf{g},\boldsymbol{\beta},\mathbf{x}')$ that satisfies the conditions in (13) to (18), we define an alternative user flow vector $\mathbf{x}$ as follows. For any p' with $g_{p'}>0$, we let $x_{p'}=x'_{p'}$, and for every p with $g_p=0$, we make $x_p>x'_p$ such that

$$\tilde{a}_p x_p^2=\tilde{a}_{p'}x_{p'}^2,$$

where p' is a provider with $g_{p'} > 0$. The modified triple $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x})$ must satisfy the following conditions:

$$\begin{cases} \tilde{a}_{p'} x_{p'}^2 = \tilde{a}_p x_p^2, \qquad \forall p,\, p', & (19) \\ 2\tilde{a}_{p'} x_{p'} + \hat{a}_{p'} g_{p'} x_{p'} + \beta_{p'} = 2\tilde{a}_p x_p + \hat{a}_p g_p x_p + \beta_p, \qquad \text{if } \ g_p g_{p'} > 0, & (20) \\ 2\tilde{a}_p x_p + \hat{a}_p g_p x_p + \beta_p \geq 2\tilde{a}_{p'} x_{p'} + \hat{a}_{p'} g_{p'} x_{p'} + \beta_{p'}, \qquad \text{if } \ g_p = 0, \ \ g_{p'} > 0, & (21) \\ \beta_p = (2\tilde{a}_p + \hat{a}_p g_p) x_p + \dfrac{g_p x_p}{\sum_{p' \neq p} \dfrac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}}}, & (22) \\ \sum_p g_p x_p = \lambda, & (23) \\ \sum_p g_p = 1. & (24) \end{cases}$$

Note that, for any triple $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x})$ that satisfies the preceding conditions (19)-(24), we can construct a triple $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x}')$ that satisfies the conditions in (13) to (18). It follows that a vector $(\mathbf{g}, \boldsymbol{\beta})$ that satisfies the conditions in (19)-(24) must be a distribution equilibrium. Next, we define

$$f_p(g_p, g_{-p}) = \frac{1}{\sqrt{\tilde{a}_p}} \left(2(2\tilde{a}_p + \hat{a}_p g_p) + \frac{g_p}{\sum_{p' \neq p} \dfrac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'} S}} \right). \tag{25}$$

Using the above, for a triple $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x})$ that satisfies the conditions in (19)-(24), substituting x_p and β_p to Eqs. (20) and (21), we obtain

$$\begin{array}{ll} f_p(g_p, g_{-p}) = f_{p'}(g_{p'}, g_{-p'}), & \text{if } \ g_{p'} g_p > 0, \\ f_p(0, g_{-p}) \geq f_{p'}(g_{p'}, g_{-p'}), & \text{if } \ g_p = 0, \ \ g_{p'} > 0, \\ \sum_p g_p = 1. & \end{array} \tag{26}$$

That is, a distribution equilibrium $\mathbf{g}$ must satisfy conditions (26). On the other hand, given a vector $\mathbf{g}$ that satisfies conditions (26), one can solve the triple $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x})$ by using conditions (19)-(24), and then obtain the corresponding distribution equilibrium $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x}')$. We conclude that condition (26) is necessary and sufficient for a vector $\mathbf{g}$ to be a distribution equilibrium.

It is this form that provides intuition for the distribution equilibrium concept. In particular, *we can regard condition* (26) *as a generalized Wardrop equilibrium where the latency function depends on all the components of the model.* Further, it can be verified that $f_p(g_p, g_{-p})$ is a convex function of $\mathbf{g}$, and increases with g_p and decreases with g_{-p}. This highlights that, though distribution equilibria are complicated concepts, there is intuition that serves as a guide for our analysis in the coming sections.

5. Profitability

Given the model introduced in the previous two sections, we are now ready to study the interaction of congestion and pricing in the cloud marketplace. The first question we seek to address is the following: Do the providers or services have market power, i.e., which extracts the most profit? Then, in the next section we study the impact of the cloud marketplace on the user experience.

Studying the relative profitability of services and providers requires contrasting the profits attained by services and providers at a distribution equilibrium. However, it is difficult to calculate closed form expressions that allow such a comparison for the general setting. Hence, we consider two special cases of the model here: the case of P symmetric servers, and the case of two asymmetric servers. For both cases, we assume that all providers have linear latency functions so that we can obtain simple, interpretable expressions for the service and provider profits.

5.1. Symmetric providers, P providers

The first case we consider is a symmetric model where $\tilde{a}_p = \tilde{a}$ and $\hat{a}_p = \hat{a}$, for every p. In this case it is easy to characterize the service and provider profits. Specifically, it follows from conditions (26) that there exists a symmetric distribution equilibrium such that

$$g_p = \frac{1}{P}, \qquad p = 1, \dots, P.$$

Then, from (22) we have

$$\beta_p^{PE} = \frac{P}{P-1}\lambda(2\tilde{a} + \hat{a}\frac{1}{P}).$$

Through a simple calculation (based on conditions in (13) to (18)), we obtain

$$\text{Provider-Profit(p)} = \left(\frac{2\tilde{a} + \hat{a}/P}{P-1}\right)\lambda^2, \qquad \text{Service-Profit(s)} = \tilde{a}\lambda^2.$$

These expressions for the provider and service profits are quite informative. In particular, they highlight that services extract profits only as a result of dedicated latency in this setting; while providers extract profits from both shared and dedicated latencies. However, competition among symmetric providers significantly reduces the profits providers can extract. Interestingly, competition more quickly reduces the profits that can be extracted from shared latencies than from dedicated latencies. However, as $P \to \infty$, provider profit goes to zero. In contrast, despite the fact that a continuum of services is considered, services still extract positive profit from the marketplace. This highlights that services maintain market power over providers even when services are highly competitive, and that one should not expect the cloud marketplace to support a large number of providers.

5.2. Asymmetric providers, 2 providers

The asymmetric case is more difficult to characterize explicitly, and so we are limited to the case of two providers, $P=2$. In this setting, we can prove the following proposition, which is the key to our study.

PROPOSITION 5. *Consider a case where there are two providers ($P=2$) with linear latency functions as in* (4). *There exists a unique distribution equilibrium, which is a solution to the following optimization problem:*

$$\begin{aligned}\textit{minimize}\quad & \frac{1}{\sqrt{\tilde{a}_1}}\left(2(2\tilde{a}_1 g_1+\frac{1}{2}\hat{a}_1 g_1^2)+2\tilde{a}_2(-g_1-\ln(1-g_1))+\frac{1}{2}\hat{a}_2 g_1^2\right)\\ & +\frac{1}{\sqrt{\tilde{a}_2}}\left(2(2\tilde{a}_2 g_2+\frac{1}{2}\hat{a}_2 g_2^2)+2\tilde{a}_1(-g_2-\ln(1-g_2))+\frac{1}{2}\hat{a}_1 g_2^2\right)\\ \textit{subject to}\quad & g_1+g_2=0, g_1\geq 0, g_2\geq 0.\end{aligned}$$

The proof of the preceding proposition is given in Appendix C.1. Using the preceding proposition, we can explicit calculations comparing the profitability of services and providers in two (extreme) examples. In particular, we consider examples where one provider is extremely inefficient with respect to either dedicated or shared latencies. These two examples highlight that the competition in the cloud marketplace "protects" inefficient providers. That is, the inefficient provider in both examples still achieves profits within a factor of four of the efficient provider.

Example: One provider has extremely inefficient shared latency

Consider a setting where $\tilde{a}_1$, $\hat{a}_2$, and $\tilde{a}_2$ are fixed, but the marginal shared cost of provider 1 increases to infinity, i.e., $\hat{a}_1\to\infty$. At a distribution equilibrium, it follows from the proof of Proposition 5 that

$$\frac{1}{\sqrt{\tilde{a}_1}}\left(2(2\tilde{a}_1+\hat{a}_1 g_1)+\frac{2\tilde{a}_2 g_1}{1-g_1}+\hat{a}_2 g_1\right)=\frac{1}{\sqrt{\tilde{a}_2}}\left(2(2\tilde{a}_2+\hat{a}_2 g_2)+\frac{2\tilde{a}_1 g_2}{1-g_2}+\hat{a}_1 g_2\right).$$

As $\hat{a}_1$ approaches infinity, through a simple calculation we obtain

$$g_1\to\frac{\sqrt{\tilde{a}_1}}{\sqrt{\tilde{a}_1}+2\sqrt{\tilde{a}_2}},\qquad g_2\to\frac{2\sqrt{\tilde{a}_2}}{\sqrt{\tilde{a}_1}+2\sqrt{\tilde{a}_2}}.$$

We then have

$$\begin{aligned} & x_1\to\frac{\lambda}{3g_1},\qquad x_2\to\frac{2\lambda}{3g_2},\qquad \text{service-profit}\to\frac{\lambda^2}{9}(\sqrt{\tilde{a}_1}+2\sqrt{\tilde{a}_2})^2,\\ & \text{provider-profit}(1)\sim\frac{\lambda^2}{9}\hat{a}_1,\qquad \text{provider-profit}(2)\sim\frac{4\lambda^2}{9}\hat{a}_1,\end{aligned}$$

where x_p is the equilibrium user flow at provider p.

Note that both providers' profits depend only on provider 1's marginal shared cost $\hat{a}_1$, and that the "bad" provider 1 still obtains one half of the user flow of provider 2 and one fourth of the profit of provider 2 despite providing much worse performance.

Example: One provider has extremely inefficient dedicated latency

Consider a setting where $\hat{a}_1$, $\hat{a}_2$, and $\tilde{a}_2$ are fixed, but the marginal dedicated cost of provider 1 increases to infinity, i.e., $\tilde{a}_1 \to \infty$. At a distribution equilibrium, it follows from the proof of Proposition 5 that

$$\frac{1}{\sqrt{\tilde{a}_1}}\left(2(2\tilde{a}_1+\hat{a}_1 g_1)+\frac{2\tilde{a}_2 g_1}{1-g_1}+\hat{a}_2 g_1\right)=\frac{1}{\sqrt{\tilde{a}_2}}\left(2(2\tilde{a}_2+\hat{a}_2 g_2)+\frac{2\tilde{a}_1 g_2}{1-g_2}+\hat{a}_1 g_2\right).$$

As $\tilde{a}_1$ increases to infinity, through a simple calculation we have

$$x_1 g_1 \to \frac{\lambda}{3}, \qquad x_2 g_2 \to \frac{2\lambda}{3}, \qquad \text{service-profit} \sim \frac{\lambda^2}{9}\tilde{a}_1,$$
$$\text{provider-profit}(1) \sim \frac{2\lambda^2}{9}\tilde{a}_1, \qquad \text{provider-profit}(2) \sim \frac{8\lambda^2}{9}\tilde{a}_1,$$

where x_p is the equilibrium user flow at provider p.

Note that both providers' profit depends only on provider 1's marginal dedicated cost $\tilde{a}_1$, and that, again, the "bad" provider (provider 1) still obtains half of the traffic of provider 2 and one fourth of total profit of provider 2.

6. Price of Anarchy

The second question we study about the cloud marketplace is the following: what is the effect of price competition in the cloud on the performance experienced by users?

To study this question, we measure the "performance experienced by users" by the aggregate user latency resulting from a distribution equilibrium $(\mathbf{g},\boldsymbol{\alpha},\boldsymbol{\beta})$. That is:

$$\ell(\mathbf{x},\mathbf{g}) \triangleq \sum_p g_p x_p(\tilde{\ell}_p(x_p)+\hat{\ell}_p(g_p x_p)), \tag{27}$$

where $\mathbf{x}=(x_1,\dots,x_P)$ is the user equilibrium under $\mathbf{g}$ and $\boldsymbol{\alpha}$.

To provide a baseline for comparison, we contrast the aggregate user latency at a distribution equilibrium with the optimal aggregate user latency. That is, we study the "price of anarchy," which is typically used to measure the loss of social welfare caused by the strategic behavior of market participants. In a similar spirit, we define the **price of anarchy (POA)** of a distribution equilibrium as the ratio of its resulting aggregate user latency to the minimum possible:

$$PoA \triangleq \frac{\ell(\mathbf{x},\mathbf{g})}{\ell(\mathbf{x}^*,\mathbf{g}^*)}, \tag{28}$$

where $(\mathbf{x}^*,\mathbf{g}^*)$ is an optimal solution to the following optimization problem

$$\begin{aligned}
\text{minimize} \quad & \ell(\mathbf{x},\mathbf{g}) \\
\text{subject to} \quad & \textstyle\sum_p g_p x_p = \lambda, \\
& \textstyle\sum_p g_p = 1, \\
& g_p \ge 0, \qquad x_p \ge 0, \qquad \forall p.
\end{aligned} \tag{29}$$

Note that a triple $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x})$ that satisfies conditions (19)-(24) yields the same aggregate latency cost as the corresponding distribution equilibrium $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x}')$ that satisfies conditions (13) to (18), because the p-th component of $\mathbf{x}$ is the same as that of $\mathbf{x}'$ for every p with $g_p > 0$. We therefore can (and will) use conditions (19)-(24) to analyze the efficiency of a distribution equilibrium.

The goal of this section is to bound the price of anarchy of the cloud marketplace; however bounding the price of anarchy in our model under general assumptions is difficult. Thus, throughout this section, we assume that latency functions are polynomial, i.e., $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ and $\hat{\ell}_p(y) = \hat{a}_p y^k$, for every p.

Under these assumptions, we provide two main results. First, in Section 6.1, we consider a general market model with P providers and we show by example that when one of the providers has very bad latency cost, a distribution equilibrium may yield an arbitrarily high price of anarchy. On the other hand, we prove an upper bound on the price of anarchy that depends on the minimum and maximum marginal latency costs among all providers. This result provides an efficiency guarantee for a distribution equilibrium, when all providers are nearly "symmetric".

Second, in Section 6.2 we consider an alternative formulation of the model that allows us to separate the impacts of the number of providers and the asymmetry among them. In particular, we consider a "replica economy" scaling of providers where there are P types of providers and the number of providers of each type scales with n as n increases to infinity.[7] In this context, as n increases to infinity, we show that there exists an ϵ-equilibrium with ϵ decreasing to zero. Further, in the limiting game the price of anarchy is bounded by $k+1$, which highlights that if the asymmetry of providers is "fixed", competition among providers leads to efficient performance for users.

6.1. General Bounds on the Price of Anarchy

As mentioned above, without any assumptions on the latency cost or the symmetry of the providers, the price of anarchy of the cloud marketplace can be quite large, as highlighted by the following examples.

Example: Unbounded price of anarchy

Consider a model with P providers. Provider $2, \ldots, P$ are identical, and each has very large marginal shared cost (i.e., $\hat{a}_2 = \hat{a}_3 = \ldots = \hat{a}_P \to \infty$). It is socially optimal for all users to use provider 1, and the minimum aggregate latency cost is given by

$$\ell(\mathbf{x}^*, \mathbf{g}^*) = \sum_p g_p x_p (\tilde{a}_p x_p + \hat{a}_p g_p x_p) = \tilde{a}_1 \lambda^2 + \hat{a}_1 \lambda^2. \tag{30}$$

At a distribution equilibrium, through a simple calculation we obtain

$$g_1 = \frac{2\sqrt{\tilde{a}_1}}{(2\sqrt{\tilde{a}_1}+\sqrt{\tilde{a}_2})}, \qquad g_p = \frac{\sqrt{\tilde{a}_p}}{(P-1)(2\sqrt{\tilde{a}_1}+\sqrt{\tilde{a}_p})}, \qquad p=2,\dots,P,$$

which yields a user flow of

$$x_1 = \frac{2\lambda}{3g_1}, \qquad x_p = \frac{\lambda}{(P-1)3g_p}, \qquad p=2,\dots,P. \tag{31}$$

It is easy to see that as the marginal shared cost of the $P-1$ providers increases to infinity, this distribution equilibrium yields an arbitrarily high price of anarchy.

Price of anarchy bounds for nearly symmetric providers with polynomial costs

The previous example highlights that the efficiency of the cloud marketplace depends heavily on the difference between the best and worst providers. This observation leads to the following proposition, which shows that a distribution equilibrium cannot be too inefficient if the worst provider is not "very" different from the best one when the latency cost is polynomial.

PROPOSITION 6. *Suppose that latency functions are polynomial, i.e., $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ and $\hat{\ell}_p(y) = \hat{a}_p y^k$, for every p. The price of anarchy of a distribution equilibrium cannot be higher than*

$$\frac{\tilde{a}_{\max} + \hat{a}_{\max}}{\tilde{a}_{\min} + \hat{a}_{\min}/P^k}, \tag{32}$$

where $\tilde{a}_{\min} = \min_p \tilde{a}_p$, $\hat{a}_{\min} = \min_p \hat{a}_p$, $\tilde{a}_{\max} = \max_p \tilde{a}_p$, and $\hat{a}_{\max} = \max_p \hat{a}_p$.

Proposition 6 is proved in Appendix D.1. It highlights that symmetry of providers is crucial for ensuring the efficiency of the cloud marketplace. Further, it highlights that when the number of providers is large, the ratio of dedicated latency costs to shared latency costs, i.e., $\hat{a}/\tilde{a}$, also plays a significant role in the efficiency of the marketplace.

6.2. Bounds on the price of anarchy when the number of providers is large

In this subsection we consider an alternative formulation of the model that allows us to attain more general bounds on the price of anarchy of the cloud marketplace. In particular, we consider a setting with a large number of small (non-atomic) providers. More specifically, when there are a large number of small providers, it is reasonable to expect that providers cannot anticipate the impacts of their prices on user flow, due to, for example, the lack of information or the limit of computational capability. This assumption leads us to a "non-atomic" provider equilibrium concept for this scenario, which we define below. Then, in this new model, we are able to obtain general bounds on the price of anarchy under polynomial latency cost functions.

6.2.1. Non-atomic provider price equilibrium In this section we focus on the case of polynomial latency functions, and so before stating our equilibrium concept it is useful to specialize some results from previous sections about service equilibrium prices. In particular, it follows from Proposition 3 that the service equilibrium prices are

$$\alpha_p^{SE} - \beta_p = x_p \tilde{\ell}_p'(x_p) = k\tilde{a}_p(x_p)^k,$$

which yields users of provider p an effective cost of

$$(x_p)^k((k+1)\tilde{\alpha}_p + \hat{\alpha}_p g_p^k) + \beta_p.$$

Using the above, we can define the non-atomic provider price equilibrium as follows.

Definition 7. *Given a service distribution* $\mathbf{g}$*, a provider price vector* $\boldsymbol{\beta}^{PE} = (\beta_1^{PE}, \ldots, \beta_P^{PE})$ *is a* ***non-atomic provider (price) equilibrium****, if*

$$\beta_p^{PE} \in \arg\max_{\beta_p \geq 0} \beta_p x(\beta, \boldsymbol{\beta}^{PE}), \qquad \forall p, \tag{33}$$

where

$$\begin{aligned} & x(\beta_p, \boldsymbol{\beta}^{PE}) = 0, && \text{if } \mu < \beta_p, \\ & x(\beta_p, \boldsymbol{\beta}^{PE})^k((k+1)\tilde{\alpha}_p + \hat{\alpha}_p g_p^k) = \mu - \beta_p, && \text{otherwise.} \end{aligned} \tag{34}$$

Here, μ *is the user effective cost of an active service at the unique user equilibrium induced by* $\mathbf{g}$ *and* $\boldsymbol{\beta}^{PE}$ *(cf. Definition 3).*

In the above definition we assumed that every provider p considers itself as infinitesimally small, and does not take into account its influence on the user effective cost μ. Given the (non-atomic) provider equilibrium defined above, we can define a corresponding distribution equilibrium that parallels Definition 6.

Note that the non-atomic provider equilibrium defined above can be rigorously interpreted as the limit of the original atomic provider game considered to this point of the paper. In particular, we justify the non-atomic provider equilibrium concept by considering a replica economy, where there are in total P types of providers, and the number of providers of each type scales with n as $n \to \infty$. In this context, as n increases to infinity, we show that every provider's profit is approximately maximized at a non-atomic provider equilibrium; that is, a non-atomic provider equilibrium is an ϵ^n-equilibrium with ϵ^n decreasing to zero (as $n \to \infty$). More formally, the sequence of finite models defined as follows converges to the non-atomic provider equilibrium we study in this section.

Definition 8. *Consider a sequence of models* $\mathcal{G}_n$*,* $n = 1, 2, \ldots$*. For each* $\mathcal{G}_n$*:*

(i) The aggregate user flow is $n\lambda$*, and there is a continuum of services in* $[0, n]$*.*

(ii) There are a total of P types of providers. The latency functions of a type p provider are assumed to be linear, i.e., $\tilde{\ell}_p(x) = \tilde{a}_p x$ and $\hat{\ell}_p(y) = \hat{a}_p y$.

(iii) For every p, the number of type-p providers is $q_p^n n$, where $\lim_{n\to\infty} q_p^n = q_p$. We assume that $q_p \geq q_{\min} > 0$, for every p.

Proposition 7. *In a sequence of games $\{\mathcal{G}_n\}_{n=1}^{\infty}$, every provider's profit is approximately maximized at a distribution equilibrium (on top of the provider equilibrium defined in Definition 7), as n increases to infinity.*

The proof of this proposition is given in Appendix D.2.

6.2.2. Results Before bounding the price of anarchy in this setting it is important to note that, under the provider equilibrium defined in Definition 7, both existence and uniqueness of a distribution equilibrium are guaranteed.

Proposition 8. *Suppose that latency functions are polynomial, i.e., $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ and $\hat{\ell}_p(y) = \hat{a}_p y^k$, for every p. There exists a unique distribution equilibrium when a nonatomic provider equilibrium is considered.*

The proof of this proposition is given in Appendix D.3. Given existence and uniqueness, we now move to the price of anarchy. In this setting, we obtain the following bound on the price of anarchy, which is proven in Appendix D.4.

Theorem 1. *Suppose that latency functions are polynomial, i.e., $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ and $\hat{\ell}_p(y) = \hat{a}_p y^k$, for every p. The price of anarchy of a distribution equilibrium using a non-atomic provider equilibrium is at most $k+1$.*

In contrast to Proposition 6, the above theorem highlights that the price of anarchy will be small in settings when there are a large number of providers. For example, the price of anarchy is simply 2 in the case of linear latencies, and more generally the price of anarchy is $k+1$ if congestion costs are polynomial with degree k. Interestingly, this is essentially the same price of anarchy as when no market structure exists, i.e., users directly choose providers based on congestion costs Roughgarden and Tardos (2002). Since the price of anarchy of the two-tier model (users and SaaSs) converges to one in the limit as the number of services grows Anselmi et al. (2011), Theorem 1 reveals that the addition of providers into the marketplace "undoes" the efficiency created by competition among services. Further, these results highlight that it is crucial to find ways to incentivize participation of IaaS/PaaS providers in the cloud marketplace, especially given the results in Section 5 which highlight that the profitability of providers decreases quickly with increasing competition.

7. Concluding Remarks

In this paper, we develop a novel model for the cloud computing marketplace which, for the first time includes: (i) the *three-tier* structure of the marketplace (including users, services, and providers), and (ii) the distinction between *shared* and *dedicated* latency in the cloud. The inclusion of these factors leads to novel qualitative insights about market power, user performance (the price of anarchy), and the differing impacts of shared and dedicated latencies.

We view this paper as a first step towards a deeper understanding of the cloud marketplace. As such, there are many extensions that are interesting to consider in future work. For example, we have considered one popular price structure, "charge per flow", but there are many other price structures that are available today, including "charge per instance", a fixed "membership" charge, etc. Additionally, there are many simplifications in the model considered here, e.g., that users and services are homogeneous and nonatomic, and that there is no market friction preventing services from switching providers. These assumptions are made to allow an analytic first step toward understanding the impact of market structure, and would of course be very interesting to remove with future research.

Endnotes

1. For an atomic model with linear latency functions, we have shown in Appendix A that all services that choose the same provider must set the same price at an equilibrium (cf. Proposition 10).

2. Since g_p is fixed, maximizing the objective function in the definition is equivalent to maximizing its profit $\beta_p x_p^{SE}(\beta_p, \boldsymbol{\beta}_{-p}^{PE}, \mathbf{g}) g_p$. For a provider p with $g_p = 0$, we have implicitly assumed that it aims to maximize the product of its user flow and its unit price, even if the set of services that choose this provider has a zero measure.

3. If $g_p = 1$ for some provider p, then a provider equilibrium does not exist. Since provider p is guaranteed to have a user flow of λ (regardless of the price it sets), it would like to charge an arbitrarily high price.

4. In contrast to a classical congestion game model, here the total flow of the p-th dark blue link (provider p) is $y_p = x_p g_p$.

5. Note that $\mathbf{x}^{SE}(\mathbf{g}, \boldsymbol{\beta})$ is defined as the unique user equilibrium resulting from a service equilibrium under $\mathbf{g}$ and $\boldsymbol{\beta}$. Since $\boldsymbol{\alpha}$ is such a service equilibrium, $\mathbf{x}^{SE}(\mathbf{g}, \boldsymbol{\beta})$ is the unique user equilibrium in $W(\mathbf{g}, \boldsymbol{\alpha})$.

6. In the proof of Proposition 4 we show that a provider price vector of the form (16) must be a provider equilibrium (cf. the discussion following (46)).

7. Such replica economies are studied commonly in the economics literature, for example in the context of core convergence (Hart (1979)).

Acknowledgments

This work was supported by NSF grants CNS-1319820, EPAS-1307794, and CNS-0846025.

References

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

Altman, E., T. Boulogne, R. El-Azouzi, T. Jiménez, L. Wynter. 2006. A survey on networking games in telecommunications. *Comput. Oper. Res.* **33**(2) 286–311.

Altman, Eitan, Urtzi Ayesta, Balakrishna Prabhu. 2008. Load balancing in processor sharing systems. *Proc. of ValueTools*. 1–10.

Amazon. 2014a. Ec2 pricing. *http://aws.amazon.com/ec2/pricing/* .

Amazon. 2014b. Simple queue service (amazon sqs). *http://aws.amazon.com/sqs/* .

Anselmi, J., B. Gaujal. 2011. The price of forgetting in parallel and non-observable queues. *Perform. Eval.* **68**(12) 1291–1311. doi:10.1016/j.peva.2011.07.023. URL http://dx.doi.org/10.1016/j.peva.2011.07.023.

Anselmi, Jonatha, Urtzi Ayesta, Adam Wierman. 2011. Competition yields efficiency in load balancing games. *Perform. Eval.* **68** 986–1001. doi:http://dx.doi.org/10.1016/j.peva.2011.07.005. URL http://dx.doi.org/10.1016/j.peva.2011.07.005.

Ardagna, D., B. Panicucci, M. Passacantando. 2012. Generalized nash equilibria for the service provisioning problem in cloud systems. *IEEE Trans. on Services Computing (Preprint)* .

Chen, H. L., J. R. Marden, A. Wierman. 2009. On the impact of heterogeneity and back-end scheduling in load-balancing designs. *Proc. of IEEE INFOCOM*.

Cloudtimes. 2012. Amazon EC2 outage reveals challenges of cloud computing. *http://cloudtimes.org/2012/07/03/amazon-outage-risk-computing/* .

Dafermos, S. C., F. T. Sparrow. 1969. The traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards, Series B* **73** 91–118.

Economides, N., J. Tåg. 2012. Network neutrality on the internet: a two-sided market analysis. *Information Economics and Policy* .

Feng, Y., B. Li, B. Li. 2013. Price competition in an oligopoly cloud market. *Under submission* .

Google. 2014. App engine pricing. *http://cloud.google.com/pricing/* .

Hart, O. D. 1979. Monopolistic competition in a large economy with differentiated commodities. *Review of Economic Studies* **46** 1–30.

Haviv, M. 2001. The Aumann-Shapely pricing mechanism for allocating congestion costs. *Operations Research Letters* **29**(5) 211–215.

Hayrapetyan, A., E. Tardos, T. Wexler. 2007. A network pricing game for selfish traffic. *Distributed Computing* **19** 255–266.

Hong, Yu-Ju, Jiachen Xue, Mithuna Thottethodi. 2011. Dynamic server provisioning to minimize cost in an iaas cloud. *Proc. of ACM SIGMETRICS*. 147–148.

Maskin, Eric, Jean Tirole. 1988. A theory of dynamic oligopoly, ii: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica* **56**(3) 571–99. URL `http://ideas.repec.org/a/ecm/emetrp/v56y1988i3p571-99.html`.

Musacchio, J., G. Schwartz, J. Walrand. 2009. A two-sided market analysis of provider investment incentives with an application to the net-neutrality issue. *Review of Network Economics* **8**(1).

NetworkWorld. 2012. Amazon outage one year later: are we safer? *http://www.networkworld.com/news/2012/042712-amazon-outage-258735.html* .

Roughgarden, T., E. Tardos. 2002. How bad is selfish routing? *Journal of the ACM* **49** 236–259.

Roughgarden, T., E. Tardos. 2004. Bounding the inefficiency of equilibria in nonatomic congestion games. *Games and Economic Behavior* **47**.

Song, Yang, Murtaza Zafer, Kang-Won Lee. 2012. Optimal bidding in spot instance market. *Proc. of IEEE INFOCOM*. 190–198.

Teng, F., F. Magoules. 2010. A new game theoretical resource allocation algorithm for cloud computing. *Advances in Grid and Pervasive Computing*. 321–330.

van den Nouweland, A., P. Borm, W. van Golstein Brouwers, R. Bruinderink, S. Tijs. 1996. A game theoretic approach to problems in telecommunication. *Manage. Sci.* **42**(2) 294–303.

Yolken, Benjamin, Nicholas Bambos. 2008. Game based capacity allocation for utility computing environments. *Proc. of ValueTools*. 1–8.

Electronic Companion

Appendix A: Approximation of Service Equilibrium Prices

In this appendix, we derive some approximation result for the model with non-atomic services introduced in Section 4. In particular, we will show that a symmetric (atomic) service equilibrium (at which all services that choose the same provider charge the same price) must converge to its non-atomic counterpart characterized in Proposition 3, as the number of services increases to infinity.

We first characterize the (atomic) service equilibrium (cf. Definition 2) in the following proposition.

Proposition 9. *Given a mapping* $\mathbf{f}$ *and a provider price vector* $\boldsymbol{\beta}$*, an atomic service equilibrium* $\boldsymbol{\alpha}^{SE}$ *must satisfy conditions (36), for every service* s *that has a positive flow at the unique user equilibrium induced by* $\mathbf{f}$ *and* $\boldsymbol{\alpha}^{SE}$*.*

Proof: We define $\overline{\mathcal{S}}$ as the set of services that have positive flow at the unique user equilibrium resulting from $\mathbf{f}$ and $\boldsymbol{\alpha}^{SE}$. For a service $s \in \overline{\mathcal{S}}$, its equilibrium price α_s^{SE} is an optimal solution to the following optimization problem,

$$\begin{aligned}
\text{maximize}_{\alpha_s \geq 0} \quad & (\alpha_s - \beta_{f_s})x_s \\
\text{subject to} \quad & \tilde{\ell}_{f_{s'}}(x_{s'}) + \hat{\ell}_{f_{s'}}(y_{f_{s'}}) + \alpha_{s'} = \tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(y_{f_s}) + \alpha_s, \qquad \forall\, s' \in \overline{\mathcal{S}}, \\
& \sum\nolimits_{s: f_s = p} x_s = y_p, \qquad \forall\, p, \\
& \sum\nolimits_{s} x_s = \lambda, \\
& x_{s'} > 0, \qquad \forall\, s' \in \overline{\mathcal{S}}, \\
& x_{s'} = 0, \qquad \forall\, s' \notin \overline{\mathcal{S}}.
\end{aligned} \tag{35}$$

Its KKT condition yields,

$$\alpha_s^{SE} - \beta_{f_s} = x_s \tilde{\ell}'_{f_s}(x_s) + x_s \frac{c_1 + 1}{c_2 + c_3 + c_4}, \tag{36}$$

where

$$c_1 = \tilde{\ell}'_{f_s}(y_{f_s}) \sum_{p' \neq f_s} \frac{\sum\limits_{s' \in \overline{\mathcal{S}}: f_{s'} = p'} (\tilde{\ell}'_{p'}(x_{s'}))^{-1}}{1 + \hat{\ell}'_{p'}(y_{p'}) \sum\limits_{s' \in \overline{\mathcal{S}}: f_{s'} = p'} (\tilde{\ell}'_{p'}(x_{s'}))^{-1}},$$

$$c_2 = \sum_{s' \in \overline{\mathcal{S}}, s' \neq s : f_{s'} = f_s} (\tilde{\ell}'_{f_s}(x_{s'}))^{-1},$$

$$c_3 = \tilde{\ell}'_{f_s}(y_{f_s}) \sum_{s' \in \overline{\mathcal{S}}, s' \neq s : f_{s'} = f_s} (\tilde{\ell}'_{f_s}(x_{s'}))^{-1} \sum_{p' \neq f_s} \frac{\sum_{s' \in \overline{\mathcal{S}} : f_{s'} = p'} (\tilde{\ell}'_{p'}(x_{s'}))^{-1}}{1 + \hat{\ell}'_{p'}(y_{p'}) \sum_{s' \in \overline{\mathcal{S}} : f_{s'} = p'} (\tilde{\ell}'_{p'}(x_{s'}))^{-1}},$$

$$c_4 = \sum_{p' \neq f_s} \frac{\sum_{s' \in \overline{\mathcal{S}} : f_{s'} = p'} (\tilde{\ell}'_{p'}(x_{s'}))^{-1}}{1 + \hat{\ell}'_{p'}(y_{p'}) \sum_{s' \in \overline{\mathcal{S}} : f_{s'} = p'} (\tilde{\ell}'_{p'}(x_{s'}))^{-1}}.$$

■

PROPOSITION 10. *Suppose that the cost functions, $\{\tilde{\ell}_p\}_{p=1}^P$ and $\{\hat{\ell}_p\}_{p=1}^P$, are linear with positive slopes $\{\tilde{a}_p\}_{p=1}^P$ and $\{\hat{a}_p\}_{p=1}^P$, respectively. At an atomic service equilibrium, all services that choose the same provider and have a positive user flow must charge the same price.*

Proof: Let services s_1, s_2 choose the same provider p. Without loss of generality, suppose that $\alpha_{s_1} < \alpha_{s_2}$. It follows that $x_{s_1} > x_{s_2}$, because at a user equilibrium we have $\tilde{\ell}_p(x_{s_1}) + \alpha_{s_1} = \tilde{\ell}_p(x_{s_2}) + \alpha_{s_2}$. According to Eq. (36), we have

$$\alpha_{s_1} = x_{s_1}\tilde{a}_p + x_{s_1}\frac{c_1 + 1}{c_2(\mathbf{x}_{-s_1}) + c_3 + c_4} + \beta_{s_1},$$

$$\alpha_{s_2} = x_{s_2}\tilde{a}_p + x_{s_2}\frac{c_1 + 1}{c_2(\mathbf{x}_{-s_2}) + c_3 + c_4} + \beta_{s_2},$$

where $\mathbf{x}_{-s}$ denotes the user flow of all services other than s. Here, c_1, c_3 and c_4 are the same for all services (cf. their expressions in the proof of Proposition 9), and

$$c_2(\mathbf{x}_{-s_1}) = (\tilde{a}_p)^{-1} + \sum_{s \in \overline{\mathcal{S}}, s \neq s_1 : f_s = p} (\tilde{a}_p)^{-1}$$

$$c_2(\mathbf{x}_{-s_2}) = (\tilde{a}_p)^{-1} + \sum_{s \in \overline{\mathcal{S}}, s \neq s_2 : f_s = p} (\tilde{a}_p)^{-1}$$

It can be seen that $c_2(x_{s_2}) = c_2(x_{s_1})$, which implies that $\alpha_{s_1} > \alpha_{s_2}$. A contradiction arises. The desired result follows. ■

We now consider a sequence of models indexed by $n = 1, 2, \ldots$. For each model n, we assume the following.

1. The aggregate user flow is $n\lambda$.
2. The P providers in the market charge a price vector $(\beta_1^n, \ldots, \beta_P^n)$.
3. There are nS services in the market. We define

$$g_p^n \triangleq \frac{1}{nS} \left| s : f_s^n = p \right|,$$

i.e., g_p^n is the fraction of services that choose provider p. We assume that for every p, $\lim_{n\to\infty} g_p^n = g_p$.

4. The dedicated and shared resource of each provider scales proportionally with n, and therefore the latency experienced by a user of service s obeys $\tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(y_{f_s}/n)$.

For the rest of this appendix, we use a superscript n to denote the notations associated with a model indexed by n.

Proposition 11. *Let $(\alpha_1^n, \dots, \alpha_{nS}^n)$ be a symmetric service equilibrium (where all services that choose the same provider charge the same price) of a model with nS services. We have*

$$\lim_{n\to\infty} \alpha_s^n - \beta_{f_s^n}^n - x_s^n \tilde{\ell}'_{f_s^n}(x_s^n) = 0, \qquad \forall s : x_s^n > 0, \tag{37}$$

where $\{x_s^n\}_{s=1}^{Sn}$ is the unique user equilibrium resulting from $\mathbf{f}^n$ and the service equilibrium $\boldsymbol{\alpha}^n$.

Note that we have shown in Proposition 10 that if cost functions are linear, a service equilibrium must be symmetric. The expression in Eq. (37), $\beta_{f_s^n}^n + x_s^n \tilde{\ell}'_{f_s^n}(x_s^n)$, is of the form of the (non-atomic) service equilibrium price characterized in Proposition 3.

Proof: At a symmetric service equilibrium, since all services choose the same provider charge the same price, they will have the same flow at a user equilibrium. We can therefore use $x_{f_s^n}^n$ to denote the user flow of service s, where f_s^n the provider chosen by service s. We then have following relation:

$$\sum_p nS g_p^n x_p^n = n\lambda.$$

Based on the preceding relation, we can still write service prices in the form of (36) at a symmetric service equilibrium of the model n, with the following parameters (where we let $p^n = f_s^n$ for conciseness):

$$\begin{aligned}
c_1^n &= \hat{\ell}'_{p^n}(y_{p^n}^n) \sum_{p \neq p^n} \frac{g_p^n nS(\tilde{\ell}'_p(x_{p^n}^n))^{-1}}{1 + g_p^n nS \hat{\ell}'_p(y_{p^n}^n)(\tilde{\ell}'_p(x_{p^n}^n))^{-1}}, \\
c_2^n &= (g_{p^n}^n nS - 1)(\tilde{\ell}'_{p^n}(x_{p^n}^n))^{-1}, \\
c_3^n &= \hat{\ell}'_{p^n}(y_{p^n}^n)(g_{p^n}^n nS - 1)(\tilde{\ell}'_{p^n}(x_{p^n}^n))^{-1} \sum_{p \neq p^n} \frac{g_{p^n}^n nS(\tilde{\ell}'_p(x_{p^n}^n))^{-1}}{1 + g_p^n nS \hat{\ell}'_p(y_{p^n}^n)(\tilde{\ell}'_p(x_{p^n}^n))^{-1}}, \\
c_4^n &= \sum_{p \neq p^n} \frac{g_p^n nS(\tilde{\ell}'_p(x_{p^n}^n))^{-1}}{1 + g_p^n nS \hat{\ell}'_p(y_{p^n}^n)(\tilde{\ell}'_p(x_{p^n}^n))^{-1}}.
\end{aligned}$$

Since the limit of g_p^n exists, we have

$$\lim_{n\to\infty} c_1^n \to O(1), \qquad \lim_{n\to\infty} c_2^n + c_3^n + c_4^n \to \infty.$$

It follows that the desired result holds. ■

Appendix B: Proofs of results in Section 4

B.1. Proof of Proposition 2

Since the latency functions, $\tilde{\ell}_p$ and $\hat{\ell}_p$, are increasing and convex for every p, it is easy to check that problem (7) is a strictly convex optimization problem. As a result, there is a unique optimal solution. With Lagrangian multipliers $L_{1,p}, L_2, L_{3,p} \geq 0, L_{4,p} \geq 0$ (associated with each of the four constraints, respectively), the KKT conditions are given by

$$\tilde{\ell}_p(x_p) + \alpha_p + L_{1,p} + L_2 - L_{4,p} = 0, \qquad \forall p, \tag{38}$$

$$\hat{\ell}_p(y_p) - L_{1,p} - L_{3,p} = 0, \qquad \forall p. \tag{39}$$

Substituting (39) into (38), we obtain

$$\tilde{\ell}_p(x_p) + \hat{\ell}_p(y_p) + \alpha_p = -L_2 + L_{4,p} + L_{3,p}, \qquad \forall p. \tag{40}$$

According to complementary slackness we have $L_{3,p} y_p = 0$ and $L_{4,p} x_p = 0$. It follows from the nonnegativity of the Lagrangian multipliers $L_{3,p}$ and $L_{4,p}$ that the condition in (40) coincides with the equilibrium conditions in Definition 3 (with $\mu^{UE} = -L_2$). Because the optimization problem has a unique optimizer, there is a unique solution that satisfies the preceding KKT condition. That is, there exists a unique user equilibrium.

B.2. Proof of Proposition 3

Suppose first that a service equilibrium exists, and we will argue that the equilibrium prices satisfy (11). We then show the existence and uniqueness of a service equilibrium by arguing that it is the unique solution to an optimization problem.

Let $\boldsymbol{\alpha}^{SE}$ be a service equilibrium. Given the service distribution $\mathbf{g}$ and the prices set by other market participants at the equilibrium, a service that chooses a provider p aims to maximize her profit, $(\alpha - \beta_p) x(\alpha, \boldsymbol{\alpha}^{SE})$, over all nonnegative prices α, subject to the constraint in (9). The equilibrium price α_p^{SE} is an optimal solution to this optimization problem.

Let $(x_1^{SE}, \ldots, x_P^{SE}) \in W(\mathbf{g}, \boldsymbol{\alpha}^{SE})$ be the unique user equilibrium induced by $\boldsymbol{\alpha}^{SE}$. For every p with $x_p^{SE} > 0$, from (9) we have

$$\alpha_p^{SE} = \mu^{SE} - \hat{\ell}_p(g_p x_p^{SE}) - \tilde{\ell}_p(x_p^{SE}).$$

The profit of a service that chooses provider p can be written as a function of x_p^{SE},

$$(\mu^{SE} - \hat{\ell}_p(g_p x_p^{SE}) - \tilde{\ell}_p(x_p^{SE}) - \beta_p) x_p^{SE}.$$

The derivative of the preceding expression on x_p^{SE} must equal zero, and therefore

$$-\tilde{\ell}_p'(x_p^{SE}) x_p^{SE} + (\mu^{SE} - \hat{\ell}_p(g_p x_p^{SE}) - \tilde{\ell}_p(x_p^{SE})) - \beta_p^{PE} = 0,$$

which implies that

$$\alpha_p^{SE} - \beta_p = \tilde{\ell}'_p(x_p^{SE})x_p^{SE},$$

as desired in (11).

We have shown that a service equilibrium, if exists, must satisfy (11). Based on the service equilibrium $\boldsymbol{\alpha}^{SE}$, we define a service price vector $\overline{\boldsymbol{\alpha}}$ such that

$$\overline{\alpha}_p = \tilde{\ell}'_p(x_p^{SE})x_p^{SE} + \beta_p, \qquad \forall p, \tag{41}$$

where $(x_1^{SE}, \ldots, x_P^{SE}) \in W(\mathbf{g}, \boldsymbol{\alpha}^{SE})$. It is easy to see that $\overline{\boldsymbol{\alpha}}$ remains a service equilibrium, and results in the same user equilibrium $(x_1^{SE}, \ldots, x_P^{SE})$. It is also straightforward to check (from Definition 4) that a service price vector satisfying condition (41) must be a service equilibrium. Substituting (41) to Definition 3, we obtain the necessary and sufficient condition for a flow vector $\mathbf{x} = (x_1, \ldots, x_P)$ to be a user equilibrium resulting from a service equilibrium (under $\mathbf{g}$ and $\boldsymbol{\beta}$):

$$\begin{aligned}
&\tilde{\ell}_p(x_p) + \hat{\ell}_p(g_p x_p) + x_p\tilde{\ell}'_p(x_p) + \beta_p = \mu, && \forall p : x_p > 0, \\
&\tilde{\ell}_p(x_p) + \hat{\ell}_p(g_p x_p) + \beta_p \ge \mu, && \forall p : x_p = 0, \\
&\sum_p g_p x_p = \lambda.
\end{aligned} \tag{42}$$

which coincide with the KKT condition of the following optimization problem

$$\begin{aligned}
\text{minimize} \quad & \sum_p \Big[x_p \tilde{\ell}_p(x_p) + \beta_p x_p\Big] + \sum_p \int_0^{y_p} \hat{\ell}_p(z)dz \\
\text{subject to} \quad & g_p x_p = y_p, \qquad \forall p, \\
& \textstyle\sum_p y_p = \lambda, \\
& x_p \ge 0, \qquad \forall p.
\end{aligned} \tag{43}$$

Since the optimization problem (43) is strictly convex, it has a unique optimizer. As a result, there exists a unique user equilibrium, denoted by $\mathbf{x}^{SE}(\mathbf{g}, \boldsymbol{\beta})$ in this paper, resulting from all service equilibria under $(\mathbf{g}, \boldsymbol{\beta})$. We finally note that there exists a service equilibrium (e.g., the $\overline{\boldsymbol{\alpha}}$ defined in (41) that depends only on the user equilibrium $\mathbf{x}^{SE}(\mathbf{g}, \boldsymbol{\beta})$), and that all service equilibria must satisfy (11).

B.3. Proof of Proposition 4

In Step 1, we show that for every provider p, given the prices set by other providers $\boldsymbol{\beta}_{-p}$, its best response always yields itself a positive profit. This result serves as a basis for our proof in Step 2, where we show that every provider's best response is a singleton, and apply Kakutani's fixed-point theorem to argue the existence of a provider equilibrium in a modified game in which we restrict provider prices to lie in some compact set. We then show that a provider equilibrium of the modified game must also be a provider equilibrium of the original game (where provider prices can be any arbitrary nonnegative real numbers). In Step 3, we show that at a provider equilibrium, the price of every provider p is uniquely determined.

Step 1: The best response of every provider p yields itself a positive user flow.

For a provider p, given the prices set by other providers $\boldsymbol{\beta}_{-p}$, let $B_p(\boldsymbol{\beta}_{-p})$ denote the set of $\overline{\beta}_p$ such that

$$\overline{\beta}_p \in \arg\max_{\beta_p \geq 0} \beta_p x_p^{SE}(\mathbf{g}, \beta_p, \boldsymbol{\beta}_{-p}),$$

where $x_p^{SE}(\mathbf{g}, \beta_p, \boldsymbol{\beta}_{-p})$ is the p-th component of the user equilibrium $\mathbf{x}^{SE}(\mathbf{g}, \beta_p, \boldsymbol{\beta}_{-p})$. For any nonnegative vector $\boldsymbol{\beta}_{-p}$, and every $\overline{\beta}_p \in B_p(\boldsymbol{\beta}_{-p})$, we will show that $\overline{\beta}_p x_p^{SE}(\mathbf{g}, \overline{\beta}_p, \boldsymbol{\beta}_{-p}) > 0$.

Suppose not, and we either have $\overline{\beta}_p = 0$ or $x_p^{SE}(\mathbf{g}, \overline{\beta}_p, \boldsymbol{\beta}_{-p}) = 0$. Suppose first that $\overline{\beta}_p = 0$. At the user equilibrium $\mathbf{x}^{SE}(\mathbf{g}, 0, \boldsymbol{\beta}_{-p})$, from (42) we have

$$(2\tilde{a}_p + \hat{a}_p g_p) x_p^{SE}(\mathbf{g}, 0, \boldsymbol{\beta}_{-p}) = \mu_1,$$

where μ_1 is the user effective cost at the user equilibrium. We note that for linear latency functions, the effective cost μ_1 must be positive. Provider p can charge a price of $\mu_1/2$, and guarantee a positive flow no less than $x_p^{SE}(\mathbf{g}, 0, \boldsymbol{\beta}_{-p})/2$, because the user effective cost increases with provider p's price. Since provider p can obtain a positive profit by properly setting its price, it follows that $x_p^{SE}(\mathbf{g}, \overline{\beta}_p, \boldsymbol{\beta}_{-p}) > 0$, if $\overline{\beta}_p$ is a best response against $\boldsymbol{\beta}_{-p}$.

Step 2: Existence of a provider equilibrium.

Given a provider price vector $\boldsymbol{\beta}$, we define $\mathcal{P}_{\boldsymbol{\beta}} = \{p : x_p^{SE}(\mathbf{g}, \boldsymbol{\beta}) > 0\}$ as the set of providers with positive flow at the user equilibrium $\mathbf{x}^{SE}(\mathbf{g}, \boldsymbol{\beta})$. Given the prices set by other providers $\boldsymbol{\beta}_{-p}$, provider p's best response is an optimal solution to the following problem:

$$\begin{aligned}
\text{maximize}_{\beta_p \geq 0} \quad & \beta_p x_p^{SE}(\mathbf{g}, \boldsymbol{\beta}) && (44)\\
\text{subject to} \quad & 2\tilde{a}_p x_p^{SE}(\mathbf{g}, \boldsymbol{\beta}) + \hat{a}_p g_p x_p^{SE}(\mathbf{g}, \boldsymbol{\beta}) + \beta_p \\
& \quad = 2\tilde{a}_{p'} x_{p'}^{SE}(\mathbf{g}, \boldsymbol{\beta}) + \hat{a}_{p'} g_{p'} x_{p'}^{SE}(\mathbf{g}, \boldsymbol{\beta}) + \beta_{p'}, \qquad \forall p' \in \mathcal{P}_{\boldsymbol{\beta}}, && (45)\\
& \sum\nolimits_p g_p x_p^{SE}(\mathbf{g}, \boldsymbol{\beta}) = \lambda, \\
& x_{p'}^{SE}(\mathbf{g}, \boldsymbol{\beta}) > 0, \qquad \forall p' \in \mathcal{P}_{\boldsymbol{\beta}}, \\
& x_{p'}^{SE}(\mathbf{g}, \boldsymbol{\beta}) = 0, \qquad \forall p' \notin \mathcal{P}_{\boldsymbol{\beta}},
\end{aligned}$$

where $\boldsymbol{\beta}=(\beta_p,\boldsymbol{\beta}_{-p})$, and the constraint (45) is legitimate because provider p's best response must yield it a positive user flow. It is straightforward to check that there exists an optimal solution to the problem (44), for any given $\boldsymbol{\beta}_{-p}$. Let $\overline{\beta}_p$, together with the user equilibrium $\mathbf{x}^{SE}(\mathbf{g},\overline{\beta}_p,\boldsymbol{\beta}_{-p})$, be an optimal solution to (44). We consider an alternative optimization problem with modified constraints:

$$\begin{aligned}
\text{maximize}_{\beta_p\geq 0}\quad & \beta_p z_p && (46)\\
\text{subject to}\quad & 2\tilde{a}_p z_p+\hat{a}_p g_p z_p+\beta_p=2\tilde{a}_{p'}z_{p'}+\hat{a}_{p'}g_{p'}z_{p'}+\beta_{p'},\quad \forall p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})},\\
& \sum\nolimits_p g_p z_p=\lambda,\\
& z_{p'}>0, \qquad \forall p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})},\\
& z_{p'}=0, \qquad \forall p'\notin\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})},
\end{aligned}$$

where in the first constraint (which is different from constraint (45)), the set of providers that have positive user flows depends only on the vector $(\overline{\beta}_p,\boldsymbol{\beta}_{-p})$ (instead of the variable β_p). We have shown in Step 1 that $\overline{\beta}_p$ yields provider p a positive user flow, and therefore $p\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}$. It follows that $\overline{\beta}_p$ and $\mathbf{x}^{SE}(\mathbf{g},\overline{\beta}_p,\boldsymbol{\beta}_{-p})$ is also an optimal solution to (46). Further, since the optimization problem (46) is strictly convex, it has a unique optimal solution. We therefore conclude that the set of optimizers of Problem (46) coincides with the set of optimizers of Problem (44), $B_p(\boldsymbol{\beta}_{-p})$, which must be a singleton.

The Lagrangian function associated with the optimization problem (46) is

$$\begin{aligned}
&\beta_p z_p-\sum\nolimits_{p'\neq p,p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}} L_{1,p'}\left[2\tilde{a}_{p'}z_{p'}+\hat{a}_{p'}g_{p'}z_{p'}+\beta_{p'}-(2\tilde{a}_p z_p+\hat{a}_p g_p z_p+\beta_p)\right]\\
&-L_2\left(\sum\nolimits_{p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}} g_{p'}z_{p'}-\lambda\right)-\sum\nolimits_{p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}} L_{3,p'}z_{p'}-\sum\nolimits_{p'\notin\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}} L_{4,p'}z_{p'}.
\end{aligned}$$

Since for every $p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}$, we have $z_{p'}>0$, it follows from complementary slackness that $L_{3,p'}=0$. For the preceding Lagrangian function, taking its derivative with respect to β_p, z_p, and $z_{p'}$ for every $p'\neq p$ such that $p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}$, we obtain

$$\begin{aligned}
&z_p=\sum\nolimits_{p'\neq p,p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}} L_{1,p'},\\
&\beta_p=(2\tilde{a}_p+\hat{a}_p g_p)\sum\nolimits_{p'\neq p,p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}} L_{1,p'}+L_2 g_p,\\
&L_{1,p'}(2\tilde{a}_{p'}+\hat{a}_{p'}g_{p'})-L_2 g_{p'}=0, \qquad \forall p'\neq p, \qquad p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}.
\end{aligned}$$

From these equations we obtain the following (necessary and sufficient) optimality condition for the optimization problem (44):

$$\overline{\beta}_p=(2\tilde{a}_p+\hat{a}_p g_p)z_p+\frac{g_p z_p}{\sum_{p'\neq p:p'\in\mathcal{P}_{(\overline{\beta}_p,\boldsymbol{\beta}_{-p})}}\dfrac{g_{p'}}{2\tilde{a}_{p'}+\hat{a}_{p'}g_{p'}}}, \tag{47}$$

where $(z_1,\dots,z_P)=\mathbf{x}_p^{SE}(\mathbf{g},\overline{\beta}_p,\boldsymbol{\beta}_{-p})$ is the unique user equilibrium resulting from $\mathbf{g}$ and the provider price vector $(\overline{\beta}_p,\boldsymbol{\beta}_{-p})$.

We now consider a variant of the original game among those providers p with $g_p>0$.[8] We restrict these providers' action space to lie in the compact set $[0,C]$ with

$$C=(3\hat{\alpha}_{\max}+2\tilde{\alpha}_{\max}/g_{\min}+2\tilde{\alpha}_{\max})\lambda,$$

where $\hat{\alpha}_{\max}=\max_p\{\hat{\alpha}_p\}$, $\tilde{\alpha}_{\max}=\max_p\{\tilde{\alpha}_p\}$, and $g_{\min}=\min_{p:g_p>0}g_p$. For a provider p, given the prices set by other providers $\boldsymbol{\beta}_{-p}$, let $\tilde{B}_p(\boldsymbol{\beta}_{-p})$ denote the set of $\overline{\beta}_p$ such that

$$\overline{\beta}_p\in\arg\max_{\beta_p\in[0,C]}\beta_p x_p^{SE}(\mathbf{g},\beta_p,\boldsymbol{\beta}_{-p}).$$

For a given provider price vector $\boldsymbol{\beta}$, let $\tilde{B}(\boldsymbol{\beta})=\{\tilde{B}_p(\boldsymbol{\beta}_{-p})\}_{p:g_p>0}$. By the maximum theorem, $\tilde{B}(\boldsymbol{\beta})$ is an upper semicontinuous correspondence. For any provider p with $g_p>0$, given any price vector set by other providers $\boldsymbol{\beta}_{-p}$, it is straightforward to see that its best response (in its unrestricted action space) given in (47) satisfies

$$\overline{\beta}_p\le(2\hat{\alpha}_{\max}+2\tilde{\alpha}_{\max}/g_{\min})\lambda+(2\tilde{\alpha}_{\max}+\hat{\alpha}_{\max})\lambda=(3\hat{\alpha}_{\max}+2\tilde{\alpha}_{\max}/g_{\min}+2\tilde{\alpha}_{\max})\lambda=C,$$

where the first inequality is true because $x_p^{SE}(\mathbf{g},\overline{\beta}_p,\boldsymbol{\beta}_{-p})\le\lambda/g_{\min}$. It follows that for every provider p with $g_p>0$, the price in (47) is its unique best response in its action space $[0,C]$, i.e., $\tilde{B}_p(\boldsymbol{\beta}_{-p})$ is a singleton, and is therefore convex. Since $\tilde{B}(\boldsymbol{\beta})$ is an upper semicontinuous and convex-valued correspondence, we can apply Kakutani's fixed-point theorem to conclude the existence of a provider equilibrium, for the modified game among those providers p with $g_p>0$.

Let $\{\beta_p\}_{p:g_p>0}$ be a provider equilibrium of the modified game. For every provider p with $g_p>0$, since the other providers p' with $g_{p'}=0$ have no influence on its profit, we have shown (in this step) that its price β_p of the form in (47) maximizes its profit over the unrestricted action space $[0,\infty)$. Further, the provider equilibrium $\{\beta_p\}_{p:g_p>0}$ determines a unique user equilibrium associated with a user effective cost μ. Since a provider p' with $g_{p'}=0$ has no influence on the user effective cost μ, it is straightforward to check that its best response over the action space $[0,\infty)$ (against the price vector $\{\beta_p\}_{p:g_p>0}$ and any prices set by the other providers) is of the form in (47), i.e.,

$$\beta_{p'}=2\tilde{a}_p x_p^{SE}(\mathbf{g},\beta_{p'},\boldsymbol{\beta}_{-p'}),\qquad\forall p':g_{p'}=0.$$

It follows that the price vector $\{\beta_p\}_{p:g_p>0}$, together with the price vector $\{\beta_{p'}\}_{p:g_{p'}>0}$ defined above, form a provider equilibrium of the original game (with unrestricted action space). Since for every provider p, its best response yields itself a positive user flow (cf. Step 1 of this proof), it follows that $x_p^{SE}(\mathbf{g},\boldsymbol{\beta}^{PE})$ must be positive at any provider equilibrium $\boldsymbol{\beta}^{PE}$. We also note that a provider equilibrium must satisfy the condition in (47).

Step 3: Uniqueness of a provider equilibrium.

Substituting (47) to (42), we obtain the necessary condition for a nonnegative vector $(z_1, \ldots, z_p)$ to be a user equilibrium $\mathbf{x}^{SE}(\mathbf{g}, \boldsymbol{\beta}^{PE})$ that is resulted from a provider equilibrium $\boldsymbol{\beta}^{PE}$:

$$\begin{aligned} &2(2\tilde{a}_p z_p + \hat{a}_p g_p z_p) + \frac{g_p z_p}{\sum_{p' \neq p} \dfrac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}}} = \mu, \qquad \forall p, \\ &\sum_p g_p z_p = \lambda. \end{aligned} \tag{48}$$

These conditions coincide with the optimality condition of the following optimization problem

$$\begin{aligned} \text{minimize} \quad & \sum_p \frac{1}{2} g_p \left[2(2\tilde{a}_p + \hat{a}_p g_p) + \frac{g_p}{\sum_{p' \neq p} \dfrac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}}} \right] z_p^2, \\ \text{subject to} \quad & \sum_p g_p z_p = \lambda, \\ & z_p \geq 0, \qquad \forall p. \end{aligned}$$

This is a strictly convex optimization problem that has a unique optimal solution. It follows that given a distribution $\mathbf{g}$, all provider equilibria result in a unique user equilibrium. Let $(x_1^{PE}, \ldots, x_P^{PE})$ denote this unique user equilibrium, and provider equilibrium prices can be uniquely determined by the condition in (47), i.e.,

$$\beta_p^{PE} = (2\tilde{a}_p + \hat{a}_p g_p) x_p^{PE} + \frac{g_p x_p^{PE}}{\sum_{p' \neq p} \dfrac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}}}.$$

Appendix C: Proofs of results in Section 5

C.1. Proof of Proposition 5

For the case with $P = 2$, we can write the functions $f_1(g_1, g_2)$ and $f_2(g_1, g_2)$ as (cf. the definition in (25))

$$f_1(g_1,g_2) = \frac{1}{\sqrt{\tilde{a}_1}}\left(2(2\tilde{a}_1 + \hat{a}_1 g_1) + \frac{g_1}{g_2}(2\tilde{a}_2 + \hat{a}_2 g_2)\right), \tag{49}$$

and

$$f_2(g_1,g_2) = \frac{1}{\sqrt{\tilde{a}_2}}\left(2(2\tilde{a}_2 + \hat{a}_2 g_2) + \frac{g_2}{g_1}(2\tilde{a}_1 + \hat{a}_1 g_1)\right). \tag{50}$$

Substituting $g_2 = 1 - g_1$ (and $g_1 = 1 - g_2$) into Eq. (49) (and (50), respectively), we have

$$f_1(g_1,g_2) = \frac{1}{\sqrt{\tilde{a}_1}}\left(2(2\tilde{a}_1 + \hat{a}_1 g_1) + \frac{2\tilde{a}_2 g_1}{1-g_1} + \hat{a}_2 g_1\right), \tag{51}$$

and

$$f_2(g_1,g_2) = \frac{1}{\sqrt{\tilde{a}_2}}\left(2(2\tilde{a}_2 + \hat{a}_2 g_2) + \frac{2\tilde{a}_1 g_2}{1-g_2} + \hat{a}_1 g_2\right). \tag{52}$$

In this case with $P = 2$, the equilibrium condition (26) is equivalent to the KKT condition of the following optimization problem

$$\begin{aligned}
\text{minimize}\quad & \frac{1}{\sqrt{\tilde{a}_1}}\left(2(2\tilde{a}_1 g_1 + \frac{1}{2}\hat{a}_1 g_1^2) + 2\tilde{a}_2(-g_1 - \ln(1-g_1)) + \frac{1}{2}\hat{a}_2 g_1^2\right)\\
& + \frac{1}{\sqrt{\tilde{a}_2}}\left(2(2\tilde{a}_2 g_2 + \frac{1}{2}\hat{a}_2 g_2^2) + 2\tilde{a}_1(-g_2 - \ln(1-g_2)) + \frac{1}{2}\hat{a}_1 g_2^2\right)\\
\text{subject to}\quad & g_1 + g_2 = 0,\\
& g_1 \ge 0, \qquad g_2 \ge 0.
\end{aligned}$$

It is straightforward to check that the above optimization problem is strictly convex. We therefore conclude that there exists a unique vector (g_1, g_2) (an optimal solution to the preceding problem) that satisfies condition (26). The desired result follows.

Appendix D: Proofs of results in section 6

D.1. Proof of Proposition 6

We note that for any vector $(\mathbf{x}, \mathbf{g})$,

$$\ell(\mathbf{x},\mathbf{g}) \;= \sum_p g_p x_p(\tilde{a}_p x_p^k + \hat{a}(g_p x_p)^k) \ge \tilde{a}_{\min}\sum_p g_p x_p^{k+1} + \hat{a}_{\min}\sum_p (g_p x_p)^{k+1}.$$

It follows that the minimum aggregate latency cost cannot be less than the optimal value of the following optimization problem

$$\begin{aligned}
\text{minimize}\quad & \tilde{a}_{\min}\sum_p g_p x_p^{k+1} + \hat{a}_{\min}\sum_p (g_p x_p)^{k+1}\\
\text{subject to}\quad & \textstyle\sum_p g_p x_p = \lambda,\\
& \textstyle\sum_p g_p = 1.
\end{aligned}$$

The optimal value of the preceding problem is $\tilde{a}_{\min}\lambda^k + \hat{a}_{\min}(\lambda/P)^k$. We therefore have

$$\ell(\mathbf{x}^*, \mathbf{g}^*) \geq \tilde{a}_{\min}\lambda^k + \hat{a}_{\min}(\lambda/P)^k,$$

and the equality hods when $\tilde{a}_p = \tilde{a}_{\min}$ and $\hat{a}_p = \hat{a}_{\min}$ for every p. We have obtained a lower bound of $\ell(\mathbf{x}^*, \mathbf{g}^*)$. To establish an upper bound for the price of anarchy of a distribution equilibrium, it remains to derive an upper bound on $\ell(\mathbf{x}, \mathbf{g})$. Note that

$$\begin{aligned}\ell(\mathbf{x}, \mathbf{g}) &= \sum_p g_p x_p (\tilde{a}_p x_p^k + \hat{a}_p (g_p x_p)^k) \\ &= \sum_p \tilde{a}_p g_p x_p^{k+1} + \sum_p \hat{a}_p (g_p x_p)^{k+1}.\end{aligned}$$

The second term can be easily bounded by

$$\sum_p \hat{a}_p (g_p x_p)^{k+1} \leq \hat{a}_{\max}\lambda^{k+1}.$$

Since $(x_1, \ldots, x_P)$ is the user flow at a distribution equilibrium, it follows from Proposition 3 and Definition 6 that

$$\tilde{a}_p x_p^{k+1} = L \geq 0, \qquad \forall p : g_p > 0,$$

where L is the profit of a service that chooses provider p. Since $\sum_p g_p = 1$, we have

$$\sum_p \tilde{a}_p g_p x_p^{k+1} = L.$$

To upper bound $\ell(\mathbf{x}, \mathbf{g})$, we only need to upper bound L. Since $x_p = (L/\tilde{a}_p)^{1/k+1}$ for any p with $g_p > 0$, from $\sum_p g_p x_p = \lambda$ we have

$$\sum_p g_p \left(\frac{L}{\tilde{a}_p}\right)^{1/(k+1)} = L^{\frac{1}{k+1}} \sum_p \frac{g_p}{(\tilde{a}_p)^{1/(k+1)}} = \lambda.$$

Since $\sum_p g_p = 1$, we have

$$\left(\frac{L}{\tilde{a}_{\max}}\right)^{\frac{1}{k+1}} \leq \lambda,$$

and therefore $L \leq \lambda^{k+1}\tilde{a}_{\max}$. In a summary,

$$\ell(\mathbf{x}, \mathbf{g}) = \sum_p \tilde{a}_p g_p x_p^{k+1} + \sum_p \hat{a}_p (g_p x_p)^{k+1} \leq \lambda^{k+1}\tilde{a}_{\max} + \hat{a}_{\max}\lambda^{k+1},$$

and therefore

$$\mathrm{PoA} = \frac{\ell(\mathbf{x}, \mathbf{g})}{\ell(\mathbf{x}^*, \mathbf{g}^*)} \leq \frac{\tilde{a}_{\max} + \hat{a}_{\max}}{\tilde{a}_{\min} + \hat{a}_{\min}/P^k}.$$

D.2. Proof of Proposition 7

To prove Proposition 7, we will need the following lemma.

Lemma 1. *In a game $\mathcal{G}_n$, suppose that a provider of some type p attracts a positive amount of services at a distribution equilibrium (on top of the provider equilibrium defined in Definition 7). Then, all type-p providers must attract the same (positive) amount of services, have the same user flow, and set the same price at this equilibrium.*

Proof: Let $(\mathbf{g}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ be a distribution equilibrium (on top of the provider equilibrium defined in Definition 7). Let providers i and j be of the same type p. By some abuse of notations, within this proof we will use subscript i (or j) to denote a quantity associated with provider i (or provider j, respectively).

If provider i attracts a positive amount of services, i.e., if $g_i > 0$, we first argue that g_j must be positive, too. Suppose not, and we have $g_j = 0$. Let x_i and x_j denote the user flow of provider i and j at the equilibrium, respectively. In Proposition 8 we have shown that $x_i > 0$. From (56) we have $\tilde{a}_p x_j^2 \leq \tilde{a}_p x_i^2$, i.e., provider j yields a service at most the same profit as provider i. It follows that $x_j \leq x_i$, which implies that

$$2(2\tilde{a}_p x_j + \hat{a}_p g_j x_j) < 2(2\tilde{a}_p x_i + \hat{a}_p g_i x_i),$$

i.e., users experience a less latency at provider j. This contracts with the definition of a user equilibrium (cf. conditions (57) and (58)). Since $g_i g_j > 0$, from (55) we have $\tilde{a}_p x_j^2 = \tilde{a}_p x_i^2$, i.e., both providers have the same user flow. It then follows from the conditions in (57) and (58) that $g_i = g_j$. Finally, from (58) we conclude that $\beta_i = \beta_j$. ■

Lemma 1 shows that providers of the same type can be regarded identical. For a game $\mathcal{G}_n$, we can therefore use g_p^n to denote the amount of services that choose a single provider of type p, and for every p with $g_p > 0$, we let x_p^n denote the user flow of a single service that chooses a type-p provider. The following conditions will be useful in the proof of Proposition 7:

$$\sum_p q_p^n g_p^n = 1, \qquad \sum_{p: g_p > 0} q_p^n g_p^n x_p^n = \lambda, \qquad , \forall n.$$

We now prove Proposition 7. In a game $\mathcal{G}_n$, let $(\mathbf{g}^n, \boldsymbol{\beta}^n)$ be a distribution equilibrium on top of the provider equilibrium defined in Definition 7. We must have

$$\beta_p^n = (2\tilde{a}_p + \hat{a}_p g_p^n) x_p^n, \qquad \forall p, \tag{53}$$

where $\mathbf{x}^n$ is the user equilibrium induced by $(\mathbf{g}^n, \boldsymbol{\beta}^n)$.

We will consider a case where services respond to the price change of a single provider by resetting their prices, instead of switching to some other providers. In other words, we assume that the service distribution at the distribution equilibrium, $\mathbf{g}^n$, is not affected by a single provider's price. On the other hand, the prices of services that choose a single provider, as well as the user flow at this provider, will change according to the price set by this provider.

In a game $\mathcal{G}_n$, suppose that all providers except one type-p provider set their prices according to the equilibrium $\boldsymbol{\beta}^n$. For the provider of type p, we will first bound the difference between the price in (53) and its best response, and then, show that the price in (53) yields the provider approximately its maximum profit.

Given the service distribution $\mathbf{g}^n$, we have shown in Proposition 4 that a type-p provider's best response satisfies

$$\overline{\beta}_p^n = (2\tilde{a}_p + \hat{a}_p g_p^n)\overline{x}_p^n + \frac{g_p^n \overline{x}_p^n}{\sum_{p' \neq p} n q_{p'}^n \dfrac{g_{p'}^n}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}^n} + (nq_p^n - 1)\dfrac{g_p^n}{2\tilde{a}_p + \hat{a}_p g_p^n}},$$

where $\overline{x}_p^n$ is the provider's user flow, at the unique user equilibrium induced by $\mathbf{g}^n$ and the provider price vector $(\overline{\beta}_p^n, \boldsymbol{\beta}^n)$ (here, all providers except the type-p provider set their prices according to the equilibrium $\boldsymbol{\beta}^n$). Since $\overline{\beta}_p^n$ is higher than the price β_p^n in (53), it is straightforward to check that $\overline{x}_p^n \leq x_p^n$. For large enough n, we have

$$\begin{aligned}
\overline{\beta}_p^n = {} & (2\tilde{a}_p + \hat{a}_p g_p^n)\overline{x}_p^n + \frac{g_p^n \overline{x}_p^n}{\sum_{p' \neq p} n q_{p'}^n \dfrac{g_{p'}^n}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}^n} + (nq_p^n - 1)\dfrac{g_p^n}{2\tilde{a}_p + \hat{a}_p g_p^n}} \\
= {} & (2\tilde{a}_p + \hat{a}_p g_p^n)\overline{x}_p^n + \frac{g_p^n \overline{x}_p^n}{n\sum_{p'} \dfrac{q_{p'}^n g_{p'}^n}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}^n} - \dfrac{g_p^n}{2\tilde{a}_p + \hat{a}_p g_p^n}} \\
\leq {} & (2\tilde{a}_p + \hat{a}_p g_p^n)\overline{x}_p^n + \frac{g_p^n \overline{x}_p^n}{n\sum_{p'} \dfrac{q_{p'}^n g_{p'}^n}{2\tilde{a}_{\max} + 2\hat{a}_{\max}/q_{\min}} - \dfrac{1}{\hat{a}_p}} \\
= {} & (2\tilde{a}_p + \hat{a}_p g_p^n)\overline{x}_p^n + \frac{g_p^n \overline{x}_p^n}{\dfrac{nq_{\min}}{2\tilde{a}_{\max} q_{\min} + 2\hat{a}_{\max}} - \dfrac{1}{\hat{a}_p}} \\
\leq {} & (2\tilde{a}_p + \hat{a}_p g_p^n)\overline{x}_p^n + \frac{g_p^n x_p^n}{\dfrac{nq_{\min}}{2\tilde{a}_{\max} q_{\min} + 2\hat{a}_{\max}} - \dfrac{1}{\hat{a}_p}} \\
\leq {} & (2\tilde{a}_p + \hat{a}_p g_p^n)\overline{x}_p^n + \frac{2\lambda/q_{\min}}{\dfrac{nq_{\min}}{2\tilde{a}_{\max} q_{\min} + 2\hat{a}_{\max}} - \dfrac{1}{\hat{a}_p}},
\end{aligned}$$

where the first inequality is true because for large enough n, we have $g_{p'}^n \leq 2/q_{\min}$, the third equality follows from the fact that $\sum_{p'} q_{p'} g_{p'} = 1$, and the last inequality is true because large enough n, we have $g_p^n x_p^n \leq 2\lambda/q_{\min}$ (note that $\sum_{p'} q_{p'}^n g_{p'}^n x_{p'}^n = \lambda$). For notational convenience, we define

$$\delta^n = \frac{2\lambda/q_{\min}}{\dfrac{nq_{\min}}{2\tilde{a}_{\max}q_{\min}+2\hat{a}_{\max}} - \dfrac{1}{\hat{a}_p}}.$$

We can therefore bound the provider's profit loss resulting from the price β_p^n in (53), ϵ^n:

$$\begin{aligned}\epsilon^n &= ((2\tilde{a}_p+\hat{a}_p g_p^n)\overline{x}_p^n+\delta^n)g_p^n\overline{x}_p^n - ((2\tilde{a}_p+\hat{a}_p g_p^n)x_p^n)g_p^n x_p^n \\ &\leq ((2\tilde{a}_p+\hat{a}_p g_p^n)x_p^n+\delta^n)g_p^n x_p^n - ((2\tilde{a}_p+\hat{a}_p g_p^n)x_p^n)g_p^n x_p^n \\ &= \delta^n g_p^n x_p^n \\ &\leq 2\delta^n\lambda/q_{\min},\end{aligned}$$

which converges to zero as n approaches infinity.

D.3. Proof of Proposition 8

Let $(\mathbf{g},\boldsymbol{\alpha},\boldsymbol{\beta})$ be a distribution equilibrium on top of a nonatomic provider equilibrium defined in Definition 33. It follows from Definition 7 that the resulting user flow $x_p^{SE}(\mathbf{g},\boldsymbol{\beta}^{PE})$ is positive for every p. Through a simple calculation it can be shown that an optimal solution to (33) and (34) satisfies

$$\beta_p = k((k+1)\tilde{a}_p+\hat{a}_p(g_p^k))(x_p^{SE}(\mathbf{g},\boldsymbol{\beta}))^k, \qquad \forall p, \tag{54}$$

We now show the existence and uniqueness of a distribution equilibrium by arguing that it is the unique solution to an optimization problem. It follows from the discussion after Definition 6 that the following conditions are necessary and sufficient for $(\mathbf{g},\boldsymbol{\beta})$ to be a distribution equilibrium, with $(x_1,\ldots,x_P)$ being the user equilibrium resulting from $\mathbf{g}$ and $\boldsymbol{\beta}$

$$\begin{cases} \tilde{a}_{p'}x_{p'}^{k+1} = \tilde{a}_p x_p^{k+1}, \qquad \text{if } \; g_p g_{p'} > 0, & (55)\\ \tilde{a}_p x_p^{k+1} \leq \tilde{a}_{p'}x_{p'}^{k+1}, \qquad \text{if } \; g_p = 0, \;\; g_{p'} > 0, & (56)\\ (k+1)\tilde{a}_{p'}x_{p'}^k + \hat{a}_{p'}(g_{p'}x_{p'})^k + \beta_{p'} = (k+1)\tilde{a}_p x_p^k + \hat{a}_p(g_p x_p)^k + \beta_p, \;\; \forall\, p,p', & (57)\\ \beta_p = k((k+1)\tilde{a}_p + \hat{a}_p(g_p^k))(x_p)^k, & (58)\\ \sum_p g_p x_p = \lambda, & (59)\\ \sum_p g_p = 1, & (60)\end{cases}$$

where (55) and (56) follow from Definition 6 and Proposition 3, and Eq. (57) (all providers yield users the same effective cost) is true because every provider p has a positive user flow x_p.

For a triple $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x}')$ that satisfies the conditions in (55) to (60), we define an alternative user flow vector $\mathbf{x}$ as follows. For any p' with $g_{p'} > 0$, we let $x_{p'} = x'_{p'}$, and for every p with $g_p = 0$, we make $x_p > x'_p$ such that

$$\tilde{a}_p x_p^2 = \tilde{a}_{p'} x_{p'}^2,$$

where p' is a provider with $g_{p'} > 0$. The modified triple $(\mathbf{g}, \boldsymbol{\beta}, \mathbf{x})$ must satisfy the following conditions:

$$\begin{cases} f_p(g_p) = f_{p'}(g_{p'}), & \text{if } \ g_p g_{p'} > 0, \quad (61) \\ f_p(0) = 2(\tilde{a}_p)^{\frac{1}{k+1}} \geq f_{p'}(g_{p'}), & \text{if } \ g_p = 0, \ \ g_{p'} > 0, \quad (62) \\ \sum_p g_p = 1, & \quad (63) \\ \tilde{a}_{p'} x_{p'}^2 = \tilde{a}_p x_p^2, & \forall p, \, p', \quad (64) \\ \sum_p x_p g_p = \lambda, & \quad (65) \end{cases}$$

where

$$f_p(g_p) = (\tilde{a}_p)^{-\frac{k}{k+1}} (2\tilde{a}_p + \hat{a}_p g_p). \tag{66}$$

The conditions (61)-(63) on the vector $\mathbf{g}$ coincide with the KKT condition of the following optimization problem

$$\begin{aligned} \text{minimize} \ & \sum_p \left[(\tilde{a}_p)^{-\frac{k}{k+1}} (2\tilde{a}_p g_p + \frac{1}{2} \hat{a}_p g_p^{k+1}) \right] \qquad (67) \\ \text{subject to} \ & \sum_p g_p = 1, \\ & g_p \geq 0, \qquad \forall p. \end{aligned}$$

Since the above optimization problem is strictly convex, it has a unique optimal solution. As a result, for price-taking providers, there exists a unique service distribution $\mathbf{g}$ that is resulting from a distribution equilibrium. On the other hand, given an optimal solution to (67), we can calculate the corresponding distribution equilibrium through conditions (55)-(60). It follows that there exists a unique distribution equilibrium (for price-taking providers), which can be characterized by the optimal solution to (67).

D.4. Proof of Theorem 1

We have shown in Proposition 8 that there exists a unique distribution equilibrium. We note that a vector $(\mathbf{g}, \mathbf{x})$ that satisfies conditions (61)-(65) yields the same aggregate user latency as the unique distribution equilibrium, because they are different only on the user flow of those providers p with $g_p = 0$. Hence, to prove this theorem, we only need to show that a vector that satisfies conditions (61)-(65) has a price of anarchy no more than 2.

In Step 1, we first characterize the vector that satisfies conditions (61)-(65) as an optimal solution to an optimization problem. This result enables us to establish the upper bound on its price of anarchy in Step 2.

Step 1: A vector $(\mathbf{g},\mathbf{x})$ *that satisfies conditions* (61)-(65) *is an optimal solution to the optimization problem in* (68)*.*

In this step, we aim to show that a vector $(\mathbf{g},\mathbf{x})$ that satisfies conditions (61)-(65) is an optimal solution to the following optimization problem

$$\begin{aligned}\text{minimize}\ & \textstyle\sum_p g_p x_p((k+1)\tilde{a}_p x_p^k + \hat{a}_p (g_p x_p)^k) \\ \text{subject to}\ & \textstyle\sum_p g_p x_p = \lambda, \\ & \textstyle\sum_p g_p = 1, \\ & g_p \geq 0, \qquad x_p \geq 0, \qquad \forall p.\end{aligned} \tag{68}$$

We will first argue that there exists an optimal solution to the above optimization problem. Fixing $\mathbf{g}$, we investigate the following optimization problem on $\mathbf{x}$,

$$\begin{aligned}\text{minimize}\ & \textstyle\sum_p x_p^{k+1}(g_p((k+1)\tilde{a}_p + \hat{a}_p g_p^k)) \\ \text{subject to}\ & \textstyle\sum_p g_p x_p = \lambda, \\ & x_p \geq 0, \qquad \forall p.\end{aligned} \tag{69}$$

Optimization problem (69) is convex in $\mathbf{x}$. Its KKT condition yields

$$x_p = \frac{L}{2(2\tilde{a}_p + \hat{a}_p g_p)}, \tag{70}$$

where L is the multiplier. It follows from the constraint $\sum_p g_p x_p = \lambda$ that

$$L = \left(\frac{\lambda}{\sum_p \frac{g_p}{((k+1)((k+1)\tilde{a}_p + \hat{a}_p g_p))^{\frac{1}{k}}}} \right)^k. \tag{71}$$

Substituting $\mathbf{x}$ into (68), we obtain the following optimization problem on $\mathbf{g}$

$$\begin{aligned}\text{minimize}\ & L^{\frac{k+1}{k}} \textstyle\sum_p \frac{g_p}{(k+1)((k+1)\tilde{a}_p + \hat{a}_p g_p^k)^{\frac{1}{k}}} \\ \text{subject to}\ & \textstyle\sum_p g_p = 1, \\ & g_p \geq 0, \qquad \forall p.\end{aligned} \tag{72}$$

Since the objective function (72) is continuous and the feasible region is compact, the above optimization problem has a solution $\mathbf{g}$; this solution $\mathbf{g}$, together with the vector $\mathbf{x}$ determined by (70) and (71), must be an optimal solution to the original optimization problem (68).

We now show that there exists an optimal solution to the optimization problem (68) that satisfies the equilibrium conditions (61)-(65). Let $(\mathbf{g},\mathbf{x})$ be an optimal solution to (68). We note that the optimization problem in (68) is non-convex. It is straightforward to check the regularity conditions: actually, the gradients of the constraints are linearly independent at any feasible solution to the optimization problem (68). We therefore can apply necessary KKT conditions to this optimization problem. Its Lagrangian function is given by

$$\sum_p g_p x_p((k+1)\tilde{a}_p x_p^k + \hat{a}_p g_p^k x_p^k) + L_1(\sum_p g_p x_p - \lambda) + L_2(\sum_p g_p - 1) - \sum_p G_p x_p - \sum_p Z_p x_p,$$

where L_1, L_2, $Z_p \geq 0$ and $G_p \geq 0$ are KKT multipliers. Setting the derivative of the preceding Lagrangian function (with respect to x_p and g_p) to zero, we obtain

$$\begin{aligned} &g_p((k+1)((k+1)\tilde{a}_p x_p^k + g_p^k \hat{a}_p x_p^k) + L_1) = Z_p, && \forall p, \\ &x_p((k+1)\tilde{a}_p x_p^k + (k+1) g_p^k \hat{a}_p x_p^k + L_1) + L_2 = G_p, && \forall p. \end{aligned} \tag{73}$$

For every p with $g_p > 0$, it follows from complementary slackness that $G_p = 0$; it can be easily shown that $x_p > 0$, and therefore $Z_p = 0$. Through a simple calculation, we obtain the following conditions that are equivalent to (73):

$$\begin{aligned} &(k+1)x_p^k((k+1)\tilde{a}_p + \hat{a}_p g_p^k) + L_1 = 0, && \forall p : g_p > 0, \\ &x_p((k+1)\tilde{a}_p x_p^k + (k+1) g_p^k \hat{a}_p x_p^k + L_1) + L_2 \geq 0, && \forall p : g_p = 0, \\ &k(k+1)x_p^k \tilde{a}_p = L_2, && \forall p : g_p > 0. \end{aligned} \tag{74}$$

We now consider another vector $\mathbf{x}'$ such that $x'_p = x_p$ for every p with $g_p > 0$, and for every p with $g_p = 0$, we let $k(k+1)(x'_p)^k \tilde{a}_p = L_2$. We note that $(\mathbf{g},\mathbf{x}')$ must also be an optimal solution to (68), and therefore must also satisfy the necessary conditions (74). It is easy to show that the optimal solution $(\mathbf{g},\mathbf{x}')$ satisfies the following conditions,

$$\begin{aligned} &(k+1)x_p^k((k+1)\tilde{a}_p + \hat{a}_p g_p^k) + L_1 = 0, && \forall p : g_p > 0, \\ &(k+1)x_p^k((k+1)\tilde{a}_p + \hat{a}_p g_p^k) + L_1 \geq 0, && \forall p : g_p = 0, \\ &k(k+1)x_p^k \tilde{a}_p = L_2, && \forall p : g_p > 0, \end{aligned} \tag{75}$$

where the second condition follows from the inequality (74) and the fact that $k(k+1)(x'_p)^k \tilde{a}_p = L_2$. The conditions in (75), together with the constraints that $\sum_p g_p = 1$ and $\sum_p x'_p g_p = \lambda$, are equivalent to the equilibrium conditions in (61)-(65). We have shown that there exists an optimal solution to (68) that satisfies the conditions in (61)-(65). Since there exists a unique vector $(\mathbf{g},\mathbf{x}')$ that satisfies conditions (61)-(65) (cf. the proof of Proposition 8), it follows that the vector that satisfies conditions (61)-(65) must be an optimal solution to (68).

Step 2: Price of anarchy bound.

Let $(\mathbf{g}^*,\mathbf{x}^*)$ be an optimal solution to the social planner's problem in (29), and $(\mathbf{g},\mathbf{x}')$ be a vector that satisfies conditions (61)-(65), respectively. We have

$$\begin{aligned}
\ell(\mathbf{g},\mathbf{x}') = & \sum_p g_p x'_p(\tilde{a}_p(x'_p)^k + \hat{a}_p(x'_p)^k g_p^k) \\
\leq & \sum_p g_p x'_p((k+1)\tilde{a}_p(x'_p)^k + \hat{a}_p(x'_p)^k g_p^k) \\
\leq & \sum_p g^*_p x^*_p((k+1)\tilde{a}_p(x^*_p)^k + \hat{a}_p(x^*_p g^*_p)^k) \\
\leq & \; (k+1)\sum_p g^*_p x^*_p(\tilde{a}_p(x^*_p)^k + \hat{a}_p(x^*_p g^*_p)^k) \\
= & \; (k+1)\ell(\mathbf{g}^*,\mathbf{x}^*),
\end{aligned}$$

where the second inequality is true because $(\mathbf{g},\mathbf{x}')$ is an optimal solution to the optimization problem (68).