On Competitive Provisioning Of Cloud Services

Jayakrishnan Nair
Centrum Wiskunde & Informatica

Vijay G. Subramanian
Northwestern University

Adam Wierman
California Institute of Technology

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 increasingly 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).

For the purposes of this paper, there are four main features of this growing marketplace that are important to highlight.

(i) A majority of cloud services *derive revenue primarily from advertising* and are offered for free to users. For example, companies like Google and Facebook make billions of dollars annually from ad supported online services [10].

(ii) Users of online services are *highly delay sensitive*. Small additional delays for users can be traced to significant declines in revenue for cloud services [9, 13, 14].

(iii) Cloud services have *positive network effects*, i.e., the experience of users in cloud services often is highly dependent on how many users the service has [6, 11, 12]. For example, social networking services, online gaming environments, etc.

(iv) Cloud services are often *highly competitive* [5, 15]. For example, the competition between Hotmail, Gmail, and Yahoo mail, or the competition between Facebook and GooglePlus.

The interplay of these four factors leads to a complicated cloud marketplace with significant interaction between user experience (congestion and network effects), service capacity provisioning, and market share. The goal of this paper is to investigate the interplay of these factors within an analytic model.

Network effects and congestion

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., [6,7,12,16,18,19,22], but the literature on "club theory" focuses on the interaction of network effects and congestion.

The theory of clubs, which originated from [4], deals with groups of congestion sensitive users sharing a certain resource. See [21] 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 (and the broader literature on network effects) it is assumed that the users *pay* for access, and the revenue of the provider is made up exclusively of such payments. This is very different than 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 applicability of the conclusions from the models.

The only previous piece of work to consider network effects and congestion in an ad-supported service is [17], which focuses on the capacity provisioning of a single service when faced with a strategic user population with positive network effects. In this setting, [17] 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, [17] studies only the behavior of a single, monopolistic service. Thus, no piece of prior work has investigated the interplay of all four of the factors described above.

Contributions of this paper

The goal of this paper is to understand the interplay of network effects, congestion, and competition in ad-supported services. Thus, we seek answer questions such as: Does competition lead to improved user experience in ad-supported services? Can competing firms coexist or will near-monopolies emerge?

To address such questions, we introduce a new model that extends the setting of [17] in order to capture competition

between service providers, a.k.a., firms (see Section 3).

The key novelty in this extension is how network effects are considered. We consider two variations of network effects in this paper: *firm-specific network effects* and *industry-wide network effects*.

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 the 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 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 effects of each.

Within these models of network effects we study a situation where users from a user base decide (in either a cooperative or a non-cooperative manner) 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 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., [3, 8, 20], while the scaling emerges endogenously in our model.

The main messages that result from our analysis are the following.

1. As in the case of the single-firm setting in [17], positive network effects allow firms to run fewer servers, and thus increase profits. One should expect this effect to diminish with increased competition. However, surprisingly, our results highlight that users are generally no better off due to competition. This is in contrast to results such as [1, 2], which prove that increased competition among cloud services result in improved user performance when paid services are considered.
2. Our results highlight important contrasts between the firm-specific and industry-wide network effects models. In particular, Theorems 3, 4 and 5 highlight that firms can share the market in the case of industry-wide network effects; however Theorems 6 and 7 highlight that near monopolies tend to emerge in the case of firm-specific network effects. This explains what can be informally observed in the cloud service marketplace: Facebook enjoys near-monopoly status while Gmail, Hotmail, and many other cloud-based email providers 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. A similar example is GooglePlus.

Importantly, the messages described above seem to emerge *because* of the ad-supported nature of the services we consider. In particular, they contrast with results from club theory for settings where the firms charge users directly.

The remainder of the paper is organized as follows. In Section 2 we derive preliminary results for the case of a single provider. These results are needed later for the case of competing firms. Then, in Section 3 we introduce the models of network effects that we consider in the case of competing firms. Sections 4 and 5 present our main results for each of the two network effects models. Finally, we end with a discussion of the results in Section 6.

2. PRELIMINARIES: A SINGLE FIRM

Before moving to the case of competing firms, it is useful to start with the case of a single firm. Note that the treatment in this section parallels some of the analysis of [17]. However, [17] considers a different latency function than we consider here, and so we need new analysis in order to provide tools for the analysis of the case of competing firms.

2.1 Model

In this section we consider the case of a single firm that provides a cloud service. There are three main components to the model: the latency experienced by the users, the strategic behavior of the users, and the strategic behavior of the firm. Each is described in the following.

2.1.1 Latency model

We consider a model where each user perceives a non-negative latency cost $f(\lambda, C)$ that is a function of both the arrival rate of users, λ, and the provisioned capacity of the provider, C. Specifically, we consider the following 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.,
$$f(\lambda, C) = \begin{cases} \frac{1}{C-\lambda} & \text{if } \lambda < C \\ +\infty & \text{otherwise} \end{cases}$$
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 $g(\rho) = \frac{\rho}{1-\rho}$.

2.1.2 User model

To model the user base, we assume that the users of a cloud service arrive at rate at most Λ. At an arrival rate $\lambda \in [0, \Lambda]$, the users collectively gain a utility $U(\lambda)$. To model positive network effects we assume that U is of the form
$$U(\lambda) = w\lambda^{1+\beta}, \tag{1}$$
for $w > 0$ and $\beta \in [0, 1]$, where the network effect becomes stronger as β increases.[1]

The rate of arriving users in $[0, \Lambda]$ gets determined as a consequence of the strategic decisions of the users, which seek to maximize their net payoff. We distinguish between two models: non-cooperative and cooperative. In the non-cooperative model (also known as the user-optimal model

[1]The restriction that $\beta \leq 1$ is made for ease of exposition. Our results also extend to the case $\beta > 1$.

in the Wardrop equilibrium literature), each user's payoff is her utility minus the latency cost. In the cooperative model (also known as social-optimal model in the Wardrop equilibrium literature), the net utility of all the users subtracted by the net latency cost is the payoff function that is maximized.

Before discussing the formal details of these models, note that the non-cooperative model applies when each user takes an individual and selfish decision whereas the cooperative model applies when a global decision is taken across all the users. One can argue that the non-cooperative scenario models realistic user behaviors whereas the cooperative scenario is an ideal benchmark that users would arrive at if they were able to take a collective perspective. Importantly, the cooperative scenario is not what the social planner would optimize as the social welfare maximization problem would include the profit(s) of the firm(s) as well.

Non-cooperative user behavior. Recall that, collectively, the users make a utility $U(\lambda)$ when the arrival rate is λ. Therefore, the per user utility is given by $V(\lambda) := U(\lambda)/\lambda$. Since the per user latency cost is $f(\lambda, C)$, the payoff function that the users maximize is $V(\lambda) - f(\lambda, C)$. Therefore, the arrival rate of users in the non-cooperative model is given by

$$\hat{\lambda}_\Lambda(C) = \max\{\lambda \in [0, \Lambda] \mid V(\lambda) - f(\lambda, C) \geq 0\}. \quad (2)$$

Note that we have chosen the largest Wardrop equilibrium if multiple solutions exist.[2] With $U(\cdot)$ being a convex increasing function with $U(0) = 0$, it follows from the subgradient inequality that $V(\cdot)$ is an increasing function of λ with $V(0) = U'(0)$, the right derivative of $U(\cdot)$ at 0.

Cooperative user behavior. Since the per user latency cost is $f(\lambda, C)$, collectively the latency cost is $\lambda f(\lambda, C)$ and the payoff function that the social planner maximizes is given by $U(\lambda) - \lambda f(\lambda, C)$. From the non-negativity of f it follows that the total latency cost is an increasing function of λ. Therefore, the arrival rate of users in the cooperative model is given by

$$\hat{\lambda}_\Lambda(C) = \max\{\arg\max_{\lambda \in [0,\Lambda]} U(\lambda) - \lambda f(\lambda, C)\}. \quad (3)$$

Note that we choose the largest maximizer if multiple solutions exist.[3]

2.1.3 Modeling the firm

The final piece of the model is the strategic behavior of the firm, which seeks to choose capacity so as to maximize profit. We assume that the cloud service provider makes b dollars per user served from advertising and pays a dollar per unit cost for each unit of capacity. Thus the firm's profit is given by

$$b\hat{\lambda}_\Lambda(C) - C. \quad (4)$$

Owing to the stability constraint, $\hat{\lambda}_\Lambda(C) < C$, and so we necessarily need $b > 1$ for the firm to consider offering the service. From this discussion, it is also clear that a firm will not provision capacity that is greater than $b\Lambda$. Therefore, the firm's problem can be written as choosing

$$C^*(\Lambda) = \max\{\arg\max_{C \in [0, b\Lambda]} b\hat{\lambda}_\Lambda(C) - C\}, \quad (5)$$

Note that we choose the largest capacity if multiple solutions exist.[3]

2.2 Results for a single firm

Given the model described in the previous section, our focus is on characterizing the operating point of the system when Λ is large. Specifically, our interest is in the behavior of $C^*(\Lambda)$ and $\lambda^*(\Lambda) := \hat{\lambda}_\Lambda(C^*(\Lambda))$ for large Λ, and the queueing regime (moderate, heavy or very-heavy traffic) that emerges endogenously.

As we have mentioned, the results that follow parallel those in [17], providing similar insights. However, [17] focuses on M/M/k latency functions, and we need results for M/M/1 and load based latency functions in the analysis of competing firms later in this paper.

The non-cooperative setting. We start by defining the unconstrained population response as follows:

$$\tilde{\lambda}(C) = \max\{\lambda \geq 0 \mid V(\lambda) - f(\lambda, C) \geq 0\},$$
$$\tilde{\rho}(C) = \frac{\tilde{\lambda}(C)}{C},$$

where $\tilde{\lambda}(C)$ is the arrival rate the provider can attract when the provider provisions capacity C, assuming the potential arrival rate is unlimited. This unconstrained response is a key analytical tool in the analysis of the operating point of the system. Specifically, we will see that $C^*(\Lambda)$ equals that value of C that satisfies $\tilde{\lambda}(C) = \Lambda$.

For the M/M/1 latency function we have the following characterization of $\tilde{\lambda}(C)$.

LEMMA 1. *For large enough C, $\tilde{\lambda}(C)$ is continuous and strictly increasing in C, and*

$$V(\tilde{\lambda}(C))(C - \tilde{\lambda}(C)) = 1.$$

Note that the above lemma implies that for large enough C, $\tilde{\rho}(C)$ is strictly increasing in C. Moreover, let $\tilde{C}(\lambda)$ denote the inverse of $\tilde{\lambda}(C)$, i.e., $\tilde{C}(\lambda)$ is the capacity the provider must provision in order to attract an arrival rate of λ. It follows from Lemma 1 that for large enough λ,

$$\tilde{C}(\lambda) = \lambda + \frac{1}{V(\lambda)}.$$

We omit the proof of this lemma, since it follows along similar lines as Lemma 2 in [17].

Next, we characterize $\tilde{\lambda}(C)$ for the case of load based latency functions.

LEMMA 2. *If $\beta \in [0, 1)$, then for all $C > 0$, $\tilde{\lambda}(C)$ is continuous and strictly increasing in C and is obtained by solving for the unique positive λ that satisfies*

$$w\lambda^\beta = g\left(\frac{\lambda}{C}\right).$$

If $\beta = 1$, then a non-zero solution exists if and only if $w > g'(0)/C$, i.e., for C large enough.

[2]The choice of the largest Wardrop equilibrium is made to concretely define the behavior of the user base. The results stated are somewhat robust to this assumption, e.g., they continue to hold even if the minimal Wardrop equilibrium is picked.

[3]As before, we pick the largest maximizer for concreteness. The stated results continue to hold if we pick the smallest maximizer.

We prove this lemma in Appendix A. Note that the above lemma implies that for $\beta > 0$, $\tilde{\rho}(C)$ is strictly increasing in C. Moreover, let $\tilde{C}(\lambda)$ denote the inverse of $\tilde{\lambda}(C)$, i.e., $\tilde{C}(\lambda)$ is the capacity the provider must provision in order to attract an arrival rate of λ. It follows from Lemma 2 that, for large enough λ,

$$\tilde{C}(\lambda) = \lambda/h(\lambda),$$

where

$$h(\lambda) := g^{-1}(w\lambda^{\beta}).$$

Building on the above, the following theorem characterizes the operating point of the service as Λ becomes large.

THEOREM 1. *Consider case of a single service provider. Under the non-cooperative user model, for large enough Λ, the following scaling behaviors hold.*

1. *For the M/M/1 latency function,*

$$\lambda^*(\Lambda) = \Lambda,$$
$$C^*(\Lambda) = \tilde{C}(\Lambda) = \Lambda + \frac{1}{V(\Lambda)}.$$

2. *For load based latency functions, if $\beta > 0$, or $\beta = 0$ and $b > \frac{1}{h(1)}$,*

$$\lambda^*(\Lambda) = \Lambda,$$
$$C^*(\Lambda) = \tilde{C}(\Lambda) = \frac{\Lambda}{h(\Lambda)}.$$

It is interesting to note that for the M/M/1 latency function, the provider operates the service at an extremely heavy traffic regime, with only a bounded spare capacity. To interpret the capacity provisioning for the load based latency function, consider the special case of the mean number of jobs in an M/M/1 queue, i.e., $g(\rho) = \frac{\rho}{1-\rho}$. In this case, $h(\lambda) = \frac{w\lambda^{\beta}}{1+w\lambda^{\beta}}$, implying that

$$C^*(\Lambda) = \Lambda + \frac{1}{w}\Lambda^{1-\beta}.$$

Thus, when $\beta = 0$, the service is operated at constant utilization $\frac{w}{1+w}$ (if $b > \frac{w+1}{w}$; otherwise, the provider is unable to operate the service profitably). As β increases, i.e., as the network effects grow stronger, the firm operates the service in heavier traffic regimes, with a spare capacity $\Theta(\Lambda^{1-\beta})$. Intuitively, as users derive an increased utility due to network effects, the firm can operate the service at higher levels of congestion.

We omit the proof of Theorem 1 as it follows easily from Lemmas 1 and 2 using similar arguments as in [17].

The cooperative setting. As in our analysis of the non-cooperative model, we define the unconstrained population response

$$\tilde{\lambda}(C) = \max\{\arg\max_{\lambda \geq 0} U(\lambda) - \lambda f(\lambda, C)\},$$
$$\tilde{\rho}(C) = \frac{\tilde{\lambda}(C)}{C}.$$

As before, it turns out that the unconstrained population response determines the operating point of the system, i.e., $C^*(\Lambda)$ equals that value of C that satisfies $\tilde{\lambda}(C) = \Lambda$.

The following lemma characterizes $\tilde{\lambda}(C)$ under the M/M/1 latency function.

LEMMA 3. *For large enough C, $\tilde{\lambda}(C)$ is continuous and strictly increasing with respect to C. Also, for large enough C, $\tilde{\lambda}(C)$ satisfies*

$$\sqrt{U'(\tilde{\lambda}(C))C}(1 - \tilde{\rho}(C)) = 1.$$

It follows from the above lemma that for large enough C, $\tilde{\rho}(C)$ is strictly increasing. As before, define $\tilde{C}(\lambda)$ to be the inverse of $\tilde{\lambda}(C)$. It follows from Lemma 3 that

$$\tilde{C}(\lambda) = \lambda + \sqrt{\frac{\lambda}{U'(\lambda)}} + o\left(\sqrt{\frac{\lambda}{U'(\lambda)}}\right).$$

We give the proof of Lemma 3 in Appendix B.

Next we consider load based latency functions.

LEMMA 4. *If $\beta \in [0, 1)$, then for all $C > 0$, $\tilde{\lambda}(C)$ is continuous and strictly increasing in C, and is obtained by solving for the unique positive solution of*

$$w(1+\beta)\lambda^{\beta} = \frac{\lambda}{C}g'\left(\frac{\lambda}{C}\right) + g\left(\frac{\lambda}{C}\right).$$

If $\beta = 1$, then a non-zero solution exists if and only $w > g'(0)/C$, i.e., for C large enough.

We prove the above lemma in Appendix C. Note that Lemma 4 implies that for $\beta > 0$, $\tilde{\rho}(C)$ is strictly increasing in C, since $l(x) := xg'(x) + g(x)$ is an increasing function. Moreover, let $\tilde{C}(\lambda)$ denote the inverse of $\tilde{\lambda}(C)$, i.e., $\tilde{C}(\lambda)$ is the capacity the provider must provision in order to attract an arrival rate of λ. It follows from Lemma 3 that for large enough λ,

$$\tilde{C}(\lambda) = \lambda/h(\lambda),$$

where we now have

$$h(\lambda) := l^{-1}(w(1+\beta)\lambda^{\beta}).$$

Finally, Lemmas 3 and 4 lead to the following characterization of the operating point of the service as Λ becomes large.

THEOREM 2. *Consider case of a single service provider. Under the cooperative user model, for large enough Λ, the following scaling behaviors hold.*

1. *For the M/M/1 latency function,*

$$\lambda^*(\Lambda) = \Lambda,$$
$$C^*(\Lambda) = \tilde{C}(\Lambda) = \Lambda + \sqrt{\frac{\Lambda}{U'(\Lambda)}} + o\left(\sqrt{\frac{\Lambda}{U'(\Lambda)}}\right).$$

2. *For load based latency functions, if $\beta > 0$, or $\beta = 0$ and $b > \frac{1}{h(1)}$,*

$$\lambda^*(\Lambda) = \Lambda,$$
$$C^*(\Lambda) = \tilde{C}(\Lambda) = \frac{\Lambda}{h(\Lambda)}.$$

It is instructive to compare the operating point of the cooperative user model with that for the non-cooperative model. For the M/M/1 latency function, note that the provider provisions $\Theta(\Lambda^{(1-\beta)/2})$ spare capacity in the cooperative model, compared to a bounded spare capacity in the non-cooperative case. This is a 'tragedy of the commons' phenomenon, wherein anarchy in the user base drives the system into a more congested state.

Also, note that the spare capacity in the cooperative case shrinks as β increases, i.e., as the network effects become stronger. Again, intuitively, this is because the service provider is able to exploit the stronger network effects to operate the system in a higher state of congestion.

Finally, let us turn to the load based latency function. We see that the operating point looks structurally similar to that in the non-cooperative model. Note, however, that the function h is different for both cases. To interpret the operating point, consider again the special case of the mean number of jobs in an M/M/1 queue, i.e., $g(\rho) = \frac{\rho}{1-\rho}$. In this case, $l(x) = x(2-x)/(1-x)^2$, and

$$h(\lambda) = 1 - \frac{1}{\sqrt{1 + w(1+\beta)\lambda^{\beta}}}.$$

Employing a Taylor expansion, one can show that

$$\frac{\Lambda}{h(\Lambda)} = \Lambda + \frac{1}{\sqrt{w(1+\beta)}}\Lambda^{1-\beta/2} + o(\Lambda^{1-\beta/2}).$$

As expected, the spare capacity, which is $\Theta(\Lambda^{1-\beta/2})$, shrinks as the network effects grow stronger. Moreover, note that as a result of the 'tragedy of the commons' effect, the spare capacity exceeds that under the non-cooperative model.

We omit the proof of Theorem 2 as it follows easily from Lemmas 3 and 4 using similar arguments as in [17].

3. COMPETING FIRMS

So far we have discussed the case of a single cloud provider (firm) and compared the capacity scalings achieved for both the (realistic) non-cooperative and (idealistic) cooperative scenarios. Of course, in reality cloud providers always have competition. Thus, we now move from a single firm to competing firms. Our goal is to study the interplay of network effects, congestion, and competition in ad-supported services.

To maintain analytic tractability, we focus on the case of two competing firms. This is clearly a small number of firms, but it is already enough to highlight the role of competition. Of course, moving beyond two firms is an interesting, challenging direction for future work.

The model we consider in the remainder of the paper is an extension of the setting introduced in Section 2 to the case of two competing firms. In this extension, the latency model described in Section 2.1.1 remains the same. However, the models for the user population and the firm need to be adapted. We discuss these models in this section and then present our results in Sections 4 and 5.

3.1 User model

Given the existence of two firms, a total population of users Λ results in an arrival rate λ that is split across the firms such that firm i has arrival rate λ_i and $\lambda = \lambda_1 + \lambda_2$.

The key change to the user model comes in the consideration of network effects. As described in the introduction, there are two contrasting notions of network effects that are relevant to cloud services: industry-wide network effects and firm-specific network effects. If network effects are industry-wide, than the experience of a user is improved by having a larger aggregate population of users across all firms, while 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.

More formally, when network effects are industry-wide then, irrespective of where the users obtain their service, the total utility of the users is

$$U(\lambda) := w\lambda^{1+\beta}$$

and the per user utility is $V(\lambda) := U(\lambda)/\lambda$, as in the single firm case.

In contrast, when network effects are firm specific the total utility of users subscribed to firm i is

$$U_i(\lambda_i) := w_i\lambda_i^{1+\beta_i}$$

and the per user utility is $V_i(\lambda_i) = U_i(\lambda_i)/\lambda_i$. In this case, we insist on different utility functions for the users from the different firms. The latency cost, however, remains exactly the same, i.e., the collective latency cost for users subscribed to firm i is $\lambda_i f(\lambda_i, C_i)$ and the per user latency cost is $f(\lambda_i, C_i)$.

In each case, the specific split of traffic across the two firms depends on whether users are cooperative or non-cooperative. We discuss the details of these two models in the case of industry-wide and firm-specific network effects Sections 4 and 5, respectively, where we also present our results for each setting.

3.2 Modeling competing firms

Not much change is needed to extend the model of a single firm in Section 2.1.3 to competing firms. In particular, each firm $i = 1, 2$ uses the same model described in Section 2.1.3, except now the traffic they receive is λ_i and the revenue per user served is b_i dollars. Thus, firm i chooses capacity C_i such that

$$C_i \in \arg\max_{C_i \geq 0} b_i\lambda_i - C_i \tag{6}$$

Note that λ_i is a function $\hat{\lambda}_i(C_1, C_2)$ of (C_1, C_2) that depends on the particular user model considered, i.e., industry-wide vs. firm-specific network effects and cooperative vs. non-cooperative behavior.

The key change to the model of the firms comes from the interaction due to competition. Since the traffic split depends on the capacity provisioning of both firms, the decisions of the two firms get coupled. We study this via a game and consider the Nash equilibria.

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.,

$$\lambda_i = \hat{\lambda}_i(C_1, C_2) \text{ for } i = 1, 2$$

$$C_1 \in \arg\max_{c \geq 0} b_1\hat{\lambda}_1(c, C_2) - c$$

$$C_2 \in \arg\max_{c \geq 0} b_2\hat{\lambda}_2(C_1, c) - c.$$

Given this definition, our goal in the remainder of the paper is to study the equilibria that can emerge among firms when Λ is large.

4. INDUSTRY-WIDE NETWORK EFFECTS

In this section we present our results characterizing two competing, ad-supported firms in the context where *network effects are industry-wide*. We study both the case of a cooperative and a non-cooperative user population. In each case, we start by describing the details of the user model

and then present our results. Results are presented tersely here and then discussed in Section 6.

4.1 The non-cooperative setting

Non-cooperative user behavior. In the non-cooperative scenario with industry-wide network effects, users getting service from firm $i = 1, 2$ see a net per-user-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 (λ_1, λ_2) such that $\lambda_1 + \lambda_2 = \lambda$ is a Wardrop equilibrium if the following condition hold for $i = 1, 2$

$$\lambda_i > 0 \Rightarrow V(\lambda) - f(\lambda_i, C_i) = \max_{j=1,2}\{V(\lambda) - f(\lambda_j, C_j)\} \quad (7)$$

Note that $(\lambda, 0)$ and $(0, \lambda)$ can also be Wardrop equilibria, but only if they also satisfy (7).

For a given (λ, C_1, C_2), where $\lambda < C_1 + C_2$, it is easy to verify that the Wardrop equilibrium $[\lambda_1(\lambda, C_1, C_2), \lambda_2(\lambda, C_1, C_2)]$ is the unique solution of the following convex optimization.

$$\begin{aligned} &\min_{\lambda_1, \lambda_2} \sum_{i=1}^{2} \int_0^{\lambda_i} f(x, C_i)dx \\ &\text{subject to } \sum_{i=1}^{2} \lambda_i = \lambda \end{aligned} \quad (8)$$

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

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

$$\begin{aligned} \hat{\lambda}(C_1, C_2) = \max \Big\{ &\lambda \in [0, \Lambda] \cap [0, C_1 + C_2) \mid \\ &\max_{j=1,2} \left\{ V(\lambda) - f(\lambda_j(\lambda, C_1, C_2), C_j) \right\} \geq 0 \Big\} \end{aligned} \quad (9)$$

The user behavior $[\hat{\lambda}_1(C_1, C_2), \hat{\lambda}_2(C_1, C_2)]$ is then defined by

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

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.

Results. For the non-cooperative setting, we provide results for both the M/M/1 latency and load based latency functions. The basic message of both results is that the competing firms can co-exist in the market when network effects are industry-wide.

We start by presenting results for the M/M/1 latency function.

THEOREM 3. *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 $i = 1, 2$.

Note that the cases of the theorem correspond to differing comparisons of the 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 not incredibly good, advertising efficiencies, which results in a multitude of possible ways for the firms to divide the market.

PROOF. We 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, it is beneficial for any firm to provision capacity $\Lambda + \frac{1}{V(\Lambda)}$, and command the full market, as in the single provider case.

Consider then the possibility of an equilibrium satisfying $C_1 > 0$, $C_2 = 0$. Such an equilibrium must necessarily have $C_1 = \Lambda + \frac{1}{V(\Lambda)}$, since that is the optimal provisioning for Firm 1, given that Firm 2 does not offer the service. However, this cannot be an equilibrium, since Firm 2 can increase its capacity to match Firm 1, leading to $\lambda_2 = \Lambda/2$, and a profit of $(\frac{b_2}{2} - 1)\Lambda - \frac{1}{V(\Lambda)}$, which is positive for large enough Λ.

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 arrivals would just decrease capacity to zero. Consider now the action of Firm 1 increasing its capacity by a small ϵ. Since the Wardrop split keeps the spare capacity balanced, 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 $(\frac{b_2}{2} - 1)\Lambda - \frac{1}{V(\Lambda)}$, 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) above. 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 the provider to increase capacity further, since any increase c in capacity would lead to an increase of $c/2$ in the arrival rate (since the Wardrop split balances the spare capacity across providers); this is not favorable given $b_i \leq 2$. Consider next the action of decreasing capacity by c, 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

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

where $\lambda' = \lambda_1' + \lambda_2'$. The above inequalities imply that $\lambda_1' \leq \lambda_1 - c$, which implies a decrease in profit. Thus, each provider has no 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 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 either provider has an incentive to decrease capacity. □

It is important to note that in every equilibrium configuration demonstrated in Theorem 3, the user base sees exactly the same congestion (measured, for the M/M/1 latency function, in terms of the spare capacity) as in the case of a single provider (See Theorem 1). Thus, we see that competition in the marketplace does not improve the payoff experienced by the user base.

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

THEOREM 4. *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*

$$b_1, b_2 > \frac{1}{h(\Lambda)},$$

and

$$\tfrac{1}{b_1} + \tfrac{1}{b_2} \geq h(\Lambda),$$

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

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

This theorem 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. Interestingly, though the overall message matches that of Theorem 3, 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 4 to hold for sufficiently large Λ.

PROOF. 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). Then we must have

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

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 λ_1'. Let $\rho_1' = \frac{\lambda_1'}{C_1'}$ denote its new utilization. Clearly, $\rho_1' \leq h(\Lambda)$. 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, which completes the proof. □

Once again, we note that in every equilibrium configuration demonstrated in Theorem 4, 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 (See Theorem 1).

4.2 The cooperative setting

Cooperative user behavior. In the cooperative scenario with industry-wide network effects, the traffic split is determined by maximizing the collective payoff. That is, the user base solves the following optimization problem.

$$\begin{aligned} \max \quad & U(\lambda) - \sum_{i=1}^{2} \lambda_i f(\lambda_i, C_i) \\ \text{subject to } & \sum_{i=1}^{2} \lambda_i = \lambda, \\ & \lambda \leq \Lambda, \\ & \lambda_1, \lambda_2 \geq 0 \end{aligned} \tag{10}$$

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.

$$\begin{aligned} \min_{\lambda_1,\lambda_2} \quad & \sum_{i=1}^{2} \lambda_i f(\lambda_i, C_i) \\ \text{subject to } & \sum_{i=1}^{2} \lambda_i = \lambda, \\ & \lambda_1, \lambda_2 \geq 0 \end{aligned} \tag{11}$$

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

Results. For the cooperative setting, our results focus only on the class of load based latency functions, e.g., mean occupancy of an M/M/1. The key message of the results parallels that in the non-cooperative setting: multiple competing firms can coexist when network effects are industry-wide. In fact, the equilibria that emerge behave similarly to the equilibria in the non-cooperative case. This is a consequence of the traffic split in both cases being proportional to the capacities provisioned.

Recall that for this set of results, $h(\lambda) = l^{-1}(w(1+\beta)\lambda^{\beta})$ where $l(x) = xg'(x) + g(x)$, and $\lim_{\lambda\to\infty} h(\lambda) = 1$ for $\beta > 0$.

THEOREM 5. *Consider a 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*

$$b_1, b_2 > \frac{1}{h(\Lambda)},$$

and

$$\tfrac{1}{b_1} + \tfrac{1}{b_2} \geq h(\Lambda),$$

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

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

It is important to realize that, even though the equilibria in the cooperative case behave in a similar manner as the equilibria in the non-cooperative case, the actual scaling is different owing to the different $h(\lambda)$ functions.

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.

PROOF. Recall that, under the cooperative model, the traffic is the unique solution of the optimization (11). Specializing to the load based latency function, the first order condition for this optimization is the following.

$$g\left(\frac{\lambda_1}{C_1}\right) + \frac{\lambda_1}{C_1} g'\left(\frac{\lambda_1}{C_1}\right) = g\left(\frac{\lambda_2}{C_2}\right) + \frac{\lambda_2}{C_2} g'\left(\frac{\lambda_2}{C_2}\right)$$

Now, it is easy to see that, as in the non-cooperative case, the optimal split of traffic is given by $\lambda_i/C_i = \lambda/(C_1 + C_2)$.

Next, we note that the optimization objective of the firm can be written as

$$U(\lambda) - \lambda_1 g(\frac{\lambda_1}{C_1}) - \lambda_2 g(\frac{\lambda_1}{C_1}) = U(\lambda) - \lambda g(\frac{\lambda}{C}),$$

where $C = C_1 + C_2$. Thus, we see that the total arrival rate is determined exactly as in the single provider case, taking the capacity of the single provider to be C. It then follows from Theorem 2 that for a net capacity $C = \frac{\Lambda}{h(\Lambda)}$ provisioned, the net arrival rate equals Λ, and the traffic split is given by $\lambda_i = \Lambda\frac{C_i}{C}$. From here on, the proof follows exactly along the same lines as the proof of Theorem 4. □

Just as we saw before in the case of a non-cooperative user base, note that in each equilibrium configuration demonstrated in Theorem 5, the user base experiences the same congestion (measured here in terms of the load) as in the single provider configuration (see Theorem 2). Thus, we conclude that phenomenon of competition not improving the payoff of users cannot be attributed to anarchy in the user base.

5. FIRM-SPECIFIC NETWORK EFFECTS

In this section we present our results characterizing two competing, ad-supported firms in the context where *network effects are firm-specific*. We study both the case of a cooperative and a non-cooperative user population. In each case, we start by describing the details of the user model and then present our results. Results are presented tersely here and then discussed in Section 6.

5.1 The non-cooperative setting

Non-cooperative user behavior. In the non-cooperative scenario with firm-specific network effects the traffic split is once again obtained using a Wardrop equilibrium. However, in contrast to the industry-wide network effects model,

the utility obtained and the latency costs are both different across firms. Specifically, with a traffic split of (λ_1, λ_2), users subscribed to Firm 1 have a per-user payoff given by $V_1(\lambda_1) - f(\lambda_1, C_1)$ whereas the users subscribed to Firm 2 have a per-user payoff of $V_2(\lambda_2) - f(\lambda_2, C_2)$.

Given this distinction, the Wardrop equilibrium condition now becomes

$$\lambda_i > 0 \Rightarrow V_i(\lambda_i) - f(\lambda_i, C_i) = \max_{j=1,2} \left\{ V_j(\lambda_j) - f(\lambda_j, C_j) \right\} \tag{12}$$

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.

In this section, we restrict ourselves to 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 $i = 1, 2$. This assumption allows us to formally define the Wardrop split as follows. For a given (λ, C_1, C_2), where $\lambda < C_1 + C_2$, the Wardrop equilibrium $[\lambda_1(\lambda, C_1, C_2), \lambda_2(\lambda, C_1, C_2)]$ is the unique solution of the following convex optimization.

$$\begin{aligned} \min_{\lambda_1, \lambda_2} & \sum_{i=1}^{2} \int_0^{\lambda_i} (-w_i + f(x, C_i)) dx \\ \text{subject to } & \sum_{i=1}^{2} \lambda_i = \lambda \end{aligned} \tag{13}$$

Note that the objective function above is strictly convex, implying the solution is unique.[4]

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.

$$\hat{\lambda}(C_1, C_2) = \max \Big\{ \lambda \in [0, \Lambda] \cap [0, C_1 + C_2) \mid \max_{j=1,2} \left\{ V_j(\lambda_j(\lambda, C_1, C_2)) - f(\lambda_j(\lambda, C_1, C_2), C_j) \right\} \geq 0 \Big\}$$

The user behavior $[\hat{\lambda}_1(C_1, C_2), \hat{\lambda}_2(C_1, C_2)]$ is then defined by

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

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.

Results. In this case, we present results for the M/M/1 latency function and linear utilities.

Note that if $w_1 = w_2$, then we recover the industry-wide network effects model. Recall from Theorem 3 that in this case, a continuum of equilibria are possible, with both firms sharing the market. We therefore focus here on the case $w_1 \neq w_2$. Let us assume, without loss of generality, that $w_1 > w_2$.

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.[5]

THEOREM 6. *Consider a non-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. For large enough Λ, any equilibrium $(\lambda_1, C_1, \lambda_2, C_2)$ must satisfy*

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

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 $(\lambda_1, C_1, \lambda_2, C_2)$, 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

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

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 users.

PROOF. Suppose that $(\lambda_1, C_1, \lambda_2, C_2)$ is an equilibrium. We will show that

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

This implies the statement of the lemma, 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 $C_1 = \Lambda + \frac{1}{w_1}$. Now, if $C_2 \leq \frac{1}{w_2}$, then $\lambda_1 = \Lambda$, and our claim on the profit follows easily. If $C_2 > \frac{1}{w_2}$, then the population split is determined by

$$w_1 - \frac{1}{C_1 - \lambda_1} = w_2 - \frac{1}{C_2 - \lambda_2}.$$

Define the 'spare capacities' $s_1 = C_1 - \lambda_1$, $s_2 = C_2 - \lambda_2$. Note that the spare capacities uniquely determine the population split. We have

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

along with $s_1 \geq \frac{1}{w_1}$, $s_2 \geq \frac{1}{w_2}$. Solving the above equations yields

$$s_1 = \frac{s}{2} + \frac{1}{\Delta} - \sqrt{\frac{s^2}{4} + \frac{1}{\Delta^2}} \leq \frac{1}{\Delta}.$$

Accordingly, we obtain

$$\lambda_1 = C_1 - s_1 \geq \Lambda + \frac{1}{w_1} - \frac{1}{\Delta}.$$

[4] The reason we are unable to handle the case $\beta_i > 0$ 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, making a precise definition of the user behavior problematic.

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

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

$$b_1\left(\Lambda+\frac{1}{w_1}-\frac{1}{\Delta}\right)-(\Lambda_1+\frac{1}{w_1})$$
$$=(b_1-1)\Lambda-\left(\frac{b_1w_2}{w_1(w_1-w_2)}+\frac{1}{w_1}\right).$$

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

5.2 The cooperative setting

Cooperative user behavior. In the cooperative scenario with firm-specific network effects the traffic split is chosen to maximize the collective payoff. Since users that subscribe to firm $i=1,2$ receive a collective payoff of $U(\lambda_i)-\lambda_i f(\lambda_i,C_i)$, the user behavior $[\hat{\lambda}_1(C_1,C_2),\hat{\lambda}_2(C_1,C_2)]$ is defined to be the solution of the following optimization problem.

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

As in the previous section we restrict attention to linear utilities ($\beta_1=\beta_2=0$), i.e., $U_i(\lambda_i)=w_i\lambda_i$ with $w_i>0$ for $i=1,2$. Since the objective function is strictly convex in (λ_1,λ_2) it follows that the traffic split is unique.

Results. In this case, we again present results for the M/M/1 latency function and linear utilities.

In order to make a comparison with our results for the non-cooperative user model, we restrict ourselves to the case of linear utilities ($\beta_1=\beta_2=0$.) Recall that that the case of $w_1=w_2$ reduces to the industry-wide network effects model. Therefore, the case of interest is $w_1>w_2$.

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.[6] 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.

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 to be the M/M/1 latency function. As Λ becomes large, any equilibrium $(\lambda_1,C_1,\lambda_2,C_2)$ (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 that follows 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.

PROOF. For any action C_2 by Firm 2, consider Firm 1's response of setting $C_1=C^*_{w_1}(\Lambda)$, where $C^*_{w_1}$ is the equilibrium response in the single provider case (given by Theorem 2). Note that we emphasize here the dependence of this response on w_1;

$$C^*_{w_1}(\Lambda)=\Lambda+\sqrt{\frac{\Lambda}{w_1}}+o\left(\sqrt{\Lambda}\right).$$

Let us denote the emerging population split resulting from Firm 1's action by $(\lambda_1(\Lambda,C_2),\ \lambda_2(\Lambda,C_2))$. We will obtain a lower bound on $\lambda_1(\Lambda,C_2)$, and therefore a lower bound on Firm 1's profit. Specifically, we will show that that

$$\lambda_1(\Lambda,C_2)\ge\lambda_1(\Lambda,\infty)=\Lambda-\sqrt{\Lambda}\left(\frac{1}{\sqrt{\Delta}}-\frac{1}{\sqrt{w_1}}\right)+o(\sqrt{\Lambda}). \tag{14}$$

Above, $\lambda_1(\Lambda,\infty)$ represents the split if Firm 2 has infinite capacity, i.e., it has no congestion cost. To prove (14), we first note that irrespective of the value of C_2, for large enough Λ, $\lambda_1(\Lambda,C_2)+\lambda_2(\Lambda,C_2)=\Lambda$. Indeed, if $\lambda_1(\Lambda,C_2)+\lambda_2(\Lambda,C_2)<\Lambda$, then it is easy to see that the user base could further increase its payoff by increasing the arrival rate into Firm 1. Now, if $\lambda_1(\Lambda,C_2)=\Lambda$, the inequality (14) is trivially true. Assuming then that $\lambda_1(\Lambda,C_2)<\Lambda$, we note that $\lambda_1(\Lambda,C_2)$ is the solution of the following convex optimization.

$$\begin{array}{ll}\max & \psi(\lambda_1):=w_1\lambda_1+w_2(\Lambda-\lambda_1)-\frac{\lambda_1}{C_1-\lambda_1}-\frac{\Lambda-\lambda_1}{C_2-\Lambda+\lambda_1}\\ s.t. & 0\le\lambda_1\le\Lambda\end{array} \tag{15}$$

Now, it is clear that $\lambda_1(\Lambda,C_2)>0$, for large enough market size. This is because

$$\psi(0)\le w_2\Lambda$$
$$\le\psi(\Lambda)=w_1\Lambda-\sqrt{w_1\Lambda}+o(\sqrt{\Lambda})$$

for large enough Λ. Therefore, $\lambda_1(\Lambda,C_2)$ must satisfy the first order condition corresponding to (15):

$$w_1-w_2-\frac{C_1}{(C_1-\lambda_1)^2}+\frac{C_2}{(C_2-\Lambda+\lambda_1)^2}=0.$$

This in turn implies that $\lambda_1(\Lambda,C_2)$ satisfies

$$w_1-w_2-\frac{C_1}{(C_1-\lambda_1)^2}\le 0.$$

It then follows that $\lambda_1(\Lambda,C_2)\ge\lambda_1(\Lambda,\infty)$, since $\lambda_1(\Lambda,\infty)$ satisfies the first order condition corresponding to the following optimization.

$$\begin{array}{ll}\max & w_1\lambda_1+w_2(\Lambda-\lambda_1)-\frac{\lambda_1}{C_1-\lambda_1}\\ s.t. & 0\le\lambda_1\le\Lambda\end{array} \tag{16}$$

Having shown that $\lambda_1(\Lambda,C_2)\ge\lambda_1(\Lambda,\infty)$, it now remains to characterize $\lambda_1(\Lambda,\infty)$. Note that the objective function of (16) can be equivalently written as $\Delta\lambda_1-\frac{\lambda_1}{C_1-\lambda_1}$. This implies that $\lambda_1(\Lambda,\infty)$ can be characterized using our analysis

[6] Once again, we do not deal here with the issue of existence of an equilibrium.

of the single provider case. Since $\Delta < w_1$, it follows that

$$\begin{aligned}\lambda_1(\Lambda,\infty) &= C_1 - \sqrt{\frac{C_1}{\Delta}} \\ &= C^*_{w_1}(\Lambda) - \sqrt{\frac{C^*_{w_1}(\Lambda)}{\Delta}} \\ &= \Lambda + \sqrt{\frac{\Lambda}{w_1}} - \sqrt{\frac{\Lambda}{\Delta}} + o(\sqrt{\Lambda}).\end{aligned}$$

This completes the proof of (14).

Now, (14) implies that

$$\begin{aligned}\text{Profit of Firm 1} &= b_1\lambda_1(\Lambda, C_2) - C_1 \\ &\geq b_1\left[\Lambda + \sqrt{\frac{\Lambda}{w_1}} - \sqrt{\frac{\Lambda}{\Delta}} + o(\sqrt{\Lambda})\right] \\ &\quad - \left(\Lambda + \sqrt{\frac{\Lambda}{w_1}} + o(\sqrt{\Lambda})\right) \\ &= (b_1 - 1)\left(\Lambda + \sqrt{\frac{\Lambda}{w_1}}\right) - b_1\sqrt{\frac{\Lambda}{\Delta}} + o(\sqrt{\Lambda}).\end{aligned}$$

It follows then that the above lower bound on profit must hold at any equilibrium $(\lambda_1, C_1, \lambda_2, C_2)$, implying that

$$(b_1 - 1)\lambda_1 \geq (b_1 - 1)\left(\Lambda + \sqrt{\frac{\Lambda}{w_1}}\right) - b_1\sqrt{\frac{\Lambda}{\Delta}} + o(\sqrt{\Lambda}),$$

which implies that

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

This completes the proof. □

It is possible to extend the above result to the case of super-linear utilities, i.e., $\beta_i > 0$. Indeed, it can be shown that if $\beta_1 > \beta_2$, or $\beta_1 = \beta_2$ and $w_1 > w_2$, any equilibrium is necessarily a near-monopoly for Firm 1.

6. DISCUSSION AND CONCLUSIONS

We end with a discussion contrasting the results obtained in Sections 4 and 5, which have characterized the equilibria that emerge in settings capturing all mixtures of cooperative and non-cooperative users with industry-wide and firm-specific network effects.

Perhaps the most striking conclusion from these results is about the impact of competition on the user experience. The message that emerges is that *competition does not help*. 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 positive network effects make users more tolerant of congestion. Importantly, this is not a result of anarchy (non-cooperation among the users). Competition does not help even when the user base is cooperative. Thus, an important take away from the results is that adding competition does not help reduce congestion if firms cannot compete on prices. This is in contrast to settings such as [1,2] where price competition does reduce congestion.

Another striking message that emerges from the results in Sections 4 and 5 is about the impact of competition on market structure. In particular, the structure of the market, i.e., the market shares of the firms, is highly dependent on the type of service, i.e., 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. 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 help. Thus, for example, the advertising efficiency of Google may seem to give an advantage to GooglePlus over Facebook, but the models suggest that the impact of this is outweighed by the service quality comparison. Further, even minor differences in utilities are sufficient to ensure a monopoly. Again, the results for the cooperative setting highlight that these conclusions are not a result of anarchy in the user base, since they appear in cooperative model as well as the non-cooperative model.

The results in this paper provide important messages about the role of network effects, congestion, and competition in ad-supported services; however, further analytic work remains in order to assess the robustness and generality of the conclusions described above. In particular, we have focused on two classes of latency functions here, M/M/1 latencies and load based latencies, and it would be interesting to understand how the results are impacted when other latency functions are considered. The results in this paper have already highlighted that the form of the latency function can have an impact.

Additionally, this paper has focused entirely on the case of two competing firms. It is important to extend the analysis to larger markets. While it is intimidating to attempt exact analysis in the case of an arbitrary number of firms, it may be possible to study the asymptotic behavior of the market for a large number of firms.

Finally, we have considered two extreme forms of network effects when in reality many services experience a combination of both industry-wide and firm-specific network effects. It would be interesting to extend the analysis here to such a setting to understand how these types of network effects interact and, specifically, to understand when the interaction is such that multiple firms can coexist in the market and when the interaction is such that near-monopolies emerge.

Acknowledgements

The authors gratefully acknowledge the support of the NSF through grants CNS-1319820 and CNS-0846025. The first author also acklowledges support from an NWO VIDI grant.

7. REFERENCES

[1] J. Anselmi, D. Ardagna, J. C. Lui, A. Wierman, Y. Xu, and Z. Yang. The economics of the cloud: Price competition and congestion. In *Proceedings of NetEcon*, 2013.

[2] J. Anselmi, U. Ayesta, and A. Wierman. Competition yields efficiency in load balancing games. *Performance Evaluation*, 68(11):986–1001, 2011.

[3] R. Atar. A diffusion regime with non-degenerate slowdown, 2013. To appear.

[4] J. Buchanan. An economic theory of clubs. *Economica*, 32(125):1–14, 1965.

[5] D. Durkee. Why cloud computing will never be free. *Queue*, 8(4):20, 2010.

[6] J. Farrell and P. Klemperer. Coordination and lock-in: Competition with switching costs and network effects. *Handbook of Industrial Organization*, 3:1967–2072, 2007.
[7] J. Farrell and G. Saloner. Standardization, compatibility, and innovation. *The RAND Journal of Economics*, 16(1):70–83, 1985.
[8] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.
[9] J. Hamilton. The cost of latency, October 2009. URL:`http://perspectives.mvdirona.com/2009/10/31/TheCostOfLatency.asp`.
[10] IAB Internet Advertising Revenue Report, 2011.
[11] R. Johari and S. Kumar. Congestible services and network effects, 2009.
[12] M. Katz and C. Shapiro. Network externalities, competition, and compatibility. *The American Economic Review*, 75(3):424–440, 1985.
[13] R. Kohavi, R. Longbotham, D. Sommerfield, and R. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.
[14] S. Lohr. For impatient web users, an wye blink is just too long to wait. *New York Times*, 2012. Published Feb. 29.
[15] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi. Cloud computingthe business perspective. *Decision Support Systems*, 51(1):176–189, 2011.
[16] B. Metcalfe. Metcalfe's law: A network becomes more valuable as it reaches more users. *Infoworld*, 17(40):53–54, 1995.
[17] J. Nair, A. Wierman, and B. Zwart. Provisioning of large scale systems: The interplay between network effects and strategic behavior in the user base. In *Under submission*, 2013.
[18] A. Odlyzko and B. Tilly. A refutation of metcalfe's law and a better estimate for the value of networks and network interconnections, 2005.
[19] S. Oren and S. Smith. Critical mass and tariff structure in electronic communications markets. *The Bell Journal of Economics*, pages 467–487, 1981.
[20] J. Reed. The G/GI/N queue in the Halfin-Whitt regime. *Annals of Applied Probability*, 19:2211–2269, 2009.
[21] T. Sandler and J. Tschirhart. Club theory: Thirty years later. *Public Choice*, 93(3):335–355, 1997.
[22] A. Sundararajan. Network effects, nonlinear pricing and entry deterrence, 2003.

APPENDIX

A. PROOF OF LEMMA 2

For $\beta = 0$, the result follows directly from $g(\cdot)$ be function increasing in λ with $g(0) = 0$ and $\lim_{x\to 1} g(x) = +\infty$.

Next consider $\beta \in (0, 1)$. Since $g(\cdot)$ is an increasing convex function, its inverse $g^{-1}(\cdot)$ is also increasing but concave. Note also that $w\lambda^\beta$ is a concave function of λ for $\beta \in (0, 1]$. Therefore, by the composition rule for concave functions, $g^{-1}(w\lambda^\beta)$ is also a concave function.

Note that $g^{-1}(0) = 0$ by our assumptions on $g(\cdot)$. Therefore, using the sub-gradient inequality for concave function $h(\lambda) := g^{-1}(w\lambda^\beta)$, it follows that $h(\lambda)/\lambda$ is a decreasing function of λ with $\lim_{\lambda\to 0} h(\lambda)/\lambda = \infty$ and $\lim_{\lambda\to\infty} h(\lambda)/\lambda = 0$. Thus there is a unique $\tilde{\lambda}(C)$ such that $h(\lambda) = \lambda/C$ and this is the largest λ such that $h(\lambda) \geq \lambda/C$; by the earlier properties of $h(\cdot)$, $\lambda = 0$ automatically satisfies the inequality.

Finally consider the case of $\beta = 1$, it immediately follows that $w \leq g'(0)$ implies that $\lambda = 0$ is the only solution to $w\lambda^\beta \geq g(\frac{\lambda}{C})$ and otherwise there exists a unique positive λ that satisfies $w\lambda = g(\frac{\lambda}{C})$ which is also continuous and increasing in C. Thus, the result also follows for C sufficiently large when $\beta = 1$.

B. PROOF OF LEMMA 3

The choice of arrival rate is made by optimizing $\lambda\kappa(\lambda)$, where $\kappa(\lambda) = V(\lambda) - f(\lambda, C)$.

From Lemma 1, there is a $\lambda'(C)$ such that for all $\lambda \geq \lambda'(C)$, $\kappa(\lambda) \leq 0$. Let $\bar{\lambda}(C)$ be the largest λ in $[0, \lambda'(C)]$ such that $\kappa'(\lambda) = 0$.

Clearly $\tilde{\lambda}(C)$ is the $\lambda \in [\bar{\lambda}(C), \lambda'(C)]$ such that the derivative of $\lambda\kappa(\lambda)$ is 0.

C. PROOF OF LEMMA 4

For $\beta = 0$, the arrival rate is chosen to maximize a strictly concave function $w\lambda - \lambda g(\frac{\lambda}{C})$. It is easily verified from the first-order conditions that the maximizer is the unique non-negative solution to

$$w = \frac{\lambda}{C} g'\left(\frac{\lambda}{C}\right) + g\left(\frac{\lambda}{C}\right).$$

By differentiating the above equation, we can readily verify that $\tilde{\lambda}(C)$ is strictly increasing in C.

Next consider $\beta \in (0, 1)$. The arrival rate is chosen to be the maximizer of $\lambda\kappa(\lambda)$ where $\kappa(\lambda) = w\lambda^\beta - g(\frac{\lambda}{C})$ is a strictly concave function of λ.

From Lemma 2 it follows that $h(\lambda) \leq 0$ for $\lambda \geq \lambda'(C)$ where $\lambda'(C)$ is the unique positive solution to $w\lambda^\beta = g(\frac{\lambda}{C})$.

Additionally, $\kappa(\cdot)$ is maximized at $\bar{\lambda}(C) \in (0, \tilde{\lambda}(C))$ where $h(\bar{\lambda}(C)) > 0$ and $\bar{\lambda}(C)$ is the unique positive solution to

$$w\beta\lambda^{\beta-1} = \frac{\lambda}{C} g'\left(\frac{\lambda}{C}\right) + g\left(\frac{\lambda}{C}\right).$$

The derivative of the objective function of the cooperative scenario traffic split is given by

$$\lambda\kappa'(\lambda) + \kappa(\lambda),$$

which is positive at $\bar{\lambda}(C)$ and negative at $\tilde{\lambda}(C)$, and by taking a derivative it can be verified that the second derivative is negative. Therefore, the objective function is strictly concave between $\bar{\lambda}(C)$ and $\lambda'(C)$. Thus, the unique optimizer is the solution in $[\bar{\lambda}(C), \lambda'(C)]$ of $\lambda\kappa'(\lambda) + \kappa(\lambda) = 0$, i.e.,

$$w(1+\beta)\lambda^\beta = \frac{\lambda}{C} g'\left(\frac{\lambda}{C}\right) + g\left(\frac{\lambda}{C}\right). \tag{17}$$

Finally we note that $\lambda\kappa(\lambda)$ is increasing in λ for $\lambda \in [0, \bar{\lambda}(C)]$. Therefore, $\tilde{\lambda}(C)$ is also the unique positive solution to (17). For $\beta = 1$, following Lemma 2 if C is large enough, then same proof can be used.

Differentiating in (17) and rearranging terms, we obtain

$$\frac{\partial\tilde{\lambda}(C)}{\partial C} = \frac{\left(\frac{\tilde{\lambda}(C)}{C}\right)^2 \left[2g'\left(\frac{\tilde{\lambda}(C)}{C}\right) + \frac{\tilde{\lambda}(C)}{C} g''\left(\frac{\tilde{\lambda}(C)}{C}\right)\right]}{\frac{\tilde{\lambda}(C)}{C}\left[2g'\left(\frac{\tilde{\lambda}(C)}{C}\right) + \frac{\tilde{\lambda}(C)}{C} g''\left(\frac{\tilde{\lambda}(C)}{C}\right)\right] - w(1+\beta)\tilde{\lambda}^\beta(C)\beta}.$$

Using (17) to substitute for $w(1+\beta)\tilde{\lambda}^\beta(C)$ the denominator

is now

$$\left(\frac{\tilde{\lambda}(C)}{C}\right)^2 g''\left(\frac{\tilde{\lambda}(C)}{C}\right) + (1-\beta)\frac{\tilde{\lambda}(C)}{C} g'\left(\frac{\tilde{\lambda}(C)}{C}\right)$$
$$+ \frac{\tilde{\lambda}(C)}{C} g'\left(\frac{\tilde{\lambda}(C)}{C}\right) - \beta g\left(\frac{\tilde{\lambda}(C)}{C}\right),$$

which in turn is positive as $\beta \in [0,1]$, $g(\cdot)$ is convex and increasing with $g(0) = 0$. This proves that $\tilde{\lambda}(C)$ strictly increases in C.