Management Science

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

Provisioning of Large-Scale Systems: The Interplay Between Network Effects and Strategic Behavior in the User Base

Jayakrishnan Nair, Adam Wierman, Bert Zwart

To cite this article:
Jayakrishnan Nair, Adam Wierman, Bert Zwart (2016) Provisioning of Large-Scale Systems: The Interplay Between Network Effects and Strategic Behavior in the User Base. Management Science 62(6):1830-1841. http://dx.doi.org/10.1287/mnsc.2015.2210

MANAGEMENT SCIENCE
Vol. 62, No. 6, June 2016, pp. 1830–1841
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2015.2210

Provisioning of Large-Scale Systems: The Interplay Between Network Effects and Strategic Behavior in the User Base

Jayakrishnan Nair
Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, 400076, India, jayakrishnan.nair@ee.iitb.ac.in

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

Bert Zwart
Centrum voor Wiskunde en Informatica, 1098 XG 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 noncooperative 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 noncooperation 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 noncooperation, or "anarchy" in the user base strongly dominates the impact of network effects.

Keywords: online services; capacity provisioning; network effects; large-scale queueing systems
History: Received August 7, 2012; accepted February 8, 2015, by Noah Gans, stochastic models and simulation. Published online in *Articles in Advance* November 20, 2015.

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. Overprovisioning enhances the user-perceived quality of the service, but is also expensive. Therefore, the service provider must strategically provision the correct service capacity 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 (Interactive 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 2010). 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 milliseconds 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 because of 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 we analyze the service capacity (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), and 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) noncooperative 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 noncooperative model, in which users independently pursue their own interest, and a cooperative model, in which the user base seeks to maximize its aggregate 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 noncooperation 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 the full potential market base using the service, for both the cooperative and the noncooperative 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 noncooperative 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 §4.1) that as the market size becomes large, the strength of the positive network effects impacts the service capacity provisioned by the service provider in the order sense. As a result, network effects strongly influence the emergent heavy-traffic regime and the profit made by the firm. On the other hand, under the noncooperative model, we show (sed §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 capacity more than the minimum 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 although network effects and noncooperation in the user base are both profitable for the service provider, the impact of noncooperation 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 §2. We introduce our model and notation in §3. We state and interpret our results, for both the cooperative and the noncooperative model of the user base in §4. Finally, we conclude in §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

[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.

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), and 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), and Randhawa and Kumar (2008). However, none of the above mentioned papers consider network effects or the comparison between cooperative and noncooperative user bases, as we do in the current work. Moreover, none of these papers study an advertising-supported service model (i.e., the revenue of the service provider does not come from payments from users), which is the dominant model for online services today.

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) and 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), and 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 consider 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. 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 operates a queueing system that serves 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 service capacity provisioned. The firm decides the service capacity to provision so as to maximize its own profit.

3.1. Model for Congestion

Formally, let k denote the service capacity provisioned by the firm. In the following, we interpret k as either the capacity/speed of a single server, or as the number of servers of unit capacity provisioned. Let $\xi(\lambda, k)$ denote the disutility experienced by a single user because of congestion, when the arrival rate of user requests for the service equals λ. We consider two models for the function $\xi(\lambda, k)$, corresponding to the two interpretations of service capacity k. Our results, presented in §4, apply to both models.

3.1.1. $M/M/1$ Model. As the name suggests, in this model, we take the user disutility $\xi(\lambda, k)$ to be the average (stationary) response time in an $M/M/1$ queue, with arrival rate λ and server speed k. Without loss of generality, we assume that user requests have a unit service requirement on average. Thus, we have

$$\xi(\lambda, k) = \begin{cases} \dfrac{1}{k-\lambda} & \text{for } \lambda < k, \\ \infty & \text{for } \lambda \geq k. \end{cases}$$

3.1.2. $M/M/k$ Model. In this model, we take the user disutility $\xi(\lambda, k)$ to be the average (stationary) waiting time in an $M/M/k$ queue with arrival rate λ. Without loss of generality, we assume a unit service

capacity for each server, and that user requests have a unit service requirement on average. Thus, we have

$$\xi(\lambda, k) = \begin{cases} \dfrac{C(\lambda, k)}{k-\lambda} & \text{for } \lambda < k, \\ \infty & \text{for } \lambda \geq k, \end{cases} \tag{1}$$

where $C(\lambda, k)$ is the Erlang waiting probability.[2] For simplicity, and to allow us to state our results for both congestion models together, we permit k to take values in $\mathbb{R}_+$, using the following well-known analytical extension of the Erlang C formula (Jagers and Van Doorn 1991):

$$C(\lambda, k) = \left[\int_0^\infty \lambda t e^{-\lambda t}(1+t)^{k-1}\, dt\right]^{-1}. \tag{2}$$

It is well known that when k is a positive integer, and $\lambda < k$, the above formula agrees with the classical Erlang C formula. We note that in the large market-size regimes we study in this paper, treating the service capacity k to be a real variable is not a limitation. Indeed, it is easy to show that the correct "integral" service capacity can be obtained from the real-valued k we derive by rounding. To summarize, in the $M/M/k$ model, we define the congestion disutility $\xi(\lambda, k)$ via (1), where $C(\lambda, k)$ is defined in (2).

3.2. Model of the User Base

In this paper, we study two models for the behavior of the user base: a noncooperative 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 payoff of the user base. 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.

3.2.1. Noncooperative Population Model. The noncooperative model postulates the following functional form for $\hat{\lambda}_\Lambda(k)$:

$$\hat{\lambda}_\Lambda(k) = \max\{\lambda \in [0, \Lambda] \mid V(\lambda) - \xi(\lambda, k) \geq 0\}. \tag{3}$$

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. As defined earlier, $\xi(\lambda, k)$ is the disutility derived by a user because of the congestion experienced in accessing the service. Therefore, Equation (3) 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 (3) corresponds to a Wardrop equilibrium between the users with respect to their individual payoffs.[3]

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 nondecreasing function, i.e., the utility derived by a single user grows as the overall usage of the service grows.

3.2.2. Cooperative Population Model. The cooperative model postulates the following functional form for $\hat{\lambda}_\Lambda(k)$:

$$\hat{\lambda}_\Lambda(k) := \max\left\{\arg\max_{\lambda \in [0, \Lambda]} [U(\lambda) - \lambda\xi(\lambda, k)]\right\}. \tag{4}$$

Here, 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, according to Equation (4), the usage level of the service is set in order to maximize the aggregate payoff of the user base.[4]

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.

3.3. Model of the Firm

By provisioning service capacity k, the firm derives revenue $b_1\hat{\lambda}_\Lambda(k)$ 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 service capacity provisioned is given by

$$k^*_\Lambda := \max\left\{\arg\max_{k \in \mathbb{R}_+} [b_1\hat{\lambda}_\Lambda(k) - k]\right\},^{5}$$

[2] Note that in the $M/M/k$ queue, the average response time differs from the average waiting time by an additive constant. It is straightforward to extend our results taking the average response time to be the congestion metric.

[3] Note that the payoff function $V(\lambda) - \xi(\lambda, k)$ is not in general monotone with respect to λ, and so there could be multiple Wardrop equilibria. Our definition of the user base behavior in Equation (3) picks the maximal Wardrop equilibrium. However, this is done for concreteness alone. Our analysis reveals that once the service system becomes large enough, the user base has a unique Wardrop equilibrium.

[4] Note that the aggregate payoff $U(\lambda) - \lambda\xi(\lambda, k)$ is not necessarily unimodal, and so could in general have multiple maximizers. Our definition in Equation (4) picks the largest of these maximizers. However, this is done for concreteness alone. Our analysis reveals that once the service system becomes large enough, the aggregate payoff does indeed become unimodal and has a unique maximizer.

[5] Once again, note that the profit function $b_1\hat{\lambda}_\Lambda(k) - k$ could in general have multiple maximizers and we define k^*_Λ to be the largest of these maximizers for concreteness. Our analysis reveals that once the market size becomes large enough, the profit function has a unique maximizer.

and corresponding request arrival rate is given by

$$\lambda_\Lambda^* := \hat{\lambda}_\Lambda(k_\Lambda^*).$$

The tuple $(\lambda_\Lambda^*, k_\Lambda^*)$ characterizes the equilibrium between the firm and the user base.[6] Since $\hat{\lambda}_\Lambda(k) < k$, a necessary condition for the firm to make positive profit is $b_1 > 1$. Since the case $b_1 \le 1$ 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, our main contribution is to 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 noncooperative and the cooperative population model.

Our results highlight the role played by noncooperative behavior among users, network effects, and economies of scale in large-scale service systems. To contrast these issues, we focus first on understanding the role of network effects when the user base is cooperative, and then we move to the noncooperative model.

4.1. Cooperative Population Model

We begin by studying the cooperative population model, defined in Equation (4). Our results highlight the strong influence of network effects on the capacity provisioning decision as well as the profit made by the firm. In particular, network effects influence the *order of magnitude* of the scaling that emerges for the equilibrium, and depending on the magnitude of network effects, three distinct scaling regimes emerge.

Our main result is stated in Theorem 1. However, to state it formally, we need to first describe some technical assumptions on the form of $U(\cdot)$.

Assumption 1. *The function* $U\colon \mathbb{R}_+ \to \mathbb{R}_+$ *is twice continuously differentiable over* $[0,\infty)$ *with* $U(0)=0$, $\lim_{\lambda\to\infty} U(\lambda)=\infty$. *Further,* $U'(\cdot)$ *is nonnegative, concave, and nondecreasing over* $[0,\infty)$.

The assumption that $U'(\cdot)$ is nondecreasing (i.e., $U(\cdot)$ is convex) seeks to capture positive network effects, with user utility growing with the usage of the service. The assumption that $U'(\cdot)$ is concave seeks to capture a diminishing growth rate in user utilities, as has been suggested in the literature (Odlyzko and Tilly 2005). As a special case, the above assumption allows for a quadratic growth of U with respect to λ; this is referred to in the literature as Metcalfe's Law (Metcalfe 1995).

Note that Assumption 1 implies that

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

Assumption 1 also implies that $w =: \lim_{\lambda\to\infty} U'(\lambda)/\lambda \in [0,\infty)$.

We are now ready to state our main result for this section, which highlights that, depending on the degree of network effects, three different operating regimes emerge for the profit-maximizing firm.

Theorem 1. *Consider the cooperative model of the user base and suppose that Assumption 1 holds. The following holds for both the M/M/1 and the M/M/k congestion model. For large enough* Λ, $\lambda_\Lambda^* = \Lambda$. *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.[7]

(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* $\omega\in(0,\infty)$, *then*

$$k_\Lambda^* = \Lambda + \frac{1}{\sqrt{\omega}} + o(1).$$

The three cases in the theorem correspond to different magnitudes of network effects in the user base, i.e., different growth rates of the aggregate utility $U(\cdot)$. Specifically,

Case (i). Corresponds to little or no network effects, i.e., asymptotically linear growth.

Case (ii). Corresponds to an asymptotically superlinear but subquadratic growth, which can be interpreted as "moderate" network effects.

Case (iii). Corresponds to an asymptotically quadratic growth, which can be interpreted as "extreme" network effects.

Thus, we consider progressively stronger positive network effects in Cases (i)–(iii).

There are a number of important insights highlighted by Theorem 1. First, it is easy see that in all three cases,

$$\lim_{\Lambda\to\infty} \frac{\lambda_\Lambda^*}{k_\Lambda^*} = 1.$$

[6] We loosely use the term "equilibrium" to describe the outcome $(\lambda_\Lambda^*, k_\Lambda^*)$ of the interaction between the service provider and the user base. Note that this hides the leader-follower nature of this interaction.

[7] For the $M/M/1$ model, $\beta(\alpha) = 1/\alpha$. For the $M/M/k$ model, $\beta(\alpha) = \zeta^{-1}(\alpha)$, where $\zeta(\cdot)$ is defined in Equation (B5) in Appendix B.

This means that, regardless of network effects, 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 service 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 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 aggregate utility $U(\cdot)$. Intuitively, if the aggregate utility is greater, then the user base becomes more tolerant to congestion. This means the firm can attract the full potential market base by provisioning lesser capacity, thereby making a higher profit.

More specifically, let us consider each of the three cases individually. Case (i) of Theorem 1 corresponds to small or no network effects, i.e., an asymptotically linear growth of the aggregate utility $U(\cdot)$. 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 $\sqrt{\Lambda}$. It is interesting to note that in this regime, the congestion disutility $\xi(\lambda^*_\Lambda, k^*_\Lambda)$ decays as $\Theta(1/\sqrt{\Lambda})$ 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

$$(b_1 - 1)\Lambda - \sqrt{\beta(\alpha)\Lambda} - o(\sqrt{\Lambda}). \tag{5}$$

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 service capacity required to maintain stability, i.e., Λ. Equation (5) implies that the service provider makes a profit $\Theta(\sqrt{\Lambda})$ less than the maximum possible because of the congestion sensitivity of the user base.

Moving to the case of "moderate" network effects, Case (ii) of Theorem 1 corresponds to an asymptotically super-linear, but subquadratic growth of $U(\cdot)$. This means, roughly, that the per user utility $V(\cdot)$ grows sublinearly. 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 $\sqrt{\Lambda/U'(\Lambda)}$. Note that under Case (ii), $U'(\Lambda) = o(\Lambda)$, and $U'(\Lambda) \to \infty$ as $\Lambda \to \infty$. Therefore, the spare capacity under Case (ii) grows to infinity, but slower than under Case (i). As a result, the congestion disutility $\xi(\lambda^*_\Lambda, k^*_\Lambda)$ decays as $\Theta(\sqrt{U'(\Lambda)/\Lambda})$ 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 $\Theta(\sqrt{\Lambda/U'(\Lambda)})$ less than the maximum possible.

Finally, let us consider the extreme network effects in Case (iii) of Theorem 1, which correspond to a quadratic growth of $U(\cdot)$. This means, roughly, that the per user utility $V(\cdot)$ grows linearly. In this case, the firm operates the system in a very heavy-traffic regime; it only needs to provision a bounded spare capacity. 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.

The proof of Theorem 1, provided in Appendix B, is based on first analysing the following "unconstrained" response of the user base, parameterised by the service capacity k:

$$\tilde{\lambda}(k) := \max\left\{\arg\max_{\lambda \geq 0}[U(\lambda) - \lambda\xi(\lambda, k)]\right\}. \tag{6}$$

Note that the above response function gives us the arrival rate attracted by the firm (and therefore its profit) as a function of the service capacity, ignoring the upper bound Λ on the arrival rate. We prove Theorem 1 by showing that for large enough Λ, the optimal strategy for the service provider is to provision capacity k^*_Λ that satisfies $\tilde{\lambda}(k^*_\Lambda) = \Lambda$. Intuitively, this may be explained as follows. Under the "unconstrained" user-response (6), statistical economies of scale result in an increasing profit for the firm with growing system size. Therefore, the firm's optimal policy is to grow the system, until a further increase in arrival rate is no longer possible because of the upper bound Λ.

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. Specifically, as the

network effects become stronger, the firm provisions lesser service capacity, users experience a higher congestion disutility, and the firm makes a greater profit. This behavior is intuitive, and represents a stark contrast to what we see in the next section for the case of a noncooperative user base.

4.2. Noncooperative Population Model

We now consider the noncooperative model of the user base, defined by Equation (3). 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. The main result of this section is that, compared to the case of cooperative users, for noncooperative 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 is stated in Theorem 2. However, before we can state it formally, we need to state the following technical assumption on the function $V(\cdot)$.

Assumption 2. *$V\colon \mathbb{R}_+ \to (0,\infty)$ is continuous, nondecreasing, and concave.*

Similar to the coopertive model, the above assumption seeks to capture network effects that are growing, but with a diminishing growth rate. Define $v := \lim_{\lambda\to\infty} V(\lambda)$.

We are now ready to state our main result for this section, which highlights that the impact of network effects when the user base is noncooperative is much smaller than what we observed in the case of a cooperative user base.

Theorem 2. *Consider the noncooperative model of the user base and suppose that Assumption 2 holds. The following holds for both the M/M/1 and the M/M/k congestion model. For large enough Λ, $\lambda^*_\Lambda = \Lambda$. Further, as $\Lambda \uparrow \infty$, the optimal capacity provisioning satisfies*

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

where $1/v$ is understood to be zero if $v = \infty$.

A first remark about Theorem 2 is that, like in the case of a cooperative user base, under the noncooperative model of the user base, it is optimal for the firm to provision enough capacity to attract the full potential user base. However, unlike the case of a cooperative user base, it is asymptotically optimal for the firm to operate the service in an extremely heavy-traffic regime: with just a bounded spare capacity. Remarkably, this is true irrespective of the strength of the network effects. In particular, the firm provisions a bounded spare capacity 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 spare capacity. On the other hand, under the noncooperative 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.

Though network effects do not influence the order of the scaling regime that emerges when the user base is noncooperative, as per Theorem 2, there is some impact from network effects. Specifically, as the market size becomes large, the spare capacity provisioned is approximately equal to $1/v$ and the component $1/v$ decreases as the network effects get stronger. Since the spare capacity remains bounded, the congestion disutility experienced by users does not approach zero as the market size grows to infinity, but remains bounded below by a positive constant.

Finally, Theorem 2 also highlights 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.

The statement of Theorem 2 can be explained intuitively as follows. It is natural to expect that the profit-maximizing firm would operate the service such that the net payoff for users is zero, i.e., $V(\lambda^*_\Lambda) = \xi(\lambda^*_\Lambda, k^*_\Lambda)$. Consider, for simplicity, the $M/M/1$ congestion model. The above condition can then be rewritten as $k^*_\Lambda - \lambda^*_\Lambda = 1/V(\lambda^*_\Lambda)$, which implies an asymptotically constant spare capacity if v is finite (recall that $v := \lim_{\lambda\to\infty} V(\lambda)$), and an asymptotically vanishing spare capacity if $v = \infty$. The proof of Theorem 2 is given in Appendix C. The proof technique is similar to that for Theorem 1: we first analyse the unconstrained response of the user base and then establish a connection between the unconstrained response and the optimal service capacity for the firm.

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 an order sense), irrespective of the strength of the network effects. The minimal impact of network effects on the capacity provisioning decision and the profit of the service provider is perhaps surprising given the previous literature on network effects and the order-of-magnitude impact of network effects when the user base is cooperative.

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 noncooperative 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 noncooperative 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 noncooperation, 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 service systems. In particular, such work typically imposes scalings exogenously, e.g., Halfin and Whitt (1981), Reed (2009), and 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.

Finally, it is important to note that the conclusions of this paper are specific to the advertising-supported service model. This model is mathematically equivalent to a model in which users pay the firm a fixed, exogenously determined price for using the service; this price can simply be absorbed into the utility function. Thus, our conclusions hold under the assumption that the firm cannot *strategically* set the price for the service. We note that very different scaling regimes emerge under service models where the firm can strategically set the price for the service (see, e.g., Maglaras and Zeevi 2003, Randhawa and Kumar 2008, Kumar and Randhawa 2010). Understanding the impact of network effects and noncooperative behavior in the user base in such settings is an interesting direction for future work.

Appendix

This appendix is devoted to the proofs of Theorems 1 and 2. For conciseness, we prove both theorems considering the M/M/k congestion model. The proofs for the (simpler) $M/M/1$ model follow along the same lines. We first collect some useful properties of the Erlang C formula in Appendix A. We then prove Theorem 1 and Theorem 2 in Appendix B and Appendix C, respectively.

A. Erlang C Properties

In this section, we collect properties of the continuous extension of the Erlang C formula (2) that are used in our proofs. Throughout, we use λ to denote the arrival rate and k to denote the (real) service capacity. We use $\rho := \lambda/k$ to denote the utilization per server.

The following lemma states if the utilization is held constant, then the probability of waiting diminishes as the service capacity grows.

Lemma 1. *For any fixed $\rho \in (0,1)$, $C(k\rho, k)$ is a strictly decreasing function of k.*

Proof. It was proved in Smith and Whitt (1981) that the continuous extension of the Erlang B formula (Jagerman 1974) is strictly decreasing with respect to (wrt) the service capacity when the utilization is held constant. Lemma 1 follows easily from this result, given the well-known relationship between the continuous extensions of the Erlang B and the Erlang C formulas (see, e.g., Janssen et al. 2011). □

The following lemma generalizes Proposition 1 in Halfin and Whitt (1981) to the continuous extension of the Erlang C waiting probability.

Lemma 2. *Consider a sequence of tuples $\{(\lambda_n, k_n)\}$ such that $\rho_n := \lambda_n/k_n < 1$ for all n and $\lim_{n\to\infty} k_n = \infty$. As $n \to \infty$, $C(\lambda_n, k_n)$ scales as follows:*

1. *Quality-driven regime: If $\lim_{n\to\infty} k_n(1-\rho_n)^2 = \infty$, then $\lim_{n\to\infty} C(\lambda_n, k_n) = 0$.*
2. *Quality-efficiency-driven regime: If $\lim_{n\to\infty} k_n(1-\rho_n)^2 = \beta \in (0,\infty)$, then*

$$\lim_{n\to\infty} C(\lambda_n, k_n) = \psi(\beta),$$

where

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

Here, $\Phi(\cdot)$ denotes the cumulative distribution function corresponding to the standard normal distribution.

3. *Efficiency-driven regime: If $\lim_{n\to\infty} k_n(1-\rho_n)^2 = 0$, then $\lim_{n\to\infty} C(\lambda_n, k_n) = 1$.*

The above lemma follows easily from Proposition 1 in Halfin and Whitt (1981) and Lemma 1.

The following lemma gives the derivative of the Erlang C waiting probability wrt the arrival rate.

Lemma 3. *For $0 < \lambda < k$,*

$$\frac{\partial C(\lambda,k)}{\partial \lambda} = \frac{(1-\rho)C(\lambda,k)}{\rho} + \frac{C(\lambda,k)(1-C(\lambda,k))}{k(1-\rho)}. \quad \text{(A1)}$$

A proof of this lemma can be found in Harel (2014).

The following lemma shows that the Erlang C waiting probability is convex wrt the arrival rate.

Lemma 4. *For fixed $k > 1$, $C(\lambda, k)$ is a strictly convex function of λ over $0 \le \lambda < k$.*

A proof of this lemma can be found in Harel (2014).

Finally, the following lemma is used in the proof of Theorem 1.

Lemma 5. *For fixed $k > 1$, the function*

$$\frac{C(\lambda,k)}{k-\lambda} + C(\lambda,k) + \frac{\lambda C(\lambda,k)(2-C(\lambda,k))}{(k-\lambda)^2} \tag{A2}$$

is strictly convex wrt λ over $0 \le \lambda < k$.

Proof. Let us denote the terms in (A2) as T_1, T_2, and T_3. Lemma 4 states that T_2 is strictly convex wrt λ. Moreover, since $C(\lambda, k)$ is strictly increasing in λ, it also follows from Lemma 4 that T_1 is strictly convex wrt λ. To prove the lemma, it therefore suffices to prove that T_3 is convex wrt λ.

Using Lemma 3, we may compute the derivative of T_3 wrt λ. After some simplification, we obtain

$$\frac{\partial T_3}{\partial \lambda} = \frac{k+\lambda}{(k-\lambda)^3} C(\lambda,k)(2-C(\lambda,k)) + 2C(\lambda,k)\frac{1-C(\lambda,k)}{k-\lambda} + \frac{2\lambda C^2(\lambda,k)}{k-\lambda}\left(\frac{1-C(\lambda,k)}{k-\lambda}\right)^2. \tag{A3}$$

Since $C(\lambda, k)$ is increasing in λ, and the function $x(2-x)$ is strictly increasing over $x \in [0, 1]$, it follows that the first term in (A3) is increasing in λ. Also, since $C(k, k) = 1$, it follows from Lemma 4 that the function $(1 - C(\lambda, k))/(k - \lambda)$ is increasing in λ. This implies that the second and the third term in (A3) are also increasing in λ. Since $\partial T_3/\partial\lambda$ is increasing in λ, we conclude that T_3 is convex wrt λ. □

B. Proof of Theorem 1

To prove Theorem 1, we first analyse the unconstrained multiserver scaling regime (6) parameterised by service capacity $k > 0$. Define, $\tilde{\rho}(k) := \tilde{\lambda}(k)/k$. 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 6. *Suppose Assumption 1 holds. Then $\tilde{\lambda}(k)$ is a continuous, nondecreasing function of k over $k \in (0, \infty)$. As $k \uparrow \infty$, $\tilde{\lambda}(k) \uparrow \infty$ as follows.*

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

$$\lim_{k\to\infty} k(1-\tilde{\rho}(k))^2 = \beta(\alpha), \tag{B1}$$

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

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

$$\lim_{k\to\infty} kU'(\tilde{\lambda}(k))(1-\tilde{\rho}(k))^2 = 1. \tag{B2}$$

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

We defer the proof of Lemma 6 to later in this section, and use it first to prove Theorem 1. Lemma 6 implies that for large enough k, $\tilde{\lambda}(k)$ is strictly increasing. We define the following inverse of $\{\tilde{\lambda}(k)\}$. Taking $\tilde{\lambda}(0) := 0$, we define $\tilde{k}\colon \mathbb{R}_+ \to \mathbb{R}_+$ as follows:

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

Since $\tilde{\lambda}(k) \overset{k\uparrow\infty}{\to} \infty$, $\tilde{k}(\lambda)$ is well defined for all $\lambda \in \mathbb{R}_+$. Moreover, for large enough λ, $\tilde{k}(\lambda)$ is continuous and strictly increasing with $\tilde{\lambda}(\tilde{k}(\lambda)) = \lambda$, and $\tilde{k}(\lambda) \overset{\lambda\uparrow\infty}{\to} \infty$.

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

Lemma 7. *For large enough Λ, $k^*_\Lambda = \tilde{k}(\Lambda)$ and $\lambda^*_\Lambda = \Lambda$.*

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

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

Now from Lemma 6, 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 wrt k, and $k(b_1\tilde{\rho}(k) - 1) \overset{k\uparrow\infty}{\to} \infty$. This in turn implies that for large enough Λ,

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

which implies that

$$k^*_\Lambda \ge \tilde{k}(\Lambda).$$

Moreover, we know that for large enough Λ, $\hat{\lambda}_\Lambda(\tilde{k}(\Lambda)) = \tilde{\lambda}(\tilde{k}(\Lambda)) = \Lambda$. This means that by provisioning a capacity of $\tilde{k}(\Lambda)$, the service provider attracts the maximum possible arrival rate. It then follows easily that the service provider has no incentive to provision a capacity exceeding $\tilde{k}(\Lambda)$. This completes the proof.

We are now ready to prove the statements of Theorem 1. We show now that Lemmas 6 and 7 imply statement (ii) of Theorem 1. Statements (i) and (iii) can be proved on similar lines.

Let us assume therefore, that $\alpha = \infty$, and $\lim_{\lambda\to\infty} U'(\lambda)/\lambda = 0$. In the following, we use the notation $f(\Lambda) \sim g(\Lambda)$ to mean that $\lim_{\Lambda\to\infty} f(\Lambda)/g(\Lambda) = 1$. Since $\tilde{k}(\Lambda) \overset{\Lambda\uparrow\infty}{\to} \infty$, it follows from statement (ii) of Lemma 6 that

$$\lim_{\Lambda\to\infty} \tilde{k}(\Lambda)U'(\tilde{\lambda}(\tilde{k}(\Lambda)))\left(1 - \frac{\tilde{\lambda}(\tilde{k}(\Lambda))}{\tilde{k}(\Lambda)}\right)^2 = 1.$$

From Lemma 7, the above statement can be rewritten as

$$\lim_{\Lambda\to\infty} k^*_\Lambda U'(\Lambda)\left(1 - \frac{\Lambda}{k^*_\Lambda}\right)^2 = 1.$$

Now, noting that $k^*_\Lambda \sim \Lambda$, we have

$$k^*_\Lambda - \Lambda \sim \sqrt{\frac{k^*_\Lambda}{U'(\Lambda)}} \sim \sqrt{\frac{\Lambda}{U'(\Lambda)}},$$

which implies 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 6. The remainder of this section is devoted to this proof.

Proof of Lemma 6. The proof uses four main steps:

1. We first show that $\tilde{\lambda}(k)$ is a continuous function. Let $f(\lambda, k) := U(\lambda) - \lambda\xi(\lambda, k)$. Since f is continuous, it is easy to see that a sufficient condition for continuity of $\tilde{\lambda}(k)$ is that the optimization in (6) has a unique maximizer for any $k > 0$.

We now prove that optimization in (6) has a unique maximizer. Using Lemma 3, the partial derivative of f wrt λ may be computed as follows:

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

Since $U'(\cdot)$ is concave, it follows from Lemma 5 that $(\partial/\partial\lambda)f(\lambda, k)$ is strictly concave wrt λ. Moreover, this partial derivative is nonnegative at $\lambda = 0$, and approaches $-\infty$ as $\lambda \uparrow k$. Thus, we conclude that $f(\lambda, k)$ is a unimodal wrt λ, and that the optimization in (6) has a unique maximizer.

2. Next, we show that $\tilde{\lambda}(k)$ is nondecreasing, and $\tilde{\lambda}(k) \uparrow \infty$ as $k \uparrow \infty$.

It is easy to see that there exists $k_1 > 0$ such that $\tilde{\lambda}(k_1) > 0$. Pick $k_2 > k_1$. Invoking Lemma 8, which proves that f is supermodular, it follows that for $0 \le \lambda < \tilde{\lambda}(k_1)$,

$$f(\tilde{\lambda}(k_1), k_2) - f(\lambda, k_2) > f(\tilde{\lambda}(k_1), k_1) - f(\lambda, k_1) \ge 0.$$

This implies that $\tilde{\lambda}(k_2) \ge \tilde{\lambda}(k_1)$. This proves that $\tilde{\lambda}(k)$ is a nondecreasing function, which implies that $\lim_{k\to\infty}\tilde{\lambda}(k)$ exists. 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 $k' > 0$ such that $U(\lambda_1) > f(\lambda_1, k') > \max_{0\le\lambda\le\nu} U(\lambda)$. This means that

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

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

3. Next, we show that for large enough k, $\tilde{\rho}(k)$ is strictly increasing. Let $\rho := \lambda/k$. We know that there exists $\bar{k} > 0$ such that for all $k > \bar{k}$, $\tilde{\lambda}(k) > 0$. Since f is continuously differentiable wrt λ, for $k > \bar{k}$, $\tilde{\lambda}(k)$ satisfies the following first order condition:

$$\begin{aligned}U'(\lambda) &= \frac{\partial}{\partial\lambda}\left(\frac{\lambda C(\lambda,k)}{k-\lambda}\right)\\ &= \frac{C(\lambda,k)}{k(1-\rho)} + C(\lambda,k) + \frac{\rho C(\lambda,k)(2-C(\lambda,k))}{k(1-\rho)^2}\\ &= \frac{C(k\rho,k)}{k(1-\rho)} + C(k\rho,k)\\ &\qquad + \frac{\rho C(k\rho,k)(2-C(k\rho,k))}{k(1-\rho)^2} =: h(\rho,k). \qquad \text{(B3)}\end{aligned}$$

The function h defined above has the following properties.

(i) For fixed $\rho \in (0, 1)$, $h(\rho, k)$ is strictly decreasing in k, since $C(k\rho, k)$ is strictly decreasing in k (see Lemma 1).

(ii) For fixed $k > 0$, $h(\rho, k)$ is strictly increasing in ρ, since $C(k\rho, k)$ is a strictly increasing in ρ.

Now, pick $k_2 > k_1 > \bar{k}$. Since $\tilde{\lambda}(k_2) \ge \tilde{\lambda}(k_1)$, we have, from Assumption 1, that $U'(\tilde{\lambda}(k_2)) \ge U'(\tilde{\lambda}(k_1))$. Therefore,

$$\begin{aligned}&h(\tilde{\rho}(k_2), k_2)\\ &\quad = U'(\tilde{\lambda}(k_2)) \ge U'(\tilde{\lambda}(k_1)) = h(\tilde{\rho}(k_1), k_1) > h(\tilde{\rho}(k_1), k_2),\end{aligned}$$

where the last inequality follows from property (i). Since $h(\tilde{\rho}(k_2), k_2) > h(\tilde{\rho}(k_1), k_2)$, property (ii) implies that $\tilde{\rho}(k_2) > \tilde{\rho}(k_1)$.

4. Finally, we prove Statements (B1) and (B2). Let $\theta(k) := kU'(\tilde{\lambda}(k))(1-\tilde{\rho}(k))^2$. Note that the function $\theta(k)$ must have a limit point in $[0, \infty]$ as $k \uparrow \infty$. 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 (B3), for large enough k, $\tilde{\lambda}(k)$ satisfies

$$\begin{aligned}1 &= \frac{C(\lambda,k)}{kU'(\lambda)(1-\rho)} + \frac{C(\lambda,k)}{U'(\lambda)} + \frac{\rho C(\lambda,k)(2-C(\lambda,k))}{kU'(\lambda)(1-\rho)^2}\\ &=: T_1 + T_2 + T_3. \qquad \text{(B4)}\end{aligned}$$

Let us suppose (for the sake of obtaining a contradiction) that there exists a strictly increasing sequence $\{k_n\}$ satisfying $\lim_{n\to\infty} k_n = \infty$ such that $\lim_{n\to\infty}\theta(k_n) = \infty$. In (B4), it is easy to see that along this sequence, $\lim_{n\to\infty} T_1 = 0$ and $\lim_{n\to\infty} T_3 = 0$. We now show that $\lim_{n\to\infty} T_2 = 0$. If $\alpha := \lim_{\lambda\uparrow\infty} U'(\lambda) = \infty$, then this is obvious. If $\alpha \in (0, \infty)$, then $\lim_{n\to\infty} k_n(1-\tilde{\rho}(k_n))^2 = \infty$. This corresponds to the *quality driven regime*, and Lemma 2 implies that $\lim_{n\to\infty} C(\tilde{\lambda}(k_n), k_n) = 0$. This in turn implies $\lim_{n\to\infty} T_2 = 0$. Since the right-hand side of (B4) approaches 0 as $n \uparrow \infty$, we have a contradiction. Therefore, ∞ is not a limit point of $\theta(k)$ as $k \uparrow \infty$.

Next, let us suppose (for the sake of obtaining a contradiction) that there exists a strictly increasing sequence $\{k_n\}$ satisfying $\lim_{n\to\infty} k_n = \infty$ such that $\lim_{n\to\infty}\theta(k_n) = 0$. In this case, $\lim_{n\to\infty} k_n(1-\tilde{\rho}(k_n))^2 = 0$. This corresponds to the *efficiency driven regime*, and Lemma 2 implies that $\lim_{n\to\infty} C(\tilde{\lambda}(k_n), k_n) = 1$. Therefore, in (B4), along the sequence under consideration, $\lim_{n\to\infty} T_3 = \infty$. Since T_1 and T_2 are nonnegative sequences, this gives us a contradiction. Therefore, 0 is not a limit point of $\theta(k)$ as $k \uparrow \infty$.

Since ∞ and 0 are not limit points of $\theta(k)$ as $k \uparrow \infty$, there exists a limit point $\gamma \in (0, \infty)$. We now show that this limit point is unique. Let $\{k_n\}$ be a strictly increasing sequence satisfying $\lim_{n\to\infty} k_n = \infty$ such that $\lim_{n\to\infty}\theta(k_n) = \gamma$. Consider the following two cases.

Case 1. $\alpha = \infty$. Along the sequence $\{k_n\}$, $k_n(1-\tilde{\rho}(k_n))^2 \to 0$, which (as we have seen before) implies $C(\tilde{\lambda}(k_n), k_n) \to 1$. Therefore, along this sequence, $T_1 \to 0$, $T_2 \to 0$, $T_3 \to 1/\gamma$ (see (B4)). This implies $\gamma = 1$, which proves (B2).

Case 2. $\alpha \in (0, \infty)$. Along the $\{k_n\}$, $k_n(1-\tilde{\rho}(k_n))^2 \to \beta$, where $\beta := \gamma/\alpha$. This corresponds to the *quality-efficiency driven regime*, and Lemma 2 implies that $C(\tilde{\lambda}(k_n), k_n) \to \psi(\beta)$, where

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

In the above expression, $\Phi(\cdot)$ denotes the cumulative distribution function corresponding to the standard normal

distribution. Therefore, along the sequence under consideration, $T_1 \to 0$, $T_2 \to \psi(\beta)/\alpha$, $T_3 \to \psi(\beta)(2-\psi(\beta))/\beta\alpha$ (see (B4)). Therefore, we must have

$$\alpha = \psi(\beta) + \frac{\psi(\beta)(2-\psi(\beta))}{\beta} =: \zeta(\beta). \tag{B5}$$

As we prove in Lemma 9, $\zeta(\cdot)$ is strictly decreasing, with $\lim_{x\downarrow 0}\zeta(x)=\infty$, $\lim_{x\uparrow\infty}\zeta(x)=0$. Therefore, there is a unique $\beta(\alpha)\in(0,\infty)$ that satisfies (B5). Moreover, it is clear that $\beta(\alpha)$ is strictly decreasing wrt α. This proves (B1). This completes the proof of Lemma 6.

Lemma 8. *$f(\lambda,k) := U(\lambda) - \lambda\xi(\lambda,k)$ 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. For $\lambda < k$,

$$\frac{\partial\xi(\lambda,k)}{\partial\lambda} = \frac{\partial}{\partial\lambda}\left(\frac{C(\lambda,k)}{k-\lambda}\right) = \frac{C(\lambda,k)}{\lambda} + \frac{C(\lambda,k)(2-C(\lambda,k))}{(k-\lambda)^2}.$$

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

$$\xi(\lambda_2,k_1) - \xi(\lambda_1,k_1) \ge \xi(\lambda_2,k_2) - \xi(\lambda_1,k_2) \implies f(\lambda_2,k_2) - f(\lambda_2,k_1) > f(\lambda_1,k_2) - f(\lambda_1,k_1).$$

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

Lemma 9. *$\zeta(\cdot)$ as defined in (B5) is strictly decreasing, with $\lim_{x\downarrow 0}\zeta(x)=\infty$, $\lim_{x\uparrow\infty}\zeta(x)=0$.*

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

C. Proof of Theorem 2

To prove Theorem 2, we first analyse the following unconstrained multiserver scaling regime parameterised by the service capacity k. Define, for $k>0$,

$$\tilde\lambda(k) := \max\{\lambda \ge 0 \mid V(\lambda) - \xi(\lambda,k) \ge 0\}, \quad \tilde\rho(k) := \frac{\tilde\lambda(k)}{k}.$$

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 10. *Given Assumption 2, $\tilde\lambda(k)$ is a strictly increasing and continuous function of k. As $k\uparrow\infty$, $\tilde\lambda(k)\uparrow\infty$ such that*

$$\lim_{k\to\infty} kV(\tilde\lambda(k))(1-\tilde\rho(k)) = 1. \tag{C1}$$

Moreover, $\tilde\rho(k)$ is strictly increasing in k.

Proof. The proof uses three main steps:

1. First, we show that $\tilde\lambda(k)$ is continuous and strictly increasing, with $\lim_{k\to\infty}\tilde\lambda(k)=\infty$. Let $f(\lambda,k) := V(\lambda) - \xi(\lambda,k)$. Since V is concave, it follows from Lemma 4 that for fixed $k>0$, $f(\lambda,k)$ is strictly concave wrt λ. Moreover, since f is continuous, $f(0,k)>0$, and $\lim_{\lambda\uparrow k} f(\lambda,k) = -\infty$, it follows that
   (i) $\tilde\lambda(k)$ is the unique solution of $f(\lambda,k)=0$,
   (ii) $f(\lambda,k)<0\ \forall\,\lambda\in(\tilde\lambda(k),k)$.

Statement (i), coupled with the continuity of f, implies that $\tilde\lambda(k)$ is continuous. It also follows from statements (i) and (ii) that $\tilde\lambda(k)$ is strictly increasing. Indeed, for $k_2 > k_1 > 0$, $f(\tilde\lambda(k_1),k_2) > f(\tilde\lambda(k_1),k_1) = 0$, which implies that $\tilde\lambda(k_2) > \tilde\lambda(k_1)$.

We now argue that $\lim_{k\to\infty}\tilde\lambda(k)=\infty$. 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 large enough such that $f(\lambda_1,k_1)>0$. Statement (ii) then implies that $\tilde\lambda(k_1)>\lambda_1$, which is a contradiction.

1. Next, we prove that (C1) holds. Let $\theta(k) := kV(\tilde\lambda(k))\cdot(1-\tilde\rho(k))$. From statement (i), we know that

$$V(\tilde\lambda(k)) = \xi(\tilde\lambda(k),k) = \frac{C(\tilde\lambda(k),k)}{k-\tilde\lambda(k)} \implies \theta(k) = C(\tilde\lambda(k),k). \tag{C2}$$

Equation (C2) 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, $kV(\tilde\lambda(k))(1-\tilde\rho(k))^2$ and thus also $k(1-\tilde\rho(k))^2$ converges to 0. This corresponds to the *efficiency driven regime*, and Lemma 2 implies that $\lim_{k\to\infty}C(\tilde\lambda(k),k)=1$, which implies (using (C2)) that $\lim_{k\to\infty}\theta(k)=1$.

2. Finally, it remains to show that $\tilde\rho(k)$ is strictly increasing. From statement (i), we know that for any $k>0$, $\tilde\lambda(k)$ satisfies

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

where $\rho=\lambda/k$. The function h has the following properties.

(i) For fixed $\rho\in(0,1)$, $h(\rho,k)$ is strictly decreasing in k, since $C(k\rho,k)$ is strictly decreasing in k.

(ii) For fixed k, $h(\rho,k)$ is strictly increasing in $\rho\in(0,1)$, since $C(k\rho,k)$ is strictly increasing in ρ.

Now, pick $k_2>k_1>0$. $h(\tilde\rho(k_2),k_2) = V(\tilde\lambda(k_2)) \ge V(\tilde\lambda(k_1)) = h(\tilde\rho(k_1),k_1) > h(\tilde\rho(k_1),k_2)$. The last inequality above follows from Property (i). Since $h(\tilde\rho(k_2),k_2) > h(\tilde\rho(k_1),k_2)$, it follows from Property (ii) that $\tilde\rho(k_2)>\tilde\rho(k_1)$. □

The remainder of the proof follows along the same lines as the proof of Theorem 1. Lemma 10 allows us to define the following inverse of $\{\tilde\lambda(k)\}$. Taking $\tilde\lambda(0) := 0$, we define $\tilde k\colon \mathbb{R}_+ \to \mathbb{R}_+$ as follows:

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

As in the proof of Theorem 1, we have the following connection between $(\lambda^*_\Lambda, k^*_\Lambda)$ and $\tilde\lambda(k)$.

Lemma 11. *For large enough Λ, $k^*_\Lambda = \tilde k(\Lambda)$ and $\lambda^*_\Lambda = \Lambda$.*

The proof of Lemma 11 is identical to that of Lemma 7, and it can be applied to prove the statements of Theorem 2 in a similar way as Lemma 7 was applied in the proof of Theorem 1. We omit the details.

References

Atar R (2012) A diffusion regime with nondegenerate slowdown. *Oper. Res.* 60(2):490–500.
Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Oper. Res.* 52(1):17–34.
Buchanan JM (1965) An economic theory of clubs. *Economica* 32(125):1–14.
Farrell J, Klemperer P (2007) Coordination and lock-in: Competition with switching costs and network effects. *Handbook Indust. Organ.* 3:1967–2072.
Farrell J, Saloner G (1985) Standardization, compatibility, and innovation. *RAND J. Econom.* 16(1):70–83.
Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.
Hamilton J (2009) The cost of latency. *Perspective.* (October 31), http://perspectives.mvdirona.com/2009/10/31/TheCostOfLatency.asp.
Harel A (2014) Convexity result for the analytic continuation of the Erlang delay formula. Working paper, Baruch College, City University of New York, New York.
Interactive Advertising Bureau (2011). IAB internet advertising revenue report conducted by PricewaterhouseCoopers (PWC). Report, http://www.iab.net/adrevenuereport.
Jagerman DL (1974) Some properties of the Erlang loss function. *Bell System Technical J.* 53(3):525–551.
Jagers AA, Van Doorn EA (1991) Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Rev.* 33(2):281–282.
Janssen AJEM, Van Leeuwaarden JSH, Zwart B (2011) Refining square-root safety staffing by expanding Erlang C. *Oper. Res.* 59(6):1512–1522.
Johari R, Kumar S (2010) Congestible services and network effects. *Proc. 11th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 93–94.
Katz ML, Shapiro C (1985) Network externalities, competition, and compatibility. *Amer. Econom. Rev.* 75(3):424–440.
Kohavi R, Longbotham R, Sommerfield D, Henne RM (2009) Controlled experiments on the Web: Survey and practical guide. *Data Mining and Knowledge Discovery* 18(1):140–181.
Kumar S, Randhawa RS (2010) Exploiting market size in service systems. *Manufacturing Service Oper. Management* 12(3): 511–526.
Lohr S (2012) For impatient Web users, an eye blink is just too long to wait. *New York Times* (February 29), http://www.nytimes.com/2012/03/01/technology/impatient-web-users-flee-slow-loading-sites.html.
Maglaras C, Zeevi A (2003) Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* 49(8):1018–1038.
Metcalfe B (1995) Metcalfe's law: A network becomes more valuable as it reaches more users. *Infoworld* 17(40):53–54.
Odlyzko A, Tilly B (2005) A refutation of Metcalfe's law and a better estimate for the value of networks and network interconnections. Preprint, University of Minnesota, Minneapolis.
Oren SS, Smith SA (1981) Critical mass and tariff structure in electronic communications markets. *Bell J. Econom.* 12(2): 467–487.
Randhawa RS, Kumar S (2008) Usage restriction and subscription services: Operational benefits with rational users. *Manufacturing Service Oper. Management* 10(3):429–447.
Reed J (2009) The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* 19:2211–2269.
Sandler T, Tschirhart J (1997) Club theory: Thirty years later. *Public Choice* 93(3):335–355.
Smith DR, Whitt W (1981) Resource sharing for efficiency in traffic systems. *Bell System Technical J.* 60(1):39–55.
Sundararajan A (2003) Network effects, nonlinear pricing and entry deterrence. Working Paper IS-03-01, NYU Center for Digital Economy Research, New York University, New York.
Vanderbilt T (2009) Data center overload. *New York Times Magazine* (June 8), http://www.nytimes.com/2009/06/14/magazine/14search-t.html.
Whitt W (2003) How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* 51(4):531–542.