Performance Evaluation 120 (2018) 36–48

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Provisioning of ad-supported cloud services: The role of competition☆

Jayakrishnan Nair [a,*], Vijay Subramanian [b], Adam Wierman [c]

[a] IIT Bombay, India
[b] University of Michigan, United States
[c] California Institute of Technology, United States

ARTICLE INFO

Article history:
Received 27 February 2017
Received in revised form 19 August 2017
Accepted 8 January 2018
Available online 16 January 2018

Keywords:
Capacity provisioning
Network effects
Competition
Congestion
Cloud services
Monopoly

ABSTRACT

Motivated by cloud services, we consider the interplay of network effects, congestion, and competition in ad-supported services. We study the strategic interactions between competing service providers and a user base, modeling congestion sensitivity and two forms of positive network effects: network effects that are either "firm-specific" or "industry-wide." Our analysis reveals that users are generally no better off due to the competition in a marketplace of ad-supported services. Further, our analysis highlights an important contrast between firm-specific and industry-wide network effects: Firms can coexist in a marketplace with industry-wide network effects, but near-monopolies tend to emerge in marketplaces with firm-specific network effects.

1. Introduction

Cloud based services are increasingly becoming the norm. While cloud-based email applications have been around for decades at this point, other cloud services are quickly replacing a wide variety of applications that used to be run locally, e.g., office applications (GoogleDocs, Office365) and even our hard drives (Dropbox, GoogleDrive, iCloud).

Competition between these emerging cloud services is extreme [3,4], e.g., competition between Hotmail, Gmail, and Yahoo mail, or the competition between GoogleDocs and Office365. In many cases, competition has driven the price of cloud services to zero (or a small nominal fee), leaving services to make profits primarily from advertising rather than subscription fees. Indeed, companies like Google and Facebook make billions of dollars annually from ad-supported online services [5].

Given the fact that many cloud services are offered for free, price does not provide a way for firms to compete to draw users. Instead, firms typically compete through a mixture of *performance* and *network effects*. In particular, users of online services are *highly delay sensitive*. Small additional delays for users can be traced to significant declines in revenue [6–8]. Further, cloud services have strong *positive network effects*, i.e., the experience of users in cloud services often is highly dependent on how many users the service has [9–11], e.g., collaborative document tools, online gaming environments, etc. Thus, cloud services compete to magnify network effects while providing good performance in order to grow more quickly than competitors.

The above highlights three key factors that lead to a complex cloud marketplace: ad-supported revenue models, delay sensitive customers, and positive network effects. Clearly, these three factors play a crucial role in determining the user experience, the optimization of service capacity provisioning, and the division of market share between competing services.

☆ Extended abstracts corresponding to this work appear in the proceedings of IFIP Performance 2014 [1] and Allerton 2014 [2].

* Corresponding author.
E-mail address: jayakrishnan.nair@ee.iitb.ac.in (J. Nair).

https://doi.org/10.1016/j.peva.2018.01.001
0166-5316/

Contributions of this paper

This paper studies the interplay of network effects, congestion, and competition in ad-supported services. The goal is to answer questions such as: Does competition lead to improved user experience in ad-supported services? How does competition impact service capacity provisioning in the cloud? Can competing firms coexist or will near-monopolies emerge in cloud marketplaces?

The last question can be rephrased as also determining the most natural market structure for cloud marketplaces. Historically, different cloud services have exhibited vastly different market structures: many competing firms coexisting (with the exact identities, ranks and market-shares changing over time) for services such as email versus near-monopolies for social networking services (e.g., MySpace initially, and Facebook at present). From this viewpoint, one of our goals is also to shed light, using a mathematical model, on some of the underlying reasons for the different market structures seen in different cloud services.

To address these questions, we introduce a new model in which to study competition between service providers, a.k.a., firms (see Section 2), in the cloud environment. A key new feature of the model is how network effects are considered in the competitive environment. In particular, we model network effects as either *firm-specific* or *industry-wide*. Firm-specific network effects capture settings where the utility of a user of a particular firm depends only on the population of users of that specific firm. This captures settings like Facebook, where a user's utility from joining Facebook grows as the number of people using Facebook grows. On the other hand, industry-wide network effects model situations where the utility of a user of a particular firm depends on the number of users across all the firms in the industry, not just the number of users at that specific firm. This captures applications such as email, where a user's utility grows with the number of people that use email, not just with the number of people that use the same email client. Of course, many applications have a combination of these two forms of network effects, but we focus on the extreme situations in this paper in order to contrast the impact most dramatically.

Given these two models of network effects, we study a situation where users from a user base decide strategically which, if any, of two services to join based on the congestion and network effects available at each service. Each of these is a function of the capacity decision of the profit-maximizing, competing firms.

Our analysis leads to a number of novel insights about the cloud marketplace, answering each question posed above.

First, our results provide insight into capacity management in competing cloud services. In particular, our results highlight that positive network effects allow firms to run fewer servers, and thus increase profits—at the expense of user performance. This parallels what has been observed in the case of a single firm in [12], but is the first such result in a setting that models competition.

Second, our results shed light on the interplay between competition and user experience. Contrasting the case of competing firms with the case of a single firm, one should expect the impact of positive network effects (discussed above) to diminish with increased competition; thus improving the user experience due to reduced congestion. Surprisingly, our results highlight that users are generally no better off due to competition between service providers. This is in contrast to results such as [13,14], which argue that increased competition among cloud services leads to improved user performance when paid services are considered. The contrast in our results stems from the ad-supported nature of the services we model.

Third, our results highlight that the form of network effects (firm-specific or industry-wide) has a significant impact on the market share of competing firms, and hence, the resulting market structure. Specifically, Theorems 1 and 2 highlight that firms can share the market in the case of industry-wide network effects; however Theorems 3 and 4 show that near monopolies tend to emerge under firm-specific network effects.

These results provide an analytic explanation for what can be informally observed in the cloud service marketplace: Facebook, which has firm-specific network effects, enjoys near-monopoly status while Gmail, Hotmail, and many other cloud-based email providers, which have industry-wide network effects, coexist. It also highlights that, in order to compete in areas where network effects are firm-specific, services must build a user base before entering the market. An example of this is Twitter, which, after building a large user base, has started to position itself as a competitor to Facebook.

Importantly, the messages described above seem to emerge *because* of the ad-supported nature of the services we consider. This contrast can be observed in the complementary nature of the results in the current paper when compared with those from the literature on club theory, which models users that pay for access [15,16].

Related work

The impact of network effects is crucial for cloud services. The more users there are on Facebook, the more appealing it is to be on Facebook. Similarly, the more users on GoogleDrive/Dropbox/iCloud, the more value a new user gets from joining. However, network effects are not specific to cloud services, and have been studied extensively in the economics and operations management literatures. Most of this literature focuses on settings where there is no congestion, e.g., [9,10,17–20]. The literature on "club theory" focuses on the interaction of network effects and congestion, and so is most related to the current paper. The theory of clubs, which originated from [15], deals with groups of congestion sensitive users sharing a certain resource. See [16] for a survey. The setting of cloud services can be interpreted as a club good offered by competing profit maximizing firms; however, throughout the literature on club goods it is assumed that the users *pay* for access, and the revenue of the provider is made up exclusively of such payments (see, for example, [21–23]). This is very different from the

situation in cloud services, where revenue predominantly comes from advertising rather than user payments. This difference turns out to have a significant impact on the behavior of profit maximizing firms. In particular, ad-supported services tend to operate in a highly congested state, since firms do not have the mechanism to monetize a reduction in congestion achieved by adding capacity.

The only previous piece of work to consider network effects and congestion in an ad-supported service is [12], which focuses on the capacity provisioning of a single service when faced with a strategic user population with positive network effects. In this setting, [12] shows that positive network effects mean that the user base is more tolerant of congestion, which allows the service provider to run the service with fewer servers and, thus, derive a larger profit. This effect is more extreme when the strategic behavior of the user base is 'cooperative' than if it is 'non-cooperative', but in both cases positive network effects lead to a worse user experience. However, [12] studies only the behavior of a single, monopolistic service. No prior work has considered the interplay of congestion and network effects in a setting where ad-supported firms are competing.

Finally, we note that our analysis focuses on the setting where the user base is large, i.e., scaling to infinity. Thus, it is related to the literature on scaling limits of queueing systems. However, most commonly in that literature, the traffic regime is imposed exogenously, e.g., [24–26], while the scaling emerges endogenously due to competition in our model.

2. Modeling competing cloud services

The goal of this paper is to study the interplay between network effects, congestion, and competition in ad-supported services using a representative, but tractable analytic model.

To maintain analytic tractability, we focus on the case of two firms offering competing cloud-based services to a common user base. We refer to the firms as Firm 1 and Firm 2. Each firm controls its capacity allocation strategically, with C_i denoting the capacity provisioned by Firm i for serving user requests.[1]

The key feature of our model is the interaction of strategic firms with users that are strategic, sensitive to congestion (which is impacted by the provisioning of the firms), and have positive network effects. Our approach is to model the interaction between the firms the user base using a leader–follower model. The firms are the 'leaders' in this interaction. Specifically, each firm decides how much capacity to provision seeking to maximize its profit. The user base 'follows' with the tuple of arrival rates (λ_1, λ_2), which corresponds to a Wardrop equilibrium, i.e., a non-cooperative equilibrium. We describe this in detail below.

2.1. User model

The user model prescribes the response of the user base to the capacity provisioning decision of the firms. In particular, it defines the arrival rate λ_i seen by each Firm i in response to the capacity provisioning decisions of the firms.

We assume that the aggregate arrival rate across firms is at most Λ, which is an exogenous parameter that can be interpreted as the 'size' of the user base. In other words, $\lambda_1 + \lambda_2 \leq \Lambda$. The user split (λ_1, λ_2) is determined by the latency experienced by users, as well as the utility derived from the each service. We first describe our latency model.

2.1.1. Latency model

We consider a model where a user of Firm i perceives a non-negative latency cost $f(\lambda_i, C_i)$ that is a function of both the arrival rate λ_i of users of Firm i, and the provisioned capacity C_i of the provider. We consider the following two classes of latency functions.

1. *M/M/1 latency:* Here, the latency cost is the average (stationary) response time in an M/M/1 queue, i.e.,
$$f(\lambda, C) = \begin{cases} \dfrac{1}{C-\lambda} & \text{if } \lambda < C \\ +\infty & \text{otherwise.} \end{cases}$$
2. *Load-based latencies:* Here, the latency cost is a general function of the load $\rho = \lambda/C$, i.e.,
$$f(\lambda, C) = g\left(\frac{\lambda}{C}\right)$$
for a strictly increasing, twice differentiable, strictly convex function g, defined over $[0, 1)$ such that $g(0) = 0$, and $\lim_{\rho \uparrow 1} g(\rho) = \infty$. An example of this class of latency functions is the average (stationary) number of jobs in an M/M/1 queue, where $g(\rho) = \frac{\rho}{1-\rho}$.

[1] Of course, moving beyond two firms is an interesting direction for future work. Given the nature of our results, we do not expect the results to change dramatically unless we consider the asymptotic setting of a perfectly competitive market, i.e., an infinite number of firms. In fact, in Section 3.2, we explicitly discuss how some of the results (and proofs) extend to the case of a large, but finite, number of firms.

Note that both the above models capture the impact of *congestion* on the latency experienced by a user. The use of the M/M/1 latency function is standard in the literature (see, for example, [13,22,27]). Our load-based latency models are novel to the best of our knowledge (though more general models have been considered for worst-case analysis in network economics; see, for example, [28]).

Given these latency models, we are now ready to describe the user behavior. We distinguish between two classes of network effects—firm specific, and industry-wide. If network effects are industry-wide, then the experience of a user is improved by having a larger aggregate population of users across all firms. Whereas if network effects are firm-specific, then the experience of a user is improved by having a large population of users at the same firm. A canonical example of industry-wide network effects is email, and a canonical example of firm-specific network effects is social networking.

2.1.2. User behavior under industry-wide network effects

When network effects are industry-wide, the utility of each user is a growing function of the aggregate usage $\lambda := \lambda_1 + \lambda_2$ across both firms. Specifically, we assume that the utility is of the form $V(\lambda) = w\lambda^\beta$, where $w > 0$ and $\beta \in [0, 1]$. Note that the network effect becomes stronger as β increases. $\beta = 0$ corresponds to no positive network effects, and $\beta = 1$ corresponds to very strong network effects.[2] Thus, users of Firm i see a net payoff of $V(\lambda) - f(\lambda_i, C_i)$, where it is only the latency cost that is different across providers. Whichever firm yields a higher payoff sees an increase in the number of subscribers as the users are driven to maximize their individual payoffs. Then, since the latency cost increases with the arrival rate, it is easy to argue that the net payoffs will equalize, if possible.

More formally, a traffic split $(\hat{\lambda}_1, \hat{\lambda}_2)$ such that $\hat{\lambda}_1 + \hat{\lambda}_2 = \hat{\lambda}$ is a Wardrop equilibrium if the following condition holds for $i = 1, 2$

$$\hat{\lambda}_i > 0 \Rightarrow V(\hat{\lambda}) - f(\hat{\lambda}_i, C_i) = \max_{j=1,2}\{V(\hat{\lambda}) - f(\hat{\lambda}_j, C_j)\}. \tag{1}$$

Note that $(\hat{\lambda}, 0)$ and $(0, \hat{\lambda})$ can also be Wardrop equilibria.

For a given $(\hat{\lambda}, C_1, C_2)$, where $\hat{\lambda} < C_1 + C_2$, it is easy to verify that the Wardrop equilibrium $(\hat{\lambda}_1(\hat{\lambda}, C_1, C_2), \hat{\lambda}_2(\hat{\lambda}, C_1, C_2))$ is the unique solution of the following convex optimization.

$$\begin{aligned} &\min_{(x_1,x_2)} \sum_{i=1}^{2} \int_0^{x_i} f(y, C_i)dy \\ &\text{subject to } \sum_{i=1}^{2} x_i = \hat{\lambda}. \end{aligned} \tag{2}$$

Note that the objective function above is strictly convex, since $f(\cdot, C)$ is strictly increasing.

We define the net arrival rate of the users as follows.

$$\begin{aligned} \lambda(C_1, C_2) := \max \Big\{ &\hat{\lambda} \in [0, \Lambda] \cap [0, C_1 + C_2) \mid \\ &\max_{j=1,2}\big\{V(\hat{\lambda}) - f(\hat{\lambda}_j(\hat{\lambda}, C_1, C_2), C_j)\big\} \geq 0\Big\}. \end{aligned} \tag{3}$$

The user behavior $(\lambda_1(C_1, C_2), \lambda_2(C_1, C_2))$ is then

$$\lambda_j(C_1, C_2) := \hat{\lambda}_j(\lambda(C_1, C_2), C_1, C_2),$$

for $j = 1, 2$. Note that we have suppressed the dependence of the user behavior on Λ for simplicity. Note also that, in the above definition, not only do we choose the highest possible arrival rate that yields non-negative payoff, but also the best possible feasible traffic split corresponding to each arrival rate.

2.1.3. User behavior under firm-specific network effects

When the network effects are firm-specific, user utility is distinct for each firm. Specifically, we consider the following utility model: users of Firm i obtain a utility $w_i > 0$. Thus, users of Firm i get a net payoff equal to $w_i - f(\lambda_i, C_i)$.

With this distinction, given $(\hat{\lambda}, C_1, C_2)$ such that $\hat{\lambda} < C_1 + C_2$, a split $(\hat{\lambda}_1, \hat{\lambda}_2)$ satisfying $\hat{\lambda}_1 + \hat{\lambda}_2 = \hat{\lambda}$ is a Wardrop equilibrium if

$$\hat{\lambda}_i > 0 \Rightarrow w_i - f(\hat{\lambda}_i, C_i) = \max_{j=1,2}\big\{w_j - f(\hat{\lambda}_j, C_j)\big\}. \tag{4}$$

Note that, if non-zero traffic is present for both firms, then the per-user payoff in both firms is the same at a Wardrop equilibrium. This follows for exactly the same reasons as in the industry-wide network effects model. By the same logic, $(0, \hat{\lambda})$

[2] It is important to note that we are describing *per-user* utility here. The net utility of the user base is then naturally captured by $\lambda V(\lambda) = w\lambda^{\beta+1}$. Note that the net utility grows linearly with usage when $\beta = 0$ and quadratically with usage when $\beta = 1$. In the literature, the case of $\beta = 1$ is referred to as Metcalfe's law [17], which has been argued to be an overestimation of user utility in practice [29]. Thus, the range $\beta \in [0, 1]$ suffices to capture real-world utility scaling due to positive network effects.

and $(\hat{\lambda}, 0)$ can also be Wardrop equilibria, so long as they satisfy (4). The Wardrop equilibrium $(\hat{\lambda}_1(\hat{\lambda}, C_1, C_2), \hat{\lambda}_2(\hat{\lambda}, C_1, C_2))$ is the unique solution of the following convex optimization.

$$\begin{aligned} &\min_{(x_1,x_2)} \sum_{i=1}^{2} \int_0^{x_i} (-w_i + f(y, C_i))dy \\ &\text{subject to } \sum_{i=1}^{2} x_i = \hat{\lambda}. \end{aligned} \tag{5}$$

Note that the objective function above is strictly convex, implying the solution is unique.[3] Having defined the Wardrop split, we can now define the user behavior model as before. We define the net arrival rate of the users as follows. We define the net arrival rate of the users as follows.

$$\begin{aligned} \lambda(C_1, C_2) := \max \Big\{ &\hat{\lambda} \in [0, \Lambda] \cap [0, C_1 + C_2) \mid \\ &\max_{j=1,2} \big\{ w_j - f(\hat{\lambda}_j(\hat{\lambda}, C_1, C_2), C_j) \big\} \geq 0 \Big\}. \end{aligned} \tag{6}$$

The user behavior $(\lambda_1(C_1, C_2), \lambda_2(C_1, C_2))$ is then

$$\lambda_j(C_1, C_2) := \hat{\lambda}_j(\lambda(C_1, C_2), C_1, C_2),$$

for $j = 1, 2$. Note that as in the previous section, we choose the highest possible arrival rate that yields non-negative payoff, using a Wardrop traffic split corresponding to each arrival rate.

2.2. Modeling competing firms

Given the user model, we can now describe the competition between the strategic firms. Recall that the firms 'lead' the user base in our leader–follower interaction.

We assume that Firm i makes a revenue $b_i\lambda_i$ dollars from advertising, and incurs a cost of one dollar per unit of capacity provisioned. Thus, the profit of Firm i is given by $b_i\lambda_i - C_i$. Given that $\lambda_i < C_i$, we require that $b_i > 1$ to enable a positive profit. Since λ_i is a function of the capacities (C_1, C_2) of both firms, the decisions of the two firms get coupled. We thus study this via a game formulation, and consider Nash equilibria between the firms.

To be precise, we define $(\lambda_1, \lambda_2, C_1, C_2)$ to be an equilibrium of our system if (λ_1, λ_2) is the response of the user base to the service capacities (C_1, C_2), and (C_1, C_2) is a Nash equilibrium between the providers, i.e.,

$$\begin{aligned} C_1 &\in \operatorname*{argmax}_{c \geq 0} b_1\lambda_1(c, C_2) - c \\ C_2 &\in \operatorname*{argmax}_{c \geq 0} b_2\lambda_2(C_1, c) - c. \end{aligned}$$

Given this definition, our goal in the remainder of the paper is to study the equilibria that can emerge among firms when the market size Λ is large. Note that we consider only pure Nash equilibria, since mixed equilibria are not meaningful in our setting.

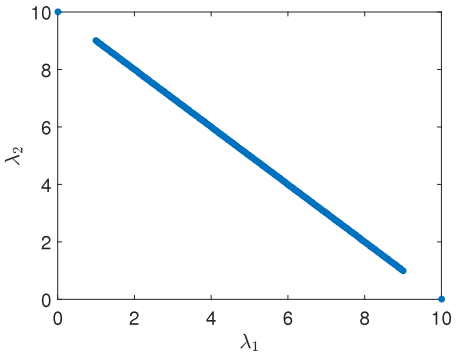
3. Results

The goal of this paper is to understand the impact of the interaction between network effects, congestion, and competition in ad-supported services. The results presented in this section contrast two forms of network effects—industry-wide and firm-specific. We begin with the case of industry-wide network effects in Section 3.1 and then present results for firm-specific network effects in Section 3.2. Following the presentation of the main results we discuss the contrast between the results for the two models in Section 4.

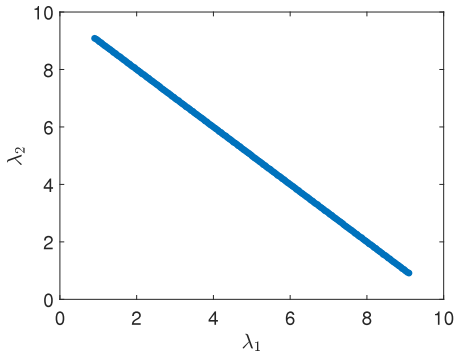
3.1. Industry-wide network effects

We begin with the case of industry-wide network effects, which capture situations where the utility of a user of a particular firm depends on the number of users across all the firms in the industry, not just the number of users at the specific firm. Email is a canonical example of an application with industry-wide network effects.

[3] From a practical standpoint, it is meaningful to also consider positive network effects when user utilities are firm-specific. The reason we are unable to handle utility functions of the form $w_i\lambda_i^{\beta_i}$ (with $\beta_i > 0$) in our analysis of firm-specific network effects is that this would make the optimization that defines the Wardrop equilibrium non-convex. Indeed, it can be shown that multiple Wardrop equilibria are possible in this case, making analytical treatment challenging.



(a) Equilibrium market shares under the M/M/1 latency function, with $\Lambda = 10$, $b_1 = b_2 = 2$, $V(\lambda) \equiv 1$.



(b) Equilibrium market shares under the load based latency function, with $g(\rho) = \frac{\rho}{1-\rho}$, $\Lambda = 10$, $b_1 = b_2 = 2.2$, $V(\lambda) \equiv 1$.

Fig. 1. Illustration of the continuum of equilibria under industry-side network effects.

The basic message of the results in this section is that the competing firms can co-exist in the market if network effects are industry-wide. This is consistent with what we informally observe in the case of cloud email services—there is a competitive market with many players.

In the following, we provide results for both the case of M/M/1 latency functions and general load-dependent latency functions. We start with the M/M/1 case.

Theorem 1. *Consider a non-cooperative user base with industry-wide network effects, and take f to be the M/M/1 latency function. For large enough Λ, the following statements hold.*

1. *If $b_1, b_2 > 2$, then no equilibrium exists.*
2. *If $b_1 > 2$, $b_2 \leq 2$, then the only equilibrium is a full monopoly of Provider 1:*

$$\lambda_1 = \Lambda, C_1 = \Lambda + \frac{1}{V(\Lambda)}, \lambda_2 = C_2 = 0.$$

3. *If $b_1, b_2 \in (1, 2]$, then a continuum of equilibria exist, including monopoly configurations. Moreover, any equilibrium is of one of the following forms.*
   (a) *Monopoly for Firm 1:* $\lambda_1 = \Lambda,\ C_1 = \Lambda + \frac{1}{V(\Lambda)},\ \lambda_2 = C_2 = 0$
   (b) *Monopoly for Firm 2:* $\lambda_2 = \Lambda,\ C_2 = \Lambda + \frac{1}{V(\Lambda)},\ \lambda_1 = C_1 = 0$
   (c) *Firms 1 and 2 share the market such that $\lambda_1 + \lambda_2 = \Lambda$, and*

$$\lambda_i \geq \frac{1}{(b_i - 1)V(\Lambda)},$$
$$C_i = \lambda_i + \frac{1}{V(\Lambda)},$$

for $i = 1, 2$.

Interestingly, the theorem highlights that there are important contrasts depending on the relative advertising efficiencies of the firms. In Case 1, the firms are both extremely efficient. In contrast, Case 2 corresponds to the case when the firms have differing advertising efficiencies, with one being extremely efficient ($b_1 > 2$). In this case the more profitable firm is able to drive the competitor out of the market. Finally, in Case 3, the two firms have differing, but intermediate, advertising efficiencies, which results in a multitude of possible ways for the firms to divide the market, as illustrated in Fig. 1(a).

Additionally, it is important to note that in every equilibrium configuration demonstrated in Theorem 1, the user base sees exactly the same congestion (measured, for the M/M/1 latency function, in terms of the spare capacity $C_i - \lambda_i$) as in the case of a single provider (see Theorem 2 in [12]). Thus, competition in the marketplace does not improve the payoff experienced by the user base.

Proof of Theorem 1. We begin with the following preliminary observations that will be used repeatedly in the proof. Firstly, given that one of the firms provisions zero capacity, then the optimal provisioning for the other firm equals $\Lambda + \frac{1}{V(\Lambda)}$. This is easy to prove directly, and also follows from the analysis of the monopolistic setting in [12]. Next, we note that if the capacity provisioning (C_1, C_2) results in a Wardrop split $\lambda_1, \lambda_2 > 0$, then $C_1 - \lambda_1 = C_2 - \lambda_2$, i.e., the spare capacity across

the firms is matched (under the M/M/1 latency model). We can now also deduce the impact of increasing C_1 by small $\epsilon > 0$. If $\lambda_1 + \lambda_2 = \Lambda$, then this perturbation would result in λ_1 increasing by $\epsilon/2$ and λ_2 decreasing by $\epsilon/2$. On the other hand, if $\lambda_1 + \lambda_2 < \Lambda$, then this perturbation would result in λ_1 increasing by at least ϵ.

We now prove the three claims in the theorem in turn.

Claim 1. *Suppose that $(\lambda_1, \lambda_2, C_1, C_2)$ is an equilibrium. Note that it not possible that $C_1 = C_2 = 0$, since in such a configuration, Firm i has an incentive to change its action to $C_i = \Lambda + \frac{1}{V(\Lambda)}$ as discussed before.*

Consider then the possibility of an equilibrium satisfying $C_1 > 0$, $C_2 = 0$. Such an equilibrium must necessarily have $C_1 = \Lambda + \frac{1}{V(\Lambda)}$. However, this cannot be an equilibrium, since if Firm 2 is to increase its provisioning to match that of Firm 1, we would have $\lambda_1 = \lambda_2 = \Lambda/2$, resulting in a profit of $(\frac{b_2}{2} - 1)\Lambda - \frac{1}{V(\Lambda)}$ for Firm 2, which is positive for large enough Λ (since $b_2 > 2$). Thus, we conclude that an equilibrium with only one firm operating does not exist.

Finally, we consider the possibility of an equilibrium satisfying $C_1, C_2 > 0$. Such an equilibrium must necessarily satisfy $\lambda_1, \lambda_2 > 0$, since any provider with zero arrival rate would be better off setting its capacity to zero. Consider now the effect of Firm 1 increasing its capacity by a small ϵ. As discussed above, Firm 1 would receive an additional arrival rate of at least $\epsilon/2$ as a result of this change. Since $b_1 > 2$, this change would be profitable to Firm 1, implying that the proposed configuration is not an equilibrium.

Claim 2. *We first argue that the proposed monopoly configuration is an equilibrium. Note that the user response is consistent with the capacities provisioned, since the users see a single provider. Similarly, from our discussion of the single provider case, it is clear that the provisioning of Firm 1 is optimal, given that Firm 2 does not offer the service. Finally, note that if Firm 2 is to increase its capacity to c, the Wardrop split implies that she will receive an arrival rate of at most $c/2$, which is not profitable, since $b_2 \leq 2$.*

Next, we rule out the possibility of an equilibrium with $C_2 > 0$. Clearly, such a equilibrium would have $\lambda_2 > 0$. If $\lambda_1 > 0$, then it is easy to see that a capacity increase of ϵ by Firm 1 would increase its arrival rate by at least $\epsilon/2$, increasing its profit. On the other hand, if $\lambda_1 = 0$, the equilibrium must satisfy $C_2 = \Lambda + \frac{1}{V(\Lambda)}$. But Firm 1 could now just match the capacity of Firm 1, share the market equally, making a profit of $(\frac{b_1}{2} - 1)\Lambda - \frac{1}{V(\Lambda)}$, which is positive for large enough Λ.

Claim 3. *We first prove that the claimed configurations are in fact equilibria. The proof that the two monopoly configurations are equilibria follows along the same lines as the proof of Claim 2 . We now show that configurations $(\lambda_1, \lambda_2, C_1, C_2)$ satisfying the conditions in Part (c) are equilibria. It is easy to see that user behavior is consistent with our model. For Provider i, consider the capacity provisioning $C_i = \lambda_i + \frac{1}{V(\Lambda)}$. Note that the lower bound on λ_i implies the provider makes a non-negative profit. It is not profitable for any provider to increase capacity further, since any increase epsilon in capacity would lead to an increase of $\epsilon/2$ in the arrival rate; this is not favorable given $b_i \leq 2$. Consider next the action of Provider 1 decreasing capacity by ϵ, leading to a Wardrop split (λ_1', λ_2'). Clearly, if $\lambda_1' = 0$, the action is unfavorable. Else, from the Wardrop condition, we must have*

$$(C_1 - \epsilon) - \lambda_1' \geq \frac{1}{V(\lambda')} \geq \frac{1}{V(\Lambda)},$$

where $\lambda' = \lambda_1' + \lambda_2'$. The above inequalities imply that $\lambda_1' \leq \lambda_1 - \epsilon$, which implies a decrease in profit for Provider 1. We conclude that neither provider has an incentive to adapt capacity, implying our configuration is an equilibrium.

Finally, we have to show that any equilibrium $(\lambda_1, \lambda_2, C_1, C_2)$ satisfying $\lambda_1, \lambda_2 > 0$ must be of the form postulated. First, we argue that it must hold that $\lambda := \lambda_1 + \lambda_2 = \Lambda$. Indeed, if $\lambda < \Lambda$, it is easy to see that a capacity increase of ϵ by any provider implies an arrival rate increase of at least ϵ, which is profitable. Now, given that $\lambda = \Lambda$, Provider i must clearly provision a capacity at-least $C_i \geq \lambda_i + \frac{1}{V(\Lambda)}$. Also, we must have $C_1 - \lambda_1 = C_2 - \lambda_2 =: s$. If $s > \frac{1}{V(\Lambda)}$, then it easy to see that both providers have an incentive to decrease capacity. □

The second class of latency functions we consider are load based latencies, for which $f(\lambda, C) = g(\lambda/C)$. Define $h(\lambda) := g^{-1}(w\lambda^{\beta})$. Thus, $h(\lambda) < 1$, and $\lim_{\lambda\to\infty} h(\lambda) = 1$ for $\beta > 0$.

Theorem 2. *Consider a non-cooperative user base with industry-wide network effects, and take f to be a load based latency function. If the per unit rewards (b_1, b_2) are such that*

$$b_1, b_2 > \frac{1}{h(\Lambda)} \text{ and } \frac{1}{b_1} + \frac{1}{b_2} \geq h(\Lambda),$$

then a continuum of equilibria $(\lambda_1, \lambda_2, C_1, C_2)$ exist, characterized by the conditions

$$C_1 + C_2 = \frac{\Lambda}{h(\Lambda)},$$
$$\lambda_1 = \Lambda\frac{C_1}{C_1 + C_2},\ \lambda_2 = \Lambda\frac{C_2}{C_1 + C_2},$$
$$1 - \frac{1}{b_1 h(\Lambda)} \leq \frac{C_1}{C_1 + C_2} \leq \frac{1}{b_2 h(\Lambda)}.$$

Once again, we note that in every equilibrium configuration demonstrated in Theorem 2, the congestion experienced by the user base (measured by the load, under the load based latency functions) is exactly the same as in the case of a single provider (this can be proved along the lines of Theorem 2 in [12]). Further, like the previous theorem, this result highlights that, if the advertising efficiencies of the two providers are not too large, then they can co-exist with each gaining a significant share of the market, as illustrated in Fig. 1(b). Interestingly, though the overall message matches that of Theorem 1, the specific equilibria that emerge are quite different.

Further, note that, if $\beta > 0$, then for any b_1, b_2 such that $\frac{1}{b_1} + \frac{1}{b_2} > 1$, using the property that $\lim_{\lambda\to\infty} h(\lambda) = 1$, we can find Λ large enough such that $\min(b_1, b_2) > \frac{1}{h(\lambda)}$ for every $\lambda \geq \Lambda$. Therefore, $\frac{1}{b_1} + \frac{1}{b_2} > 1$ is sufficient for Theorem 2 to hold for sufficiently large Λ.

Proof of Theorem 2. Consider a tuple $(\lambda_1, \lambda_2, C_1, C_2)$ satisfying the conditions of the theorem. It is easy to see that the tuple is a Wardrop equilibrium for the users, since the utilization with both providers equals $h(\Lambda)$.

Next, we argue that Firm 1 has no incentive to increase her capacity. Suppose that Firm 1 increases its capacity by δ, and let us denote the new Wardrop split by $(\lambda_1 + \epsilon, \lambda_2 - \epsilon)$ (it is easy to see that the total arrival rate remains Λ). So,

$$\frac{\lambda_1 + \epsilon}{C_1 + \delta} = \frac{\lambda_2 - \epsilon}{C_2}.$$

Noting that $\frac{\lambda_1}{C_1} = \frac{\lambda_2}{C_2}$, it follows that

$$\epsilon = \frac{\lambda_2 \delta}{C_1 + C_2 + \delta} \leq \frac{\lambda_2 \delta}{C_1 + C_2}.$$

The change in profit of Firm 1 equals

$$\begin{aligned} b_1\epsilon - \delta &\leq \delta\left(\frac{b_1\lambda_2}{C_1 + C_2} - 1\right) \\ &= \delta\left(\frac{b_1 h(\Lambda) C_2}{C_1 + C_2} - 1\right) \\ &\leq 0, \end{aligned}$$

where the last inequality follows from our restriction on the value of $\frac{C_1}{C_1+C_2}$. This proves that Firm 1 has no incentive to increase its capacity.

Next, we show that Firm 1 has no incentive to decrease her capacity. Suppose that Firm 1 decreases its capacity to $C_1' < C_1$, leading to a new arrival rate vector (λ_1', λ_2'). Let $\rho_1' = \frac{\lambda_1'}{C_1'}$ denote the new utilization of Firm 1. The non-negativity of the user utility for Firm 1 implies that $V(\lambda_1' + \lambda_2') \geq g(\rho_1')$, implying that $\rho_1' \leq h(\lambda_1' + \lambda_2') \leq h(\Lambda)$. The new profit of Firm 1 now equals

$$\begin{aligned} b_1\lambda_1' - C_1' &= C_1'(b_1\rho_1' - 1) \\ &\leq C_1(b_1 h(\Lambda) - 1) = b_1\lambda_1 - C_1. \end{aligned}$$

Thus, Firm 1 has no incentive to decrease its capacity.

Symmetric arguments apply to Firm 2. □

3.2. Firm-specific network effects

We now move to the case of firm-specific network effects, which capture settings where the utility of a user of a particular firm depends only on the population of users of that specific firm. Social networks, e.g., Facebook, are a canonical example of a service with firm-specific network effects.

In our model, firm-specific network effects are captured by setting $w_1 \neq w_2$, since $w_1 = w_2$, recovers the case of industry-wide network effects. Without loss of generality, assume that $w_1 > w_2$.

In the following, we provide results for both the case of M/M/1 latency functions and general load-dependent latency functions. We start with the M/M/1 case for simplicity.

The following theorem shows that, for the M/M/1 latency function, any equilibrium is necessarily a near-monopoly for Firm 1. That is, Firm 2 can never gather more than a bounded arrival rate, and thus a negligible fraction of the user population—regardless of how close w_1 is to w_2.[4]

Theorem 3. *Consider a non-cooperative user base with firm-specific network effects such that $w_1 > w_2 > 0$. Further, take f to be the M/M/1 latency function. For large enough Λ, any equilibrium $(\lambda_1, \lambda_2, C_1, C_2)$ must satisfy*

$$\lambda_1 \geq \Lambda + \frac{1}{w_1} - \frac{b_1}{(b_1 - 1)(w_1 - w_2)}.$$

[4] We do not address here the issue of existence of an equilibrium.

While the above theorem does not provide an exact characterization of the congestion experienced by users at equilibrium (assuming one exists), the proof that follows shows that, at any equilibrium $(\lambda_1, \lambda_2, C_1, C_2)$, the arrivals into Firm 1 will only see a bounded spare capacity $C_1 - \lambda_1$.

To see this, note that the profit of Firm 1 equals

$$(b_1 - 1)\lambda_1 - (C_1 - \lambda_1) \leq (b_1 - 1)\Lambda - (C_1 - \lambda_1).$$

Combining this with the lower bound on the provider's profit from the proof below, we see that

$$C_1 - \lambda_1 \leq \frac{b_1 - 1}{w_1} - \frac{b_1}{(w_1 - w_2)} = \frac{b_1 w_2}{w_1(w_1 - w_2)} + \frac{1}{w_1}.$$

This suggests that the congestion experienced by the user base at equilibrium is of the same order as that in the case of a single firm. In other words, competition does not significantly improve the payoff of the user base.

Proof of Theorem 3. Suppose that $(\lambda_1, \lambda_2, C_1, C_2)$ is an equilibrium. We will show that

$$\text{Profit of Firm 1} \geq (b_1 - 1)\left(\Lambda + \frac{1}{w_1}\right) - \frac{b_1}{w_1 - w_2}.$$

This implies the statement of the theorem, since

$$(b_1 - 1)\lambda_1 \geq \text{Profit of Firm 1}.$$

For any C_2, consider the response by Firm 1 setting capacity to $\tilde{C}_1 = \Lambda + \frac{1}{w_1}$. Now, if $C_2 \leq \frac{1}{w_2}$, then the population split is such that $\tilde{\lambda}_1 = \Lambda$ arrives at firm 1, and our claim on the profit follows easily. If $C_2 > \frac{1}{w_2}$, then the population split is determined by

$$w_1 - \frac{1}{\tilde{C}_1 - \tilde{\lambda}_1} = w_2 - \frac{1}{C_2 - \tilde{\lambda}_2}.$$

Define the 'spare capacities' $s_1 = \tilde{C}_1 - \tilde{\lambda}_1$, $s_2 = C_2 - \tilde{\lambda}_2$. Note that the spare capacities uniquely determine the population split. We have

$$\frac{1}{s_1} - \frac{1}{s_2} = \Delta =: w_1 - w_2,$$
$$s_1 + s_2 = s =: C_2 + \frac{1}{w_1},$$

along with $s_1 \geq \frac{1}{w_1}$, $s_2 \geq \frac{1}{w_2}$. It is easily seen that

$$\frac{1}{s_1} = \frac{1}{s_2} + \Delta \Leftrightarrow s_1 = \frac{1}{\frac{1}{s_2} + \Delta} \leq \frac{1}{\Delta}.$$

Accordingly, we obtain

$$\tilde{\lambda}_1 = \tilde{C}_1 - s_1 \geq \Lambda + \frac{1}{w_1} - \frac{1}{\Delta}.$$

It follows that under the response of Firm 1, its profit is at least equal to

$$b_1\left(\Lambda + \frac{1}{w_1} - \frac{1}{\Delta}\right) - \left(\Lambda + \frac{1}{w_1}\right)$$
$$= (b_1 - 1)\left(\Lambda + \frac{1}{w_1}\right) - \frac{b_1}{\Delta}.$$

We therefore conclude that the profit of Firm 1 at an equilibrium can only be greater, and the result follows. □

The result above is presented for the case of two firms, but it easily generalizes to more than two firms. Suppose that Firm 1 has the highest utility. Using the same ideas as above, it can be shown that under any equilibrium, the profit of Firm 1 is at least $(b_1 - 1)\left(\Lambda + \frac{1}{w_1}\right) - \frac{b_1}{\Delta}$, where $\Delta := w_1 - \max_{i \neq 1} w_i$, from which the result would then follow.

Next, we investigate the case of load-based latency functions. For ease of exposition we define

$$\gamma_{nc} := g^{-1}(w_1 - w_2). \tag{7}$$

Theorem 4. *Consider a non-cooperative user base with firm-specific network effects such that $w_1 > w_2 > 0$. Further, take f to be a load-based latency function. If $b_1 > \frac{1}{\gamma_{nc}}$ with γ_{nc} given by (7), for large enough Λ, any equilibrium $(\lambda_1, \lambda_2, C_1, C_2)$ must*

satisfy

$$\lambda_1 \geq \frac{b_1 - \frac{1}{\gamma_{nc}}}{b_1 - 1}\Lambda \text{ and } \lambda_2 \leq \frac{\frac{1}{\gamma_{nc}} - 1}{b_1 - 1}\Lambda.$$

The conclusions that we can draw from the result are weaker than those in Theorem 3 for M/M/1 latency functions. For example, we only obtain an upper bound on the traffic choosing Firm 2 that is not enough to determine whether or not we have a virtual monopoly for Firm 1. We believe that this is because, under load-based latencies, firms do not operate in heavy traffic. Each firm has to operate at a utilization that is bounded from above (strictly smaller than 1) so that users subscribe to the service (i.e., obtain non-negative payoff), and also bounded from below (strictly greater than 0) to make a positive profit. Therefore, the resulting equilibrium would depend delicately on the balance between (w_1, w_2) (user utility) and (b_1, b_2) (revenue rate).

Nevertheless, Theorem 4 does provide interesting qualitative conclusions. First, Theorem 4 highlights that the better firm, i.e., Firm 1, has a guaranteed market-share. If we further insist that $b_1 > \frac{2}{\gamma_{nc}} - 1$, then it is easily seen that $\frac{\lambda_1}{\Lambda} \geq 0.5$, so that Firm 1 is guaranteed a majority market-share. Second, Theorem 4 shows that, in order to achieve positive profit, the lesser firm, i.e., Firm 2, must choose its capacity such that

$$C_2 \leq b_2\lambda \leq b_2 \frac{\frac{1}{\gamma_{nc}} - 1}{b_1 - 1}\Lambda.$$

Proof of Theorem 4. Consider the action $\tilde{C}_1 = \Lambda/\gamma_{nc}$ by Firm 1. It is easy to see that for any action C_2 of Firm 2, and any traffic split $\lambda_1, \lambda_2 > 0$, with $\lambda_1 + \lambda_2 \leq \Lambda$, $w_1 - f(\lambda_1, \tilde{C}_1) > w_2 - f(\lambda_2, C_2)$. It follows that the action $\tilde{C}_1$ would result in Firm 1 getting all user traffic, leading to a profit of

$$b_1\Lambda - \tilde{C}_1 = (b_1 - 1/\gamma_{nc})\Lambda > 0.$$

Thus, we conclude that under any equilibrium $(\lambda_1, \lambda_2, C_1, C_2)$,

$$(b_1 - 1)\lambda_1 \geq \text{Profit of Firm 1} \geq (b_1 - 1/\gamma_{nc})\Lambda,$$

which yields the statement of the theorem. □

This result also generalizes to more than two firms. Assuming n firms, with Firm 1 having the highest utility, it is not hard to show that the lower bound on λ_1 in the statement of Theorem 4 holds, with $\gamma_{nc} := g^{-1}(w_1 - \max_{i\neq 1} w_i)$.

4. Discussion

Perhaps the most striking conclusion from the results in the previous sections is about the impact of competition on the user experience. Regardless of the form of the network effects, the message that emerges is that *competition does not improve the user experience*. In particular, even with competition, the firms provision the same order of magnitude of capacity as does a single monopolistic firm, which can exploit positive network effects to increase profit by taking advantage of the fact that the network effects make users more tolerant of congestion.

One might suggest that this is the result of "anarchy", i.e. non-cooperative competition among the users. However, parallel results (summarized in Appendix) highlight that competition does not help even when the user base is cooperative. Thus, an important takeaway from the results is that adding competition does not help reduce congestion if firms cannot compete on prices. This is in contrast to price competition, e.g., [13,14], where competition reduces congestion.

Another striking message that emerges from the results in Sections 3.1 and 3.2 is about the impact of competition on market structure, and its interplay with network effects. In particular, the structure of the market, i.e., the market shares of the firms, is highly dependent on the type of service, in particular, the network effects of the service. If there are no network effects, or if the network effects are industry-wide, then multiple firms can co-exist, sharing the market (unless there is considerable asymmetry in the advertising efficiency of the firms). In other words, in such situations it is hard for firms to grab market share from each other, and the distinction between the firms disappears from user's point of view. However, if network effects are firm-specific, then near-monopolies tend to emerge. In fact, even minor differences in utilities are sufficient to ensure a monopoly. Moreover, the firm that obtains a near-monopoly is the one with the 'better' service, i.e., greater network effect.

Surprisingly, advertising efficiency does not change the story above, except for the load-based latencies. Thus, for example, the advertising efficiency of Google may seem to have given an advantage to GooglePlus over Facebook, but the results here suggest that the impact of this is outweighed by the service quality comparison. This aligns qualitatively with what has emerged in the case of competition between GooglePlus and Facebook. Again, while the results about the contrast between firm-specific and network-wide network effects in Sections 3.1 and 3.2 are focused on a non-cooperative user base. These results are not a consequence of "anarchy" among users—parallel results hold in the case of a cooperative user model, again see Appendix.

We also note that many of the results in Section 3.2 do not depend on having two firms, and easily extend to a finite number of firms.

We conclude by pointing out some interesting avenues for future work. For analytical tractability, the model in the present paper ignores two key features of real-world cloud services:

(i) User stickiness: In practice, users get locked-in to a certain service provider and do not switch to a competing provider unless the competitor offers a considerably higher utility. This has been modeled in the literature in different ways, including via the logistic function (see, for example, [30]) and via the Hotelling model [31].

(ii) Geographical effects: Network latency plays an important role in the overall latency experienced by users. The network latency in turn depends on the location of the service infrastructure (most cloud providers operate geographically distributed service infrastructures) as well as user location. The impact of location on user latency is captured in several papers, including [32].

Investigating the interplay between network effects and competition by incorporating the above features into our model is a promising avenue for future work.

Acknowledgements

Jayakrishnan Nair acknowledges the support of DST and CEFIPRA. Vijay Subramanian would like to acknowledge support from NSF via grants AST-1343381, AST-1516075, IIS-1538827 and ECCS-1608361. Adam Wierman would also like to acknowledge support from the NSF via grants CNS-1518941 and CNS-1319820.

Appendix. Cooperative user behavior

The main body of the paper considers a setting where users of the cloud services are non-cooperative. In order to highlight that the qualitative conclusions of the paper are not a result of "anarchy" in the user base induced by this modeling assumption, we consider a model of cooperative user behavior in this section. Proofs are omitted due to space constraints.

A.1. Industry-wide network effects

In the cooperative scenario with industry-wide network effects, the traffic split of users is determined by maximizing the collective payoff, which is the (aggregate) utility $U(\cdot)$ (where $U(\lambda) = \lambda V(\lambda)$) minus the congestion cost given by congestion function $f(\cdot, \cdot)$. That is, the user base solves the following optimization problem.

$$\begin{aligned} &\max \quad U(\lambda) - \sum_{i=1}^{2} \lambda_i f(\lambda_i, C_i) \\ &\text{subject to} \sum_{i=1}^{2} \lambda_i = \lambda, \quad \lambda \leq \Lambda, \quad \lambda_1, \lambda_2 \geq 0. \end{aligned} \tag{A.1}$$

If multiple solutions exist, the one with the largest value of net arrival rate λ is chosen. Note that given λ, the split (λ_1, λ_2) is the solution of the following optimization.

$$\begin{aligned} &\min_{\lambda_1, \lambda_2} \sum_{i=1}^{2} \lambda_i f(\lambda_i, C_i) \\ &\text{subject to} \sum_{i=1}^{2} \lambda_i = \lambda, \quad \lambda_1, \lambda_2 \geq 0. \end{aligned} \tag{A.2}$$

Since the objective function is strictly convex, the split (λ_1, λ_2) is unique.

The key message of the result that follows parallels that in the non-cooperative setting: multiple competing firms can coexist when network effects are industry-wide. Below we state our result for the case of load-based latency functions for the class of utility functions $U(x) = wx^{1+\beta}$ with $w, \beta > 0$. Define $h_c(\lambda) = l^{-1}(w(1+\beta)\lambda^{\beta})$ (we suppress parameters w and β), where $l(x) = xg'(x) + g(x)$, and $\lim_{\lambda \to \infty} h_c(\lambda) = 1$.

Theorem 5. *Consider a cooperative user base with industry-wide network effects, and take $f(x, C)$ to be a load based latency function determined by function $g(x/C)$. If the per unit rewards (b_1, b_2) are such that*

$$b_1, b_2 > \frac{1}{h_c(\Lambda)} \text{ and } \tfrac{1}{b_1} + \tfrac{1}{b_2} \geq h_c(\Lambda),$$

then a continuum of equilibria $(\lambda_1, \lambda_2, C_1, C_2)$ *exists, characterized by the conditions*

$$C_1 + C_2 = \frac{\Lambda}{h_c(\Lambda)},$$
$$\lambda_1 = \Lambda\frac{C_1}{C_1 + C_2},\ \lambda_2 = \Lambda\frac{C_2}{C_1 + C_2},$$
$$1 - \frac{1}{b_1 h_c(\Lambda)} \le \frac{C_1}{C_1 + C_2} \le \frac{1}{b_2 h_c(\Lambda)}.$$

It is important to realize that, even though the equilibria in the cooperative case are structurally similar to those in the non-cooperative setting, the actual scaling is different owing to the different $h_c(\lambda)$ function. Also, note that, like in the non-cooperative case, if $\beta > 0$, $\frac{1}{b_1} + \frac{1}{b_2} > 1$ is sufficient for the theorem to hold for large enough Λ. The most important thing to note is that, despite the competition, in every equilibrium, the user-base experiences the same level of congestion as in the single-firm case [12].

A.2. Firm-specific network effects

In the cooperative scenario with firm-specific network effects the traffic split of users is chosen to maximize the collective payoff. Here we let the users that subscribe to Firm $i = 1, 2$ receive a collective payoff of $U_i(\lambda_i) - \lambda_i f(\lambda_i, C_i)$, where $U_i(\cdot)$ is the (collective) utility achieved by the users that subscribe to service from Firm i. Then the user behavior $[\lambda_1(C_1, C_2), \lambda_2(C_1, C_2)]$ is defined to be the solution of the following optimization problem.

$$\max_{\lambda_1,\lambda_2} \sum_{i=1}^{2} [U_i(\lambda_i) - \lambda_i f(\lambda_i, C_i)]$$
$$\text{subject to } \sum_{i=1}^{2} \lambda_i \le \Lambda, \quad \lambda_i \ge 0,\ i = 1, 2.$$

As before, we consider the case of linear utilities ($\beta_1 = \beta_2 = 0$), i.e., $U_i(\lambda_i) = w_i\lambda_i$ with $w_i > 0$ for $i = 1, 2$. Under this assumption, since the objective function is strictly concave, the split (λ_1, λ_2) is unique. We also assume $w_1 > w_2$, without loss of generality.

We start with the case of M/M/1 latency functions. As in the non-cooperative case, the following theorem shows that, for the response time latency function, any equilibrium is necessarily a near-monopoly for Firm 1. That is, Firm 2 can never gather an arrival rate of more than $O(\sqrt{\Lambda})$ as the market size Λ grows large. Thus, we see that whether the user-base behaves cooperatively or non-cooperatively, a near-monopoly for the 'better' firm emerges in the firm-specific network effects model. Furthermore, the congestion seen by the user-base is the same as with a single firm [12].

Theorem 6. *Consider a cooperative user base with firm-specific network effects such that* $\beta_1 = \beta_2 = 0$ *and* $w_1 > w_2 > 0$. *Further, take* f *to be the M/M/1 latency function. As* Λ *becomes large, any equilibrium* $(\lambda_1, \lambda_2, C_1, C_2)$ *(if it exists) must satisfy*

$$\lambda_1 \ge \Lambda - \sqrt{\Lambda}\left(\frac{b_1}{(b_1 - 1)\sqrt{w_1 - w_2}} - \frac{1}{\sqrt{w_1}}\right) + o(\sqrt{\Lambda}).$$

While the above theorem does not give an exact characterization of the congestion experienced by the user base, the proof does give a bound on the congestion experienced by arrivals into Firm 1 at an equilibrium. Following the same line of argument as for the non-cooperative model, it follows that, at an equilibrium, Firm 1 provisions $O(\sqrt{\Lambda})$ spare capacity, which is of the same order as in the case of a single firm. This suggests that, as in the non-cooperative case, competition does not significantly improve the payoff of the user base.

We finish with our results for the case of load-dependent latency functions, which parallels Theorem 4. Again, the result is weaker than Theorem 6 but demonstrates that the better Firm 1 is always guaranteed a non-zero market-share. Here we define γ_c by $\gamma_c = l^{-1}(w_1 - w_2)$.

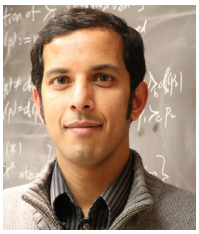
Theorem 7. *Consider a cooperative user base with firm-specific network effects such that* $\beta_1 = \beta_2 = 0$ *and* $w_1 > w_2 > 0$. *Further, take* $f(x, C)$ *to be a load-based latency function given by* $g(x/C)$. *If* $b_1 > \frac{1}{\gamma_c}$, *for large enough* Λ, *any equilibrium* $(\lambda_1, \lambda_2, C_1, C_2)$ *must satisfy*

$$\lambda_1 \ge \frac{b_1 - \frac{1}{\gamma_c}}{b_1 - 1}\Lambda \text{ and } \lambda_2 \le \frac{\frac{1}{\gamma_c} - 1}{b_1 - 1}\Lambda.$$

References

[1] J. Nair, V.G. Subramanian, A. Wierman, On competitive provisioning of cloud services, in: Proceedings of IFIP Performance, 2014.
[2] J. Nair, V.G. Subramanian, A. Wierman, On competitive provisioning of ad-supported cloud services, in: Proceedings of Allerton, 2014.
[3] D. Durkee, Why cloud computing will never be free, Queue 8 (4) (2010) 20.
[4] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, A. Ghalsasi, Cloud computing—The business perspective, Decis. Support Syst. 51 (1) (2011) 176–189.
[5] IAB Internet Advertising Revenue Report, 2011. URL http://www.iab.net/adrevenuereport.
[6] J. Hamilton, The cost of latency, (October 2009), URL http://perspectives.mvdirona.com/2009/10/31/TheCostOfLatency.asp.
[7] S. Lohr, For impatient web users, an wye blink is just too long to wait, New York Times Published Feb. 29. 2012. URL http://www.nytimes.com/2012/03/01/technology/impatient-web-users-flee-slow-loading-sites.html.
[8] R. Kohavi, R. Longbotham, D. Sommerfield, R. Henne, Controlled experiments on the web: survey and practical guide, Data Min. Knowl. Discov. 18 (1) (2009) 140–181.
[9] M. Katz, C. Shapiro, Network externalities, competition, and compatibility, Amer. Econ. Rev. 75 (3) (1985) 424–440.
[10] J. Farrell, P. Klemperer, Coordination and lock-in: Competition with switching costs and network effects, in: Handbook of Industrial Organization, Vol. 3, 2007, pp. 1967–2072.
[11] R. Johari, S. Kumar, Congestible services and network effects, 2009.
[12] J. Nair, A. Wierman, B. Zwart, Provisioning of large-scale systems: The interplay between network effects and strategic behavior in the user base, Manage. Sci. 62 (6) (2016) 1830–1841.
[13] J. Anselmi, U. Ayesta, A. Wierman, Competition yields efficiency in load balancing games, Perform. Eval. 68 (11) (2011) 986–1001.
[14] J. Anselmi, D. Ardagna, J.C. Lui, A. Wierman, Y. Xu, Z. Yang, The economics of the cloud: Price competition and congestion, in: Proceedings of NetEcon, 2013.
[15] J. Buchanan, An economic theory of clubs, Economica 32 (125) (1965) 1–14.
[16] T. Sandler, J. Tschirhart, Club theory: Thirty years later, Public Choice 93 (3) (1997) 335–355.
[17] B. Metcalfe, Metcalfe's law: A network becomes more valuable as it reaches more users, Infoworld 17 (40) (1995) 53–54.
[18] S. Oren, S. Smith, Critical mass and tariff structure in electronic communications markets, Bell J. Econ. (1981) 467–487.
[19] J. Farrell, G. Saloner, Standardization, compatibility, and innovation, Rand J. Econ. 16 (1) (1985) 70–83.
[20] A. Sundararajan, Network effects, nonlinear pricing and entry deterrence, 2003.
[21] H. Chen, Y.-w. Wan, Capacity competition of make-to-order firms, Oper. Res. Lett. 33 (2) (2005) 187–194.
[22] H. Chen, Y.-W. Wan, Price competition of make-to-order firms, IIE Trans. 35 (9) (2003) 817–832.
[23] V.V. Mazalov, A.V. Melnik, Equilibrium prices and flows in the passenger traffic problem, Int. Game Theory Rev. 18 (01) (2016) 1650001.
[24] S. Halfin, W. Whitt, Heavy-traffic limits for queues with many exponential servers, Oper. Res. 29 (3) (1981) 567–588.
[25] J. Reed, The G/GI/N queue in the Halfin-Whitt regime, Ann. Appl. Probab. 19 (2009) 2211–2269.
[26] R. Atar, A diffusion regime with non-degenerate slowdown, Oper. Res. (2013) (in press).
[27] M. Haviv, T. Roughgarden, The price of anarchy in an exponential multi-server, Oper. Res. Lett. 35 (4) (2007) 421–426.
[28] D. Acemoglu, A. Ozdaglar, Competition and efficiency in congested markets, Math. Oper. Res. 32 (1) (2007) 1–31.
[29] A. Odlyzko, B. Tilly, A refutation of Metcalfe's law and a better estimate for the value of networks and network interconnections, 2005.
[30] V. Mazalov, A. Lukyanenko, S. Luukkainen, Equilibrium in cloud computing market, Perform. Eval. 92 (2015) 40–50.
[31] H. Hotelling, Stability in competition, Econom. J. 39 (153) (1929) 41–57.
[32] Z. Liu, M. Lin, A. Wierman, S.H. Low, L.L. Andrew, Geographical load balancing with renewables, ACM Sigmetrics Perform. Eval. Rev. 39 (3) (2011) 62–66.

Jayakrishnan Nair received his BTech and MTech in Electrical Engg. (EE) from IIT Bombay (2007) and Ph.D. in EE from California Inst. of Tech. (2012). He has held post-doctoral positions at California Inst. of Tech. and Centrum Wiskunde & Informatica. He is currently an Assistant Professor in EE at IIT Bombay. His research focuses on modeling, performance evaluation, and design issues in queueing systems and communication networks.



Vijay Subramanian received the B.Tech. degree in electronics engineering from IIT Madras in 1993, the M.Sc. (Engg.) degree in electrical communication engineering from IISc Bangalore in 1995, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign. He was with the Research Arm of the Networks Business Sector, Motorola, Arlington Heights, IL, USA, in 2006. In 2006, he moved to the Hamilton Institute of the National University of Ireland, Maynooth, as a Research Fellow. In 2010, he was a Visiting Researcher with LIDS MIT. From 2010 to 2011, he was a Senior Research Associate with the EECS Department, Northwestern University. From 2011 to 2014, he was a Research Assistant Professor with the EECS Department, Northwestern University. He is currently an Associate Professor with the EECS Department, University of Michigan. His interests are in stochastic modeling, communications, information theory and applied mathematics. A large portion of his past work has been on probabilistic analysis of communication networks, especially analysis of scheduling and routing algorithms. In the past he has also done some work with applications in immunology and coding of stochastic processes. His current research interests are on game theoretic and economic modeling of socio-technological systems and networks, and the analysis of associated stochastic processes.

Adam Wierman is a Professor in the Department of Computing and Mathematical Sciences at the California Institute of Technology. His research interests center around resource allocation and scheduling decisions in computer systems and services. He received the 2011 ACM SIGMETRICS Rising Star award, the 2014 IEEE Communications Society William R. Bennett Prize, and has been coauthor on papers that received of best paper awards at ACM SIGMETRICS, IEEE INFOCOM, IFIP Performance, IEEE Green Computing Conference, IEEE Power & Energy Society General Meeting, and ACM GREENMETRICS.