Provisioning of large scale systems: The interplay between network effects and strategic behavior in the user base

Jayakrishnan Nair
Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, 91125, ujk@caltech.edu

Adam Wierman
Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, 91125, adamw@caltech.edu

Bert Zwart
Centrum voor Wiskunde en Informatica, 1098 SJ Amsterdam, The Netherlands, Bert.Zwart@cwi.nl

In this paper, we consider the problem of capacity provisioning for an online service supported by advertising. We analyse the strategic interaction between the service provider and the user base in this setting, modeling positive network effects, as well as congestion sensitivity in the user base. We focus specifically on the influence of positive network effects, as well as the impact of non-cooperative behavior in the user base on the firm's capacity provisioning decision and its profit. Our analysis reveals that stronger positive network effects, as well as non-cooperation in the user base, drive the service into a more congested state and lead to increased profit for the service provider. However, the impact of non-cooperation, or 'anarchy' in the user base strongly dominates the impact of network effects.

1. Introduction

The internet today offers a wide range of online services. Implementing these services typically requires considerable computing infrastructure, consisting of an extremely large number of servers (Vanderbilt 2009). Therefore, how much (computing) capacity to provision is a crucial decision for the service provider. Over provisioning enhances the user-perceived quality of the service, but is also expensive. Therefore, the service provider must strategically provision the correct number of servers to maximize its profit. The goal of this paper is to provide insight into this capacity provisioning decision.

In exploring the capacity provisioning of online systems, there are three features of the online services themselves that are of particular importance.

First, since a majority of online services are offered for free to the end user, the firm (or service provider) is *deriving its revenue via advertising.* Corporations like Google and Facebook make billions of dollars in revenue annually by offering advertising supported online services (Internet Advertising Bureau 2011).

Second, many online services allow for interaction between users. As a result, these services exhibit *strong positive network effects*, i.e., users obtain an increased utility from other people using the same service (Katz and Shapiro 1985, Farrell and Klemperer 2007, Johari and Kumar 2009). Examples of such services abound: social networking applications, online gaming environments, document editing services, and many others. Indeed, network effects are believed to be a primary driver of usage growth for many services.

Third, users of online services today are *highly delay sensitive* (Hamilton 2009, Lohr 2012). Even a small additional delay in accessing a service can adversely affect the user perceived quality of the service, potentially leading to a decline in usage, and thus a decline in revenue for the service provider. For example, an experiment by Google showed that adding 500 milli-seconds of delay to its search results resulted in a 20% drop in revenue (Kohavi et al. 2009).

Clearly, the capacity provisioning decision for online services is influenced by the interplay of the three factors discussed above. The objective of a service provider is to maximize profit, taking into account the revenue from advertising as well as the cost of managing its computing infrastructure. On the other hand, users care about maximizing their own payoff, which depends on the utility derived from using the service as well as the disutility due to congestion or delay. Therefore, the emergent capacity provisioning decision and the popularity of the service result from a strategic interaction between the service provider and the user base.

1.1. Our results

In this paper, we study the problem of optimal capacity provisioning for a firm operating an advertising supported online service. We model both network effects and congestion sensitivity of the user base, and analyze the number of servers (i.e., the capacity of computing infrastructure) the firm must provision to maximize its profit as the volume of the user base (or the market size) scales to infinity. A key feature of our model is that the traffic scaling regime is endogenous. That is, users in the user base use the service or not depending on the congestion and network effects, which is determined by the capacity provisioning of a profit-maximizing firm. This endogenous traffic scaling is in contrast to the majority of work on large scale systems in queueing theory, which tends to impose a scaling exogenously, e.g., Halfin and Whitt (1981), Reed (2009), Atar (2012).[1]

The key focus of this paper is to understand the impact of two factors on the firm's capacity provisioning decision: (i) the strength of positive network effects in the user base, and (ii) non-cooperative behavior in the user base, i.e., users independently seeking to maximize their own payoff. We study the impact of the latter by analyzing two different models of the behavior of the user population: a non-cooperative model, in which users independently pursue their own interest, and a cooperative model, in which the user base seeks to maximize the aggregate social payoff.

Intuitively, we would expect that stronger positive network effects would make the user base more tolerant to congestion, allowing the service provider to the run the service with fewer servers, and thus make a higher profit. Similarly, we would expect that a lack of cooperation in the user base would lead to a higher utilization of the service (tragedy of the commons), leading to higher profit for the service provider. Our analysis supports these intuitions, but reveals some surprises with respect to the relative impact of network effects and non-cooperation in the user base.

Our analysis shows that as the market size becomes large, the profit maximizing strategy for the service provider involves operating the service in heavy traffic while still having almost the full potential market base using the service, for both the cooperative and the non-cooperative population models. This is made possible by the statistical economies of scale inherent in large queueing systems: the firm can run its servers at a high utilization, and simultaneously provide good quality of service to users. Moreover, our analysis shows that stronger positive network effects lead to increased profit for the service provider. This is because the service provider can exploit the additional utility users derive from aggregation to operate the service at a higher level of congestion, thus saving on server costs.

However, the cooperative and the non-cooperative model differ in the extent to which network effects influence the capacity provisioning decision and the profit made by the firm. Under the cooperative model, we show (see Section 4.1) that as the market size becomes large, the strength of the positive network effects impacts the number of servers provisioned by the service provider in the order sense. As a result, network effects strongly influence the emergent heavy-traffic regime

[1] In the context of the literature on exogenous traffic scalings in queueing theory, the current paper provides insight about which scalings may emerge endogenously from the interaction between a service provider and its user base.

and the profit made by the firm. On the other hand, under the non-cooperative model, we show (see Section 4.2) that the very absence of coordination in the user base drives the system into an extreme heavy-traffic regime in which the firm provisions only a bounded number of servers more than the minimum number required to serve the full potential user base. Remarkably, this happens irrespective of how strong the network effects are. This 'tragedy of the commons' effect implies that the impact of network effects on the capacity provisioning decision and the firm's profit is significantly diminished, compared to the scenario in which the user base behaves cooperatively.

In other words, our results suggest that while network effects and non-cooperation in the user base are both profitable for the service provider, the impact of non-cooperation in the user base strongly dominates the impact of network effects.

The remainder of this paper is organized as follows. We review related literature in Section 2. We introduce our model and notation in Section 3. We state and interpret our results, for both the cooperative and the non-cooperative model of the user base in Section 4. Finally, we conclude in Section 5.

2. Related Literature

There are two distinct streams of literature related to this work: literature from the queueing domain, and literature focused on network effects and their consequences.

Within the queueing literature, there is a large body of work analyzing queueing systems where the arrival rate of jobs as well as the number of servers scale to infinity. Depending on how the arrival rate and the number of servers scale relative to one another, different heavy traffic regimes are possible. One well studied scaling regime is the so-called Halfin-Whitt regime (Halfin and Whitt 1981), in which the number of servers equals the minimum number required to stably support the arrival rate, plus a 'spare' that is proportional to the square root of the arrival rate. There are many other scaling regimes that are studied too; see, for instance, Halfin and Whitt (1981), Reed (2009), Atar (2012) and the references therein. In all of this work the scaling is imposed exogenously.

In contrast, very few papers take the approach of deriving an endogenous scaling regime that emerges naturally in the considered setting, as we do in this paper. One work of this type is Borst et al. (2004), which considers the problem of optimal staffing in a call center in an asymptotic regime where the call arrival rate is exogenously scaled to infinity. Other papers that focus on endogenous scalings (including this one) take the approach of scaling only the potential arrival rate to infinity. The actual arrival rate is a function of the price of the service, and/or the level of congestion. Papers in this category include Whitt (2003), Kumar and Randhawa (2010), Maglaras and Zeevi (2003), Randhawa and Kumar (2008). However, none of the above mentioned papers consider network effects or the comparison between cooperative and non-cooperative user bases, as we do in the current work.

A second body of literature that is related to the current paper studies network effects and its consequences. In this space, one line of work proposes scaling laws for the aggregate value of a network of connected users; for example, see Metcalfe (1995), Odlyzko and Tilly (2005). Another line of work focuses on firms and users interacting in a market setting, in which the utility of a user consuming a product/service increases with the number of users consuming the same, or a compatible product/service. Representative papers in this category include Oren and Smith (1981), Farrell and Saloner (1985), Katz and Shapiro (1985), Sundararajan (2003), Farrell and Klemperer (2007). However, these papers do not deal with services involving resource sharing between users. As a result, they do not consider congestion, which is a key component of our model.

There is one body of work that does considers network effects and congestion, the literature on club theory. See Sandler and Tschirhart (1997) for a survey. The theory of clubs, which originated from Buchanan (1965), deals with groups of congestion sensitive users sharing a certain resource. Indeed, the setting in this paper can be interpreted as a club good offered by a profit maximizing firm. However, a key distinction between our work and the previous work in club theory is that we consider an advertising supported service (i.e., the revenue of the service provider does not come from payments from users). Moreover, we use an explicit queueing model of the service to model congestion, something that is not typically done in this literature.

3. Model Overview

In this section, we describe our model for the interaction between a profit maximizing firm (service provider) and a congestion sensitive user base. In our model, the firm implements the service by operating a cluster of servers, which serve user requests. We assume that there is a known market size, which determines the maximum possible usage of the service. The actual usage depends on the utility that the service provides to the user base, as well as the congestion (or delay) experienced by the user base in accessing the service. The firm derives a revenue proportional to the usage of the service, which is characteristic of services that are supported by advertising, and incurs a cost proportional to the number of servers provisioned. The firm decides the number of servers to provision so as to maximize its own profit.

Formally, let k denote the number of servers provisioned by the firm. Let Λ denote the maximum possible arrival rate of requests for the service; Λ thus characterizes the market size. User requests arrive according to a Poisson process with rate $\widehat{\lambda}_\Lambda(k) \leq \Lambda$. These requests are served in a First-Come-First-Served manner by a system with k parallel servers and a single queue. The processing times of requests are independent and exponentially distributed with mean $1/\mu$. Without loss of generality, we take $\mu = 1$. Note that the function $\widehat{\lambda}_\Lambda(k)$ defines the level of 'usage' of the service, and thus models the behavior of the user base. We now formally describe our behavioral model of the user base.

3.1. Model of the user base

In this paper, we study two models for the behavior of the user base: a non-cooperative model, in which each user acts independently and in her own self interest, and a cooperative model, in which the usage level is set so as to maximize the aggregate social payoff. Comparing the results for these two models helps us understand the impact of users acting independently and in their own interest, i.e., the impact of 'anarchy'. We now formally describe the two models.

Non-cooperative population model. The non-cooperative model postulates the following functional form for $\widehat{\lambda}_\Lambda(k)$:

$$\widehat{\lambda}_\Lambda(k) = \max\left\{\lambda \in [0,\Lambda] \mid V(\lambda) - \xi(\lambda,k) \geq 0\right\}. \tag{1}$$

Here, $V(\cdot)$ is the utility derived by a single (infinitesimal) user from using the service, as a function of the overall usage (or arrival rate) λ seen by the service. $\xi(\lambda,k)$ is an indicator of the steady state delay experienced by a typical request for service, and represents the disutility experienced by a user due to congestion. We set $\xi(\lambda,k) = \infty$ for $\lambda \geq k$, since the queueing system is unstable in this case. Therefore, Equation (1) can be interpreted as follows. A single (infinitesimal) user receives a payoff equal to $V(\lambda) - \xi(\lambda,k)$ if she chooses to use the service, and zero payoff if she chooses not to. Thus, the overall usage level in Equation (1) corresponds to a Wardrop equilibrium between the users with respect to their individual payoffs.

Clearly, network effects determine the form of $V(\cdot)$. Specifically, no network effects would imply that $V(\cdot)$ is a constant. On the other hand, positive network effects would imply that $V(\cdot)$ is a non-decreasing function, i.e., the utility derived by a single user grows as the overall usage of the service grows. In this paper, for simplicity, we take $\xi(\lambda,k) = \mathbb{E}\left[W(\lambda,k)\right]$ for $\lambda < k$, where $W(\lambda,k)$ is the stationary waiting time experienced by a request.

Cooperative population model. In the cooperative model, we set the usage level of the service so as to maximize the net social payoff. Accordingly, we take $U(\lambda) := \lambda V(\lambda)$ to be the net utility derived by the user base at usage level λ. Similarly, we take $\lambda\xi(\lambda,k)$ to be the net disutility experienced by the user base on account of congestion. Therefore, the cooperative model postulates the following functional form for $\widehat{\lambda}_\Lambda(k)$:

$$\widehat{\lambda}_\Lambda(k) := \max\left\{\arg\max_{\lambda\in[0,\Lambda]}\left[U(\lambda) - \lambda\xi(\lambda,k)\right]\right\}. \tag{2}$$

Note that if there were no network effects, we would expect $U(\lambda)$ to grow linearly. Positive network effects would cause $U(\lambda)$ to grow superlinearly. We now turn to the behavioral model for the firm.

3.2. Model of the firm

By provisioning k servers, the firm derives revenue $b_1\widehat{\lambda}_\Lambda(k)$, and incurs cost $b_2 k$ per unit time. Without loss of generality, we set $b_2 = 1$. The profit maximizing firm naturally provisions capacity so as to maximize its profit. Specifically, the number of servers provisioned is given by

$$k^*_\Lambda := \max\left\{\arg\max_{k}\left[b_1\widehat{\lambda}_\Lambda(k) - k\right]\right\},$$

and the corresponding request arrival rate is given by

$$\lambda^*_\Lambda := \widehat{\lambda}_\Lambda(k^*_\Lambda).$$

The tuple $(\lambda^*_\Lambda, k^*_\Lambda)$ characterizes the equilibrium between the firm and the user base. Since $\widehat{\lambda}_\Lambda(k) < k$, a necessary condition for the firm to make positive profit is $b_1 > 1$. Since the case $b_1 \leq 1$ is uninteresting (the firm will simply not operate in this case), we assume hereafter that $b_1 > 1$.

4. Results

The goal of this paper is to provide insight into the interplay between network effects and strategic behavior in the user base. As such, we provide theorems characterizing the equilibrium $(\lambda^*_\Lambda, k^*_\Lambda)$ resulting from the interaction between a profit maximizing service provider and the congestion sensitive user base. In particular, we study the behavior of the equilibrium as the possible market size grows, i.e., as Λ scales to infinity for both the non-cooperative and the cooperative population model. Our results highlight the role played by non-cooperative behavior among users, network effects, and economies of scale in the regime of large market size. We start with the cooperative model and then move to the non-cooperative model.

4.1. Cooperative population model

We begin by studying the cooperative population model, defined in Equation (2). Our results, summarized in Theorem 1 below, highlight the strong influence of network effects on the capacity provisioning decision as well as the profit made by the firm. In particular, the network effects change the *order of magnitude* of the scaling that emerges for the equilibrium.

The statement of Theorem 1 makes the following technical assumptions on the functional form of $U(\cdot)$.

Assumption 1. *$U : \mathbb{R}_+ \to \mathbb{R}_+$ is continuously differentiable over $[0,\infty)$ with $U(0) = 0$, $\lim_{\lambda\to\infty} U(\lambda) = \infty$. $U'(\cdot)$ satisfies the following properties.*

(a) There exists $\bar{\lambda} \geq 0$ such that $U'(\cdot)$ is non-decreasing over $[\bar{\lambda},\infty)$.

(b) $\lim_{\lambda\to\infty} \frac{U'(\lambda)}{\lambda}$ exists.

(c) $\lim_{\lambda\to\infty} \frac{U'(\lambda+\nu)}{U'(\lambda)} = 1 \quad \forall\ \nu > 0.$

Condition (a) above states that $U(\lambda)$ is convex for large λ. This allows us to capture positive network effects. (b) and (c) are regularity assumptions. Note that Assumption 1 implies that

$$\alpha := \lim_{\lambda\to\infty} U'(\lambda) \in (0,\infty) \cup \{\infty\}.$$

Define $\omega := \lim_{\lambda\to\infty} \frac{U'(\lambda)}{\lambda}$. We are now ready to state our theorem. The proof of the theorem is given in Appendix A.

THEOREM 1. *Consider the cooperative model of the user base and suppose that Assumption 1 holds. Then for large enough* Λ, $\lambda^*_\Lambda \in [\Lambda-2,\Lambda]$. *Further, as* $\Lambda \uparrow \infty$, *the optimal capacity provisioning is the following.*

(i) If $\alpha \in (0,\infty)$, *then*

$$k^*_\Lambda = \Lambda + \sqrt{\beta(\alpha)\Lambda} + o(\sqrt{\Lambda}),$$

where $\beta(\alpha) \in (0,\infty)$ *is a strictly decreasing function of* α.

(ii) If $\alpha = \infty$, *and* $\omega = 0$, *then*

$$k^*_\Lambda = \Lambda + \sqrt{\frac{\Lambda}{U'(\Lambda)}} + o\left(\sqrt{\frac{\Lambda}{U'(\Lambda)}}\right).$$

(iii) If $\alpha = \infty$, *and* $\omega \in (0,\infty) \cup \{\infty\}$, *then*

$$k^*_\Lambda = \Lambda + \frac{1}{\omega} + e(\Lambda) + o(1),$$

where $e(\Lambda) \in [-2, 2b_1]$.[2]

The three cases in the theorem correspond to different growth rates of the aggregate social utility $U(\cdot)$. Specifically, Case (i) corresponds to an asymptotically linear growth, Case (ii) corresponds to an asymptotically super-linear, but sub-quadratic growth, and Case (iii) corresponds to an asymptotically quadratic/super-quadratic growth. Thus, we consider progressively stronger positive network effects in Cases (i) through (iii). We now highlight the key insights from Theorem 1.

Firstly, it is easy see that in all three cases,

$$\lim_{\Lambda\to\infty} \frac{\lambda^*_\Lambda}{k^*_\Lambda} = 1.$$

This means that it is asymptotically optimal for the profit maximizing firm to operate in heavy traffic, even though the user base is congestion sensitive. This is because as the market size becomes large, the statistical economies of scale associated with large multi-server systems allow the firm to operate the service at high utilization, and still provide a good quality of service (Whitt 2003, Borst et al. 2004, Kumar and Randhawa 2010). Moreover, the profit maximizing strategy for the firm is to provision enough capacity so as to attract (almost) the full potential market base.

Next, we observe that the heavy-traffic regime that emerges in our model, as well as the profit made by the firm, depend critically on the growth rate of the social utility $U(\cdot)$. Intuitively, if the social utility is greater, the firm can attract the full potential market base by provisioning fewer servers, thereby making a higher profit. In other words, stronger positive network effects make the user base more tolerant to congestion, allowing the firm to operate the service with fewer servers.

[2] If $\omega = \infty$, then $\frac{1}{\omega}$ is understood to be zero.

Case (i) of Theorem 1 corresponds to an asymptotically linear growth of social utility $U(\cdot)$. This means, roughly, that the per user utility $V(\cdot)$ is asymptotically constant. In this case, the optimal operating regime for the firm is the well known Halfin-Whitt regime; the firm provisions the minimum capacity to serve the full market size Λ, plus a 'spare capacity' approximately proportional to $\sqrt{\Lambda}$ servers. It is interesting to note that in this regime, the expected stationary waiting time decays as $\Theta\left(1/\sqrt{\Lambda}\right)$ as the market size Λ grows to infinity (Halfin and Whitt 1981). This means that as the market size becomes large, the congestion disutility experienced by users approaches zero. Finally, under Case (i), the profit of the firm is given by

$$(b_1-1)\Lambda-\sqrt{\beta(\alpha)\Lambda}-o(\sqrt{\Lambda}). \tag{3}$$

The above equation may be interpreted as follows. Intuitively, $(b_1-1)\Lambda$ can be interpreted as the maximum possible profit for the service provider. Indeed, if the user base was not congestion sensitive, the service provider could have attracted the maximum possible usage of Λ by provisioning the minimum number of servers required to maintain stability, i.e., Λ. Equation (3) implies that the service provider makes a profit $\Theta(\sqrt{\Lambda})$ less than the maximum possible due to the congestion sensitivity of the user base.

Case (ii) of Theorem 1 corresponds to an asymptotically super-linear, but sub-quadratic growth of $U(\cdot)$. This means, roughly, that the per user utility $V(\cdot)$ grows sub-linearly. In this case, the optimal operating regime for the firm is a 'heavier' traffic regime than the Halfin-Whitt regime: the firm provisions a 'spare capacity' of approximately $\sqrt{\frac{\Lambda}{U'(\Lambda)}}$ servers. Note that under Case (ii), $U'(\Lambda)=o(\Lambda)$, and $U'(\Lambda)\to\infty$ as $\Lambda\to\infty$. Therefore, that the number of spare servers under Case (ii) grows to infinity, but slower than under Case (i). As a result, the expected stationary waiting time decays as $\Theta\left(\sqrt{\frac{U'(\Lambda)}{\Lambda}}\right)$ as the market size Λ increases to infinity (Halfin and Whitt 1981). This means that as the market size becomes large, the congestion disutility experienced by users approaches zero, but at a slower rate than under Case (i). Finally, the profit of the firm under Case (ii) is given by

$$(b_1-1)\Lambda-\sqrt{\frac{\Lambda}{U'(\Lambda)}}-o\left(\sqrt{\frac{\Lambda}{U'(\Lambda)}}\right).$$

Note that profit is greater than that under Case (i); the firm makes a profit that is $\Theta\left(\sqrt{\frac{\Lambda}{U'(\Lambda)}}\right)$ less than the maximum possible.

Case (iii) of Theorem 1 corresponds to a quadratic/super-quadratic growth of $U(\cdot)$. This means, roughly, that the per user utility $V(\cdot)$ grows linearly/super-linearly. In this case, the firm operates the system in a very heavy-traffic regime; it only needs to provision a bounded number of 'spare servers.' As a result, as the market size becomes large, the congestion disutility experienced by users does not approach zero, but remains bounded below by a positive constant. Finally, under Case (iii), the firm makes the most profit: $(b_1-1)\Lambda-O(1)$. Thus, when the network effects are as strong as under Case (iii), the firm makes the maximum possible profit, short of a bounded amount.

To summarize, our results for the cooperative model of the user base reveal that network effects strongly influence the capacity provisioning decision, as well as the profit of the firm. As the network effects become stronger, the firm provisions fewer spare servers, users experience a higher congestion disutility, and the firm makes a greater profit.

4.2. Non-cooperative population model

We now consider the non-cooperative model of the user base, defined by Equation (1). As in the previous section, we analyze the firm's capacity provisioning decision, the congestion experienced by the user base, as well as the profit of the service provider, in the asymptotic regime of large market size. In contrast to the case of cooperative users, for non-cooperative users, the impact of network effects is significantly diminished – network effects no longer have an order-of-magnitude impact on the scaling of the equilibrium. Our main result, stated in Theorem 2 below, relies on the following technical assumption on the function $V(\cdot)$.

Assumption 2. *$V:\mathbb{R}_+ \to \mathbb{R}_+$ is continuous. There exists $\bar{\lambda} \geq 0$ such that $V(\cdot)$ is non-decreasing over $[\bar{\lambda}, \infty)$, and $\lim_{\lambda \to \infty} V(\lambda) \in (0, \infty) \cup \{\infty\}$.*

The above assumption simply states that the per user utility $V(\lambda)$ is non-decreasing for large enough λ. One would of course expect this assumption to hold if there are no network effects or if there are positive network effects. Define $v := \lim_{\lambda \to \infty} V(\lambda)$. We are now ready to state our theorem. The proof of this theorem is given in Appendix B.

Theorem 2. *Consider the non-cooperative model of the user base and suppose that Assumption 2 holds. Then for large enough Λ, $\lambda^*_\Lambda \in [\Lambda - 2, \Lambda]$. Further, as $\Lambda \uparrow \infty$, the optimal capacity provisioning satisfies*

$$k^*_\Lambda = \Lambda + \frac{1}{v} + \tilde{e}(\Lambda) + o(1),$$

where $\tilde{e}(\Lambda) \in [-2, 2b_1]$.[3]

Theorem 2 states that under the non-cooperative model of the user base, it is optimal for the firm to provision enough capacity to attract (almost) the full potential user base, similar to the cooperative model. Moreover, it is asymptotically optimal for the firm to operate the service in an extremely heavy traffic regime: with just a bounded number of spare servers. Remarkably, this is true irrespective of the strength of the network effects. In particular, the firm provisions a bounded number of spare servers even when there are no network effects (recall that $V(\cdot)$ remains constant in this case). In contrast, under the cooperative model, such a heavy traffic regime emerges only when the positive network effects are extremely strong (see Case (*iii*) of Theorem 1). In other words, under the cooperative model, extremely strong network effects are required to drive the service into the high-congestion regime with a bounded number of spare servers. On the other hand, under the non-cooperative model, the absence of coordination among users leads to a tragedy of the commons effect, driving the service into a similar high-congestion regime. As a result, the impact of network effects is diminished. Indeed, network effects do not influence the emergent scaling regime in the order sense.

As per Theorem 2, as the market size becomes large, the number of spare servers provisioned is approximately equal to $\frac{1}{v} + \tilde{e}(\Lambda)$. Note that the component $\frac{1}{v}$ decreases as the network effects get stronger. The term $\tilde{e}(\Lambda)$ is an artifact of the discrete nature of the optimization performed by the service provider. Since the number of spare servers remains bounded, the expected stationary waiting time experienced by users does not approach zero as the market size grows to infinity, but remains bounded below by a positive constant. Finally, we see that the profit of the service provider grows with the market size as $(b_1 - 1)\Lambda - O(1)$, implying the service provider makes the maximum possible profit, short of a bounded amount. Note that network effects only influence this bounded component. Indeed, it is easy to show that stronger network effects lead to a higher profit via a reduction in the value of this (bounded) component.

To summarize, when users independently pursue their own interest, the ‘anarchy’ in the user base drives the service into a highly congested regime, in which the service provider only provisions a bounded spare capacity. As a result, the service provider makes the maximum possible profit (in the order sense), irrespective of the strength of the network effects. Furthermore, the impact of network effects on the capacity provisioning decision and the profit of the service provider is significantly diminished, compared to the scenario where the user base behaves cooperatively.

[3] If $v = \infty$, then $\frac{1}{v}$ is understood to be zero.

5. Conclusion

In this paper, we consider the problem of capacity provisioning for an online service supported by advertising. We analyze the strategic interaction between the service provider and the user base in this setting, modeling positive network effects, as well as congestion sensitivity in the user base. We focus specifically on the influence of positive network effects, as well as non-cooperative behavior in the user base on the firm's capacity provisioning decision, and its profit.

Our analysis provides rigorous justification for the intuition that both stronger positive network effects and non-cooperative behavior tend to drive the service into a more congested state, leading to increased profit for the service provider. Furthermore, our analysis highlights the fact that the impact of non-cooperation, or 'anarchy', in the user base strongly dominates the impact of network effects.

Additionally, our results have impact for the literature studying large-system scalings of multi-server systems. In particular, such work typically imposes scalings exogenously, e.g., Halfin and Whitt (1981), Reed (2009), Atar (2012). Our work derives scalings that occur endogenously as a result of the interaction between a profit maximizing firm and a congestion sensitive user base with network effects. Thus, our results can provide a guide for the queueing literature on which scalings are (or are not) appropriate for a given setting.

Appendix A: Proof of Theorem 1

To prove Theorem 1, we first analyse the following 'unconstrained' multi-server scaling regime parameterised by the number of servers k. Define

$$\tilde{\lambda}(k) := \max\left\{\arg\max_{\lambda \geq 0}\big[U(\lambda) - \lambda\mathbb{E}\left[W\right]\big]\right\},$$
$$\tilde{\rho}(k) := \frac{\tilde{\lambda}(k)}{k}.$$

We prove Theorem 1 by establishing a connection between the evolution of $(\lambda^*_\Lambda, k^*_\Lambda)$ as $\Lambda \uparrow \infty$ and $(\tilde{\lambda}(k), k)$ as $k \uparrow \infty$. The following lemma characterizes the evolution of the tuple $(\tilde{\lambda}(k), k)$.

Lemma 1. *Suppose Assumption 1 holds. Then $\{\tilde{\lambda}(k)\}$ is a non-decreasing sequence. As $k \uparrow \infty$, $\tilde{\lambda}(k) \uparrow \infty$ as follows.*

(i) If $\alpha \in (0, \infty)$, then

$$\lim_{k\to\infty} k(1-\tilde{\rho}(k))^2 = \beta(\alpha), \tag{4}$$

where $\beta(\alpha) \in (0, \infty)$ is a strictly decreasing function of α.

(ii) If $\alpha = \infty$, then

$$\lim_{k\to\infty} kU'(\tilde{\lambda}(k))(1-\tilde{\rho}(k))^2 = 1. \tag{5}$$

Moreover, for large enough k, $\tilde{\rho}(k)$ is strictly increasing.

We defer the proof of Lemma 1 to later in this section, and use it first to prove Theorem 1. Lemma 1 allows us to define the following inverse of $\{\tilde{\lambda}(k)\}$. Taking $\tilde{\lambda}(0) := 0$, we define $\tilde{k} : \mathbb{R}_+ \to \mathbb{Z}_+$ as follows.

$$\tilde{k}(\lambda) := \max\{k \in \mathbb{Z}_+ \mid \tilde{\lambda}(k) \leq \lambda\}.$$

Since $\tilde{\lambda}(k) \overset{k\uparrow\infty}{\to} \infty$, $\tilde{k}(\lambda)$ is well defined for all $\lambda \in \mathbb{R}_+$. Moreover, $\tilde{k}(\lambda)$ is a non-decreasing function with $\tilde{k}(\lambda) \overset{\lambda\uparrow\infty}{\to} \infty$.

We are now ready to state the connection between $(\lambda^*_\Lambda, k^*_\Lambda)$ and $\{\tilde{\lambda}(k)\}$. Intuitively, the following lemma shows that $k^*_\Lambda \approx \tilde{k}(\Lambda)$, and $\lambda^*_\Lambda \approx \tilde{\lambda}(\tilde{k}(\Lambda))$.

LEMMA 2. *For large enough* Λ,

$$\tilde{k}(\Lambda) \leq k^*_\Lambda \leq \tilde{k}(\Lambda) + 2b_1, \text{ and} \tag{6}$$
$$\Lambda - 2 \leq \tilde{\lambda}(\tilde{k}(\Lambda)) \leq \lambda^*_\Lambda \leq \Lambda. \tag{7}$$

Proof. Note that for $k \leq \tilde{k}(\Lambda)$, $\tilde{\lambda}(k) \leq \Lambda$, implying that $\widehat{\lambda}_\Lambda(k) = \tilde{\lambda}(k)$. Therefore, for $k \leq \tilde{k}(\Lambda)$,

$$b_1\widehat{\lambda}_\Lambda(k) - k = b_1\tilde{\lambda}(k) - k = k(b_1\tilde{\rho}(k) - 1).$$

Now from Lemma 1, we know that for large enough k, $\tilde{\rho}(k)$ is strictly increasing, and $\tilde{\rho}(k) \overset{k\uparrow\infty}{\to} 1$. This means that for large enough k, $\tilde{\rho}(k) > 1/b_1$. Therefore, for large enough k, $k(b_1\tilde{\rho}(k) - 1)$ is strictly increasing with respect to k, and $k(b_1\tilde{\rho}(k) - 1) \overset{k\uparrow\infty}{\to} \infty$. This in turn implies that for large enough Λ,

$$\max\left\{\arg\max_{k\leq\tilde{k}(\Lambda)}\left[b_1\widehat{\lambda}_\Lambda(k) - k\right]\right\} = \max\left\{\arg\max_{k\leq\tilde{k}(\Lambda)}\left[b_1\tilde{\lambda}(k) - k\right]\right\} = \tilde{k}(\Lambda),$$

which implies that

$$k^*_\Lambda \geq \tilde{k}(\Lambda). \tag{8}$$

Lemma 1 implies that for large enough k, $\tilde{\lambda}(k)$ is strictly increasing, and

$$\lim_{k\to\infty} \tilde{\lambda}(k+1) - \tilde{\lambda}(k) = 1.$$

This implies that for large enough Λ,

$$\tilde{\lambda}(\tilde{k}(\Lambda) + 1) - \tilde{\lambda}(\tilde{k}(\Lambda)) < 2.$$

Also, by the definition of $\tilde{k}(\cdot)$,

$$\tilde{\lambda}(\tilde{k}(\Lambda)) \leq \Lambda < \tilde{\lambda}(\tilde{k}(\Lambda) + 1).$$

Combining the above two equations, we conclude that for large enough Λ,

$$\tilde{\lambda}(\tilde{k}(\Lambda)) \leq \Lambda < \tilde{\lambda}(\tilde{k}(\Lambda)) + 2. \tag{9}$$

Now, by provisioning $\tilde{k}(\Lambda)$ servers, the firm would derive profit $b_1\tilde{\lambda}(\tilde{k}(\Lambda)) - \tilde{k}(\Lambda)$. By definition, the profit from provisioning k^*_Λ servers can only be greater. Therefore,

$$b_1\tilde{\lambda}(\tilde{k}(\Lambda)) - \tilde{k}(\Lambda) \leq b_1\lambda^*_\Lambda - k^*_\Lambda \tag{10}$$
$$\Rightarrow\ 0 \leq k^*_\Lambda - \tilde{k}(\Lambda) \leq b_1\left(\lambda^*_\Lambda - \tilde{\lambda}(\tilde{k}(\Lambda))\right) \leq b_1\left(\Lambda - \tilde{\lambda}(\tilde{k}(\Lambda))\right) < 2b_1. \tag{11}$$

The first inequality in (11) uses (8), the last inequality uses (9). Note that (11) implies (6). Moreover, (11) implies that $\lambda^*_\Lambda \geq \tilde{\lambda}(\tilde{k}(\Lambda))$, which, in combination with (9) implies (7). □

We are now ready to prove the statements of Theorem 1. Lemma 2 implies that for large enough Λ, $\lambda^*_\Lambda \in [\Lambda - 2, \Lambda]$. Also, Lemma 2 tells us that for large enough Λ,

$$\begin{aligned} e_1(\Lambda) &:= \Lambda - \tilde{\lambda}(\tilde{k}(\Lambda)) \in [0, 2], \\ e_2(\Lambda) &:= k^*_\Lambda - \tilde{k}(\Lambda) \in [0, 2b_1]. \end{aligned} \tag{12}$$

We show now that (12) and Lemma 1 imply Statement (ii) of Theorem 1. Statements (i) and (iii) can be proved on similar lines. Let us assume therefore, that $\alpha = \infty$, and $\lim_{\lambda\to\infty} \frac{U'(\lambda)}{\lambda} = 0$. In the following, we use the notation $f(\Lambda) \sim g(\Lambda)$ to mean that $\lim_{\Lambda\to\infty} \frac{f(\Lambda)}{g(\Lambda)} = 1$. Since $\tilde{k}(\lambda) \overset{\lambda\uparrow\infty}{\to} \infty$, Statement (ii) of Lemma 1 implies that

$$\begin{aligned}
&\lim_{\Lambda\to\infty} \tilde{k}(\Lambda)U'(\tilde{\lambda}(\tilde{k}(\Lambda)))\left(1-\frac{\tilde{\lambda}(\tilde{k}(\Lambda))}{\tilde{k}(\Lambda)}\right)^2 = 1\\
\Rightarrow\ & \tilde{k}(\Lambda)\left(1-\frac{\tilde{\lambda}(\tilde{k}(\Lambda))}{\tilde{k}(\Lambda)}\right) \sim \sqrt{\frac{\tilde{k}(\Lambda)}{U'(\tilde{\lambda}(\tilde{k}(\Lambda)))}}\\
\Rightarrow\ & \tilde{k}(\Lambda)\left(1-\frac{\tilde{\lambda}(\tilde{k}(\Lambda))}{\tilde{k}(\Lambda)}\right) \sim \sqrt{\frac{\tilde{\lambda}(\tilde{k}(\Lambda))}{U'(\tilde{\lambda}(\tilde{k}(\Lambda)))}}\\
\Rightarrow\ & k^*_\Lambda - \Lambda - e_2(\Lambda) + e_1(\Lambda) \sim \sqrt{\frac{\Lambda - e_1(\Lambda)}{U'(\Lambda - e_1(\Lambda))}}.
\end{aligned}$$

The second implication above uses the fact that $\tilde{\lambda}(\tilde{k}(\Lambda)) \sim \tilde{k}(\Lambda)$. Since, for large enough Λ, $e_1(\Lambda)$ is bounded, the uniform convergence theorem (Bingham et al. 1989, Chapter 1.2) implies that

$$\sqrt{\frac{\Lambda - e_1(\Lambda)}{U'(\Lambda - e_1(\Lambda))}} \sim \sqrt{\frac{\Lambda}{U'(\Lambda)}}.$$

Therefore, we have

$$\begin{aligned}
& k^*_\Lambda - \Lambda - e_2(\Lambda) + e_1(\Lambda) \sim \sqrt{\frac{\Lambda}{U'(\Lambda)}}\\
\Rightarrow\ & k^*_\Lambda = \Lambda + \sqrt{\frac{\Lambda}{U'(\Lambda)}} + o\left(\sqrt{\frac{\Lambda}{U'(\Lambda)}}\right) + e_2(\Lambda) - e_1(\Lambda).
\end{aligned}$$

Note that our assumption on $U(\cdot)$ implies that $\sqrt{\frac{\Lambda}{U'(\Lambda)}} \overset{\Lambda\uparrow\infty}{\to} = \infty$. Therefore,

$$e_2(\Lambda) - e_1(\Lambda) = o\left(\sqrt{\frac{\Lambda}{U'(\Lambda)}}\right),$$

implying Statement (ii) of Theorem 1. Statements (i) and (iii) of the theorem can be proved similarly.

To complete the proof of Theorem 1, it remains to prove Lemma 1. The remainder of this section is devoted to this proof.

Proof of Lemma 1 The proof uses three main steps:

1. We first show that $\{\tilde{\lambda}(k)\}$ is a non-decreasing sequence, and $\tilde{\lambda}(k) \uparrow \infty$ as $k \uparrow \infty$. Let $f(\lambda, k) := U(\lambda) - \lambda\mathbb{E}[W]$. It is easy to see that there exists $k_0 \in \mathbb{N}$ such that $\tilde{\lambda}(k_0) > 0$. Invoking Lemma 3) below, which proves that f is supermodular, it follows that for $0 \le \lambda < \tilde{\lambda}(k_0)$,

$$f(\tilde{\lambda}(k_0), k_0+1) - f(\lambda, k_0+1) > f(\tilde{\lambda}(k_0), k_0) - f(\lambda, k_0) \ge 0.$$

This implies that $\tilde{\lambda}(k_0+1) \ge \tilde{\lambda}(k_0)$. Proceeding inductively, we conclude that the sequence $\{\tilde{\lambda}(k)\}$ is non-decreasing, and therefore must have a limit. For the purpose of obtaining a contradiction, assume that $\lim_{k\to\infty} \tilde{\lambda}(k) = \nu < \infty$. Pick $\lambda_1 > \nu$ such that $U(\lambda_1) > \max_{0\le\lambda\le\nu} U(\lambda)$. Since $f(\lambda_1, k) \uparrow U(\lambda_1)$ as $k \uparrow \infty$, there exists $k_1 \in \mathbb{N}$ such that $U(\lambda_1) > f(\lambda_1, k_1) > \max_{0\le\lambda\le\nu} U(\lambda)$. This means that

$$f(\lambda_1, k_1) > U(\tilde{\lambda}(k_1)) \ge f(\tilde{\lambda}(k_1), k_1),$$

which is a contradiction. Therefore, $\lim_{k\to\infty} \tilde{\lambda}(k) = \infty$.

2. Next, we show that if $U(\cdot)$ is eventually convex, then for large enough k, $\tilde{\rho}(k)$ is strictly increasing.

Let $\rho := \lambda/k$. Let $C(\lambda,k)$ denote the stationary probability of waiting in an $M/M/k$ queue with arrival rate λ, and a mean service time of 1. Then $\mathbb{E}[W(\lambda,k)] = \frac{C(\lambda,k)}{k-\lambda}$. We know that for large enough k, $\tilde{\lambda}(k) > 0$. Since f is continuously differentiable wrt λ, for large enough k, $\tilde{\lambda}(k)$ satisfies

$$\begin{aligned}
U'(\lambda) &= \frac{\partial}{\partial\lambda}\left(\frac{\lambda C(\lambda,k)}{k-\lambda}\right) \\
&= \frac{kC(\lambda,k)}{(k-\lambda)^2} + \frac{\lambda}{(k-\lambda)}\frac{\partial C(\lambda,k)}{\partial\lambda} \\
&= \frac{kC(\lambda,k)}{(k-\lambda)^2} + \frac{\lambda}{(k-\lambda)}\left[\frac{(1-\rho)C(\lambda,k)}{\rho} + \frac{C(\lambda,k)(1-C(\lambda,k))}{k(1-\rho)}\right] \\
&= \frac{C(\lambda,k)}{k(1-\rho)^2} + \frac{C(\lambda,k)}{} + \frac{\rho C(\lambda,k)(1-C(\lambda,k))}{k(1-\rho)^2} \\
&= \frac{C(\lambda,k)}{k(1-\rho)} + \frac{C(\lambda,k)}{} + \frac{\rho C(\lambda,k)(2-C(\lambda,k))}{k(1-\rho)^2} \qquad (13) \\
&=: h(\rho,k).
\end{aligned}$$

The above calculation uses Lemma 7. The function h has the following properties.

(i) For fixed $\rho \in [0,1)$, $h(\rho,k)$ is a strictly decreasing function of k. This follows from the fact that $C(k\rho,k)$ is a strictly decreasing function of k.

(ii) For fixed k, $h(\rho,k)$ is a strictly increasing function of ρ. This follows from the fact that $C(k\rho,k)$ is a strictly increasing function of ρ.

Consider k large enough so that $\tilde{\lambda}(k) > \bar{\lambda}$. Since $U(\cdot)$ is convex over $[\bar{\lambda},\infty)$, and $\tilde{\lambda}(k+1) \geq \tilde{\lambda}(k)$, we conclude that $U'(\tilde{\lambda}(k+1)) \geq U'(\tilde{\lambda}(k))$. Therefore,

$$h(\tilde{\rho}(k+1),k+1) = U'(\tilde{\lambda}(k+1)) \geq U'(\tilde{\lambda}(k)) = h(\tilde{\rho}(k),k) > h(\tilde{\rho}(k),k+1),$$

where the last inequality follows from Property (i) above. Since

$$h(\tilde{\rho}(k+1),k+1) > h(\tilde{\rho}(k),k+1),$$

Property (ii) above implies that $\tilde{\rho}(k+1) > \tilde{\rho}(k)$.

3. We now prove the statements (4) and (5). Let $\theta(k) := kU'(\tilde{\lambda}(k))(1-\tilde{\rho}(k))^2$. Note that the sequence $\{\theta(k)\}$ must have a limit point in $[0,\infty]$. We first rule out ∞ and 0 as possible limit points, and then show that the limit point is unique.

For large enough k, $U'(\tilde{\lambda}(k)) > 0$. Therefore, from (13), for large enough k, $\tilde{\lambda}(k)$ satisfies

$$\begin{aligned}
1 &= \frac{C(\lambda,k)}{kU'(\lambda)(1-\rho)} + \frac{C(\lambda,k)}{U'(\lambda)} + \frac{\rho C(\lambda,k)(2-C(\lambda,k))}{kU'(\lambda)(1-\rho)^2} \\
&=: T_1 + T_2 + T_3. \qquad (14)
\end{aligned}$$

By possibly restricting to a subsequence of $\{\theta(k)\}$, let us assume (for the sake of obtaining a contradiction) that $\lim_{k\to\infty}\theta(k) = \infty$. In (14), it is easy to see that $\lim_{k\to\infty} T_1 = 0$ and $\lim_{k\to\infty} T_3 = 0$. We now show that $\lim_{k\to\infty} T_2 = 0$. If $\alpha := \lim_{\lambda\uparrow\infty} U'(\lambda) = \infty$, then this is obvious. If $\alpha \in (0,\infty)$, then $\lim_{k\to\infty} k(1-\tilde{\rho}(k))^2 = \infty$. This corresponds to the *quality driven regime*, and Proposition 1 of Halfin and Whitt (1981) implies that $\lim_{k\to\infty} C(\tilde{\lambda}(k),k) = 0$. This in turn implies $\lim_{k\to\infty} T_2 = 0$. Since the right hand side of (14) approaches 0 as $k \uparrow \infty$, we have a contradiction. Therefore, ∞ is not a limit point of $\{\theta(k)\}$.

Next, by possibly restricting to a subsequence of $\{\theta(k)\}$, let us assume (for the sake of obtaining a contradiction) that $\lim_{k\to\infty}\theta(k)=0$. In this case, $\lim_{k\to\infty}k(1-\tilde{\rho}(k))^2=0$. This corresponds to the *efficiency driven regime*, and Proposition 1 of Halfin and Whitt (1981) implies that $\lim_{k\to\infty}C(\tilde{\lambda}(k),k)=1$. Therefore, in (14), $\lim_{k\to\infty}T_3=\infty$. Since T_1 and T_2 are non-negative sequences, this gives us a contradiction. Therefore, 0 is not a limit point of $\{\theta(k)\}$.

Since ∞ and 0 are not limit points of $\{\theta(k)\}$, there exists a limit point $\gamma\in(0,\infty)$. We now show that this limit point is unique. Consider the following two cases.

Case 1: $\alpha=\infty$. Let us restrict to a subsequence of $\{\theta(k)\}$ that converges to $\gamma\in(0,\infty)$. Along this subsequence, $k(1-\tilde{\rho}(k))^2\to 0$, which (as we have seen before) implies $C(\tilde{\lambda}(k),k)\to 1$. Therefore, along this subsequence,

$$T_1\to 0,\quad T_2\to 0,\quad T_3\to\frac{1}{\gamma}$$

(see (14)). This implies $\gamma=1$, which proves (5).

Case 2: $\alpha\in(0,\infty)$. Let us restrict to a subsequence of $\{\theta(k)\}$ that converges to $\gamma\in(0,\infty)$. Along this subsequence, $k(1-\tilde{\rho}(k))^2\to\beta$, where $\beta:=\frac{\gamma}{\alpha}$. This corresponds to the *quality-efficiency driven regime*, and Proposition 1 of Halfin and Whitt (1981) implies $C(\tilde{\lambda}(k),k)\to\psi(\beta)$, where

$$\psi(x):=\left[1+\sqrt{2\pi}x\Phi(x)e^{x^2/2}\right]^{-1}.$$

Therefore, along this subsequence,

$$T_1\to 0,\quad T_2\to\frac{\psi(\beta)}{\alpha},\quad T_3\to\frac{\psi(\beta)(2-\psi(\beta))}{\beta\alpha}$$

(see (14)). Therefore, we must have

$$\alpha=\psi(\beta\alpha)+\frac{\psi(\beta)(2-\psi(\beta))}{\beta}=:\xi(\beta). \tag{15}$$

As we prove below in Lemma 4, $\xi(\cdot)$ is strictly decreasing, with $\lim_{x\downarrow 0}\xi(x)=\infty$, $\lim_{x\uparrow\infty}\xi(x)=0$. Therefore, there is a unique $\beta(\alpha)\in(0,\infty)$ that satisfies (15). Moreover, it is clear that $\beta(\alpha)$ is strictly decreasing with respect to α. This proves (5).

This completes the proof of Lemma 1. □

Lemma 3. *$f(\lambda,k):=U(\lambda)-\lambda\mathbb{E}[W]$ is supermodular, i.e., for $0\le\lambda_1<\lambda_2<k_1<k_2$,*

$$f(\lambda_2,k_1)-f(\lambda_1,k_1)<f(\lambda_2,k_2)-f(\lambda_1,k_2).$$

Proof. Let $w(\lambda,k):=\mathbb{E}[W]$. For $\lambda<k$,

$$\begin{aligned}\frac{\partial w(\lambda,k)}{\partial\lambda}&=\frac{\partial}{\partial\lambda}\left(\frac{C(\lambda,k)}{k-\lambda}\right)\\&=\frac{C(\lambda,k)}{(k-\lambda)^2}+\frac{1}{(k-\lambda)}\left[\frac{(1-\rho)C(\lambda,k)}{\rho}+\frac{C(\lambda,k)(1-C(\lambda,k))}{k(1-\rho)}\right]\\&=\frac{C(\lambda,k)}{(k-\lambda)^2}+\frac{C(\lambda,k)}{\lambda}+\frac{C(\lambda,k)(1-C(\lambda,k)}{(k-\lambda)^2}\\&=\frac{C(\lambda,k)}{\lambda}+\frac{C(\lambda,k)(2-C(\lambda,k)}{(k-\lambda)^2}.\end{aligned}$$

Noting that $C(\lambda,k)$ decreases wrt. k, and since the function $x(2-x)$ is increasing in $[0,1]$, we conclude that $\frac{\partial w(\lambda,k)}{\partial\lambda}$ is decreasing in k. This implies

$$\begin{aligned} w(\lambda_2,k_1)-w(\lambda_1,k_1) &\geq w(\lambda_2,k_2)-w(\lambda_1,k_2) \\ \Rightarrow \lambda_2\left(w(\lambda_2,k_1)-w(\lambda_2,k_2)\right) &> \lambda_1\left(w(\lambda_1,k_1)-w(\lambda_1,k_2)\right) \\ \Rightarrow f(\lambda_2,k_2)-f(\lambda_2,k_1) &> f(\lambda_1,k_2)-f(\lambda_1,k_1). \end{aligned}$$

The final inequality above is equivalent to the statement of the lemma. □

LEMMA 4. *$\xi(\cdot)$ as defined in (15) is strictly decreasing, with $\lim_{x\downarrow 0}\xi(x)=\infty$, $\lim_{x\uparrow\infty}\xi(x)=0$.*

Proof. Since $\psi(\cdot)$ is a strictly decreasing function, and $x(2-x)$ is increasing in $[0,1]$, it is clear that $\xi(\cdot)$ is strictly decreasing. Moreover, $\lim_{x\downarrow 0}\psi(x)=1$ implies that $\lim_{x\downarrow 0}\xi(x)=\infty$, and $\lim_{x\to\infty}\psi(x)=0$ implies that $\lim_{x\to\infty}\xi(x)=0$. □

Appendix B: Proof of Theorem 2

To prove Theorem 2, we first analyse the following ‘unconstrained’ multi-server scaling regime parameterised by the number of servers k. Define

$$\begin{aligned} \tilde{\lambda}(k) &:= \max\{\lambda\geq 0 \mid V(\lambda)-\mathbb{E}[W]\geq 0\} \\ \tilde{\rho}(k) &:= \frac{\tilde{\lambda}(k)}{k}. \end{aligned}$$

We prove Theorem 2 by establishing a connection between the evolution of $(\lambda^*_\Lambda, k^*_\Lambda)$ as $\Lambda\uparrow\infty$ and $(\tilde{\lambda}(k),k)$ as $k\uparrow\infty$. The following lemma characterizes the evolution of the tuple $(\tilde{\lambda}(k),k)$.

LEMMA 5. *Suppose Assumption 2 holds. Then $\{\tilde{\lambda}(k)\}$ is a non-decreasing sequence. As $k\uparrow\infty$, $\tilde{\lambda}(k)\uparrow\infty$ such that*

$$\lim_{k\to\infty} kV(\tilde{\lambda}(k))(1-\tilde{\rho}(k))=1. \tag{16}$$

Moreover, for large enough k, $\tilde{\rho}(k)$ is strictly increasing.

Proof. First, we show that $\{\tilde{\lambda}(k)\}$ is non-decreasing with $\lim_{k\to\infty}\tilde{\lambda}(k)=\infty$. Let $f(\lambda,k):=V(\lambda)-\mathbb{E}[W(\lambda,k)]$ (note that we are highlighting the dependence of W on λ and k). Since $f(\cdot,k)$ is continuous, $f(0,k)\geq 0$, and $\lim_{\lambda\uparrow k} f(\lambda,k)=-\infty$, it is easy to see that $\tilde{\lambda}(k)$ satisfies

$$\begin{aligned} &f(\tilde{\lambda}(k),k)=0, \\ &f(\lambda,k)<0 \quad \forall \quad \lambda\in(\tilde{\lambda}(k),k). \end{aligned} \tag{17}$$

Since $\lim_{\lambda\to\infty} V(\lambda)>0$, it is easy to show that there exists $k_0\in\mathbb{N}$ such that $\tilde{\lambda}(k_0)>0$. Since $f(\tilde{\lambda}(k_0),k_0)=0$, note that $f(\tilde{\lambda}(k_0),k_0+1)>0$. It follows then from (17) that $\tilde{\lambda}(k_0+1)>\tilde{\lambda}(k_0)$. Proceeding inductively, we conclude that the sequence $\{\tilde{\lambda}(k)\}$ is non-decreasing, and hence has a limit. For the purpose of obtaining a contradiction, assume that $\lim_{k\to\infty}\tilde{\lambda}(k)=\nu<\infty$. Pick $\lambda_1>\nu$ satisfying $V(\lambda_1)>0$. Since $f(\lambda_1,k)\overset{k\uparrow\infty}{\to}V(\lambda_1)>0$, there exists $k_1\in\mathbb{N}$ such that $f(\lambda_1,k_1)>0$. From (17), this implies that $\tilde{\lambda}(k_1)>\lambda_1$, which is a contradiction. Therefore, $\lim_{k\to\infty}\tilde{\lambda}(k)=\infty$.

Next, we prove that (16) holds. Let $C(\lambda,k)$ denote the stationary probability of waiting in an $M/M/k$ queue with arrival rate λ, and a mean service time of 1. Recall that $\mathbb{E}[W(\lambda,k)]=\frac{C(\lambda,k)}{k-\lambda}$. Let $\theta(k):=kV(\tilde{\lambda}(k))(1-\tilde{\rho}(k))$. From (17), we know that

$$\begin{aligned} V(\tilde{\lambda}(k)) &= \mathbb{E}\left[W(\tilde{\lambda}(k),k)\right]=\frac{C(\tilde{\lambda}(k),k)}{k-\tilde{\lambda}(k)}, \\ \Rightarrow\quad \theta(k) &= C(\tilde{\lambda}(k),k). \end{aligned} \tag{18}$$

(18) implies that

$$\limsup_{k\to\infty} kV(\tilde{\lambda}(k))(1-\tilde{\rho}(k)) \le 1.$$

Since $\lim_{k\to\infty} V(\tilde{\lambda}(k)) > 0$, we conclude that $(1-\tilde{\rho}(k)) \overset{k\uparrow\infty}{\to} 0$. Therefore,

$$\begin{aligned} &\lim_{k\to\infty} kV(\tilde{\lambda}(k))(1-\tilde{\rho}(k))^2 = 0, \\ \Rightarrow\ &\lim_{k\to\infty} k(1-\tilde{\rho}(k))^2 = 0. \end{aligned}$$

This corresponds to the *efficiency driven regime*, and Proposition 1 of Halfin and Whitt (1981) implies that $\lim_{k\to\infty} C(\tilde{\lambda}(k),k) = 1$, which implies (using (18)) that $\lim_{k\to\infty} \theta(k) = 1$.

Finally, it remains to show that for large enough k, $\tilde{\rho}(k)$ is strictly increasing. Consider k large enough so that $\tilde{\rho}(k) > 0$. In this case, (18) implies that $\tilde{\rho}(k)$ satisfies

$$h(\rho,k) := \frac{kV(k\rho)(1-\rho)}{C(k\rho,k)} = 1.$$

The function h has the following properties.

(i) For fixed $\rho \in (0,1)$, $h(\rho,k)$ is a strictly increasing function of k. This follows from the fact that $C(k\rho,k)$ is a strictly decreasing function of k.

(ii) For fixed k, $h(\rho,k)$ is a strictly decreasing function of $\rho \in (0,1)$. This follows from the fact that $C(k\rho,k)$ is a strictly increasing function of ρ.

Now, since $h(\tilde{\rho}(k),k) = 1$, Property (i) above implies that $h(\tilde{\rho}(k),k+1) > 1$. Then, Property (ii) above implies that $\tilde{\rho}(k+1) > \tilde{\rho}(k)$. This completes the proof. □

The remainder of the proof of Therorem 2 proceeds on similar lines as the proof of Theorem 1. Lemma 5 allows us to define the following inverse of $\{\tilde{\lambda}(k)\}$. Taking $\tilde{\lambda}(0) := 0$, we define $\tilde{k} : \mathbb{R}_+ \to \mathbb{Z}_+$ as follows.

$$\tilde{k}(\lambda) := \max\{k \in \mathbb{Z}_+ \mid \tilde{\lambda}(k) \le \lambda\}.$$

Since $\tilde{\lambda}(k) \overset{k\uparrow\infty}{\to} \infty$, $\tilde{k}(\lambda)$ is well defined for all $\lambda \in \mathbb{R}_+$. Moreover, $\tilde{k}(\lambda)$ is a non-decreasing function with $\tilde{k}(\lambda) \overset{\lambda\uparrow\infty}{\to} \infty$.

We now state the connection between $(\lambda^*_\Lambda, k^*_\Lambda)$ and $\{\tilde{\lambda}(k)\}$. Intuitively, the following lemma shows that $k^*_\Lambda \approx \tilde{k}(\Lambda)$, and $\lambda^*_\Lambda \approx \tilde{\lambda}(\tilde{k}(\Lambda))$.

LEMMA 6. *For large enough* Λ,

$$\tilde{k}(\Lambda) \le k^*_\Lambda \le \tilde{k}(\Lambda) + 2b_1, \tag{19}$$

$$\Lambda - 2 \le \tilde{\lambda}(\tilde{k}(\Lambda)) \le \lambda^*_\Lambda \le \Lambda. \tag{20}$$

The proof is identical to that of Lemma 2. We are now ready to prove the statements of Theorem 2. Lemma 6 implies that for large enough Λ, $\lambda^*_\Lambda \in [\Lambda-2,\Lambda]$. Also, Lemma 6 tells us that for large enough Λ,

$$\begin{aligned} e_1(\Lambda) &:= \Lambda - \tilde{\lambda}(\tilde{k}(\Lambda)) \in [0,2], \\ e_2(\Lambda) &:= k^*_\Lambda - \tilde{k}(\Lambda) \in [0,2b_1]. \end{aligned} \tag{21}$$

In the following, we use the notation $f(\Lambda) \sim g(\Lambda)$ to mean that $\lim_{\Lambda\to\infty} \frac{f(\Lambda)}{g(\Lambda)} = 1$. Now, Lemma 5 implies that

$$\begin{aligned} &\tilde{k}(\Lambda) - \tilde{\lambda}(\tilde{k}(\Lambda)) \sim \frac{1}{V(\tilde{\lambda}(\tilde{k}(\Lambda)))} \\ \Rightarrow\ &k^*_\Lambda - \Lambda - e_2(\Lambda) + e_1(\Lambda) \sim \frac{1}{V(\Lambda - e_1(\Lambda))} \\ \Rightarrow\ &k^*_\Lambda - \Lambda - e_2(\Lambda) + e_1(\Lambda) = \frac{1}{v} + o(1) \\ \Rightarrow\ &k^*_\Lambda = \Lambda + \frac{1}{v} + \tilde{e}(\Lambda) + o(1), \end{aligned}$$

where $\tilde{e}(\Lambda) := e_2(\Lambda) - e_1(\Lambda)$. This completes our proof.

Appendix C: Erlang C properties

Consider the $M/M/k$ queue with arrival rate λ and mean service time 1. Let $\rho := \frac{\lambda}{k}$. Let $C(\lambda, k)$ denote the stationary probability of waiting.

LEMMA 7. *For* $0 < a < k$,

$$\frac{\partial C(\lambda,k)}{\partial a} = \frac{(1-\rho)C(\lambda,k)}{\rho} + \frac{C(\lambda,k)(1-C(\lambda,k))}{k(1-\rho)}. \tag{22}$$

Proof. Let $B(a,k)$ denote the Erlang B blocking probability. $C(\lambda,k)$ can be expressed in terms of $B(a,k)$ as follows (Whitt 2002).

$$C(\lambda,k) = \frac{B(a,k)}{1-\rho+\rho B(a,k)}. \tag{23}$$

Moreover, the partial derivative of $B(a,k)$ with respect to a is given by (see Whitt (2002))

$$\frac{\partial B(a,k)}{\partial a} = B(a,k)\left[\rho^{-1} - 1 + B(a,k)\right]. \tag{24}$$

(22) can be derived easily using (23) and (24). □

References

Atar, R. 2012. A diffusion regime with non-degenerate slowdown. To appear.

Bingham, N.H., C.M. Goldie, J.L. Teugels. 1989. *Regular variation*, vol. 27. Cambridge University Press.

Borst, S, A. Mandelbaum, M.I. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34.

Buchanan, J.M. 1965. An economic theory of clubs. *Economica* **32**(125) 1–14.

Farrell, J., P. Klemperer. 2007. Coordination and lock-in: Competition with switching costs and network effects. *Handbook of Industrial Organization* **3** 1967–2072.

Farrell, J., G. Saloner. 1985. Standardization, compatibility, and innovation. *The RAND Journal of Economics* **16**(1) 70–83.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3) 567–588.

Hamilton, J. 2009. The cost of latency. URL:`http://perspectives.mvdirona.com/2009/10/31/TheCostOfLatency.asp`.

Internet Advertising Bureau. 2011. IAB Internet Advertising Revenue Report. URL `http://www.iab.net/adrevenuereport`.

Johari, R., S. Kumar. 2009. Congestible services and network effects.

Katz, M.L., C. Shapiro. 1985. Network externalities, competition, and compatibility. *The American Economic Review* **75**(3) 424–440.

Kohavi, R., R. Longbotham, D. Sommerfield, R.M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* **18**(1) 140–181.

Kumar, S., R.S. Randhawa. 2010. Exploiting market size in service systems. *Manufacturing & Service Operations Management* **12**.

Lohr, S. 2012. For impatient web users, an wye blink is just too long to wait. *New York Times* URL `http://www.nytimes.com/2012/03/01/technology/impatient-web-users-flee-slow-loading-sites.html`. Published Feb. 29.

Maglaras, C, A Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* **49**(8).

Metcalfe, B. 1995. Metcalfe's law: A network becomes more valuable as it reaches more users. *Infoworld* **17**(40) 53–54.

Odlyzko, A., B. Tilly. 2005. A refutation of metcalfe's law and a better estimate for the value of networks and network interconnections.

Oren, S.S., S.A. Smith. 1981. Critical mass and tariff structure in electronic communications markets. *The Bell Journal of Economics* 467–487.

Randhawa, R.S., S Kumar. 2008. Usage restriction and subscription services: Operational benefits with rational users. *Manufacturing & Service Operations Management* **10**(3) 429–447.

Reed, J. 2009. The G/GI/N queue in the Halfin-Whitt regime. *Annals of Applied Probability* **19** 2211–2269.

Sandler, T., J. Tschirhart. 1997. Club theory: Thirty years later. *Public Choice* **93**(3) 335–355.

Sundararajan, A. 2003. Network effects, nonlinear pricing and entry deterrence.

Vanderbilt, T. 2009. Data center overload. *New York Times Magazine* URL `http://www.nytimes.com/2009/06/14/magazine/14search-t.html`. Published June 8.

Whitt, W. 2002. Solutions for the Erlang B and C formulas. URL `www.columbia.edu/~ww2040/ErlangBandCFormulas.pdf`. Class notes for IEOR 6707: Advanced topics in queueing theory: Focus on customer call centers.

Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. *Operations Research* **51**(4) 531–542.