On Competitive Provisioning Of Cloud Services

Jayakrishnan Nair
IIT Bombay

Vijay G. Subramanian
Northwestern University

Adam Wierman
California Institute of Technology

ABSTRACT

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

1. INTRODUCTION

Cloud-based services are increasingly becoming the norm. While cloud-based email applications have been around for decades at this point, other cloud-based services are increasingly replacing a wide variety of applications that used to be run locally, e.g., document editing (GoogleDocs, Office365) and file storage (Dropbox, GoogleDrive, iCloud). For the purposes of this paper, there are four main features of this growing marketplace that are important to highlight.

(i) A majority of cloud services derive *revenue primarily from advertising* and are offered for free to users: e.g., for Google and Facebook ad-driven online services are cash cows [6].

(ii) Users of online services are *highly delay sensitive* and small additional delays can be traced to significant declines in revenue [5, 8].

(iii) Cloud services have *positive network effects*, i.e., the experience of users in cloud services often is highly dependent on the number of users subscribed to the service [4, 7]: e.g., social networking services.

(iv) Cloud services are often *highly competitive* [3],: e.g. the competition between Hotmail, Gmail, and Yahoomail, or between Facebook and GooglePlus.

The interplay of these four factors leads to a complex marketplace with complicated interactions between user experience (congestion and network effects), service capacity provisioning, and market share. *The goal of this paper is to analytically investigate the influence of these factors.*

Related work. The impact of network effects is crucial for cloud services: e.g., the more users there are on Facebook, the more appealing it is to be on Facebook. However, network effects are not specific to cloud services, and have been studied extensively in the economics and operations management literatures. While most of this literature [10, 11] does not consider congestion, the literature on "club theory" focuses on the interaction of network effects and congestion. The theory of clubs [2, 12] deals with groups of congestion-sensitive users sharing certain resources. Cloud services can also be interpreted as a club good offered by competing profit maximizing firms. However, in contrast to the club goods setting, cloud service providers typically do not charge users for the service but instead obtain their revenue from advertising. This difference leads to significantly different conclusions in our setting.

The only previous piece of work to consider network effects and congestion in an ad-supported service is [9], which focuses on the capacity provisioning of a *single monopolistic service* when faced with a strategic user population with positive network effects. In this setting, [9] shows that the network effects lead to the user base being more tolerant of congestion, which allows the provider to run the service with fewer servers and, thus, derive a larger profit.

Since the "club theory" literature does not consider ad-supported services and [9] does not consider competition, no piece of prior work has investigated the interplay of all four of the factors described above.

Contributions of this paper. In this paper we seek to answer questions such as: *Does competition lead to improved user experience in ad-supported services? Can competing firms coexist or will near-monopolies emerge?* To address such questions, we introduce a new model that extends the setting of [9] in order to capture competition between service providers, a.k.a., firms. The key novelty in this extension is how network effects are considered. We consider two variations of network effects in this paper: *firm-specific* and *industry-wide* network effects.

Firm-specific network effects capture settings where the utility of a user of a particular firm depends only on the population of users of that specific firm. This captures settings like Facebook, where a user's utility from joining Facebook grows as the number of people using Facebook grows. On the other hand, industry-wide network effects model situations where the utility of a user of a particular firm depends on the number of users across all the firms in the industry, not just the number of users at the specific firm. This captures applications such as email, where a user's utility grows with the number of people that use email, not just with the number of people that use the same email client. Of course many applications have a combination of these two forms of network effects, but we focus on the extreme situations in this paper in order to contrast the effects of each. Within these models of network effects we study a situation where a user base decides selfishly which, if any, of two services to join based on the congestion and network effects available at each service. Each of these is a function of the capac-

IFIP WG 7.3 Performance 2014, October 7-9, Turin, Italy.

ity decision of the profit-maximizing, competing firms. Our analysis focuses on the setting where the user base is large, i.e., increasing to infinity.

The main messages from our analysis are the following:

(i) Depending on whether the network effects are firm-specific or industry-wide, different market structures emerge. While firms can share the market under industry-wide network effects, near monopolies tend to emerge under firm-specific network effects setting.

(ii) Generally, competition in ad-supported services does not improve the performance for the users, in contrast to competition in paid services [1].

The first conclusion explains several informal observations in the marketplace, e.g., Facebook enjoys near-monopoly status while Gmail, Hotmail, and many other email providers coexist. Moreover, in order to compete in areas where network effects are firm-specific, services must build a user base before entering the market, e.g., Twitter or GooglePlus *vis-à-vis* Facebook.

2. PRELIMINARIES: A SINGLE FIRM

Before we move to the case of competing firms, it is useful to consider the case of a single monopolistic firm, which is analysed in [9]. In this section, we recall a key result from [9], which we contrast in subsequent sections with our results for the case of competing firms. First, we present our model, which is a special case of that in [9].

User model. To model the user base, we assume that the user-requests for service arrive at rate at most Λ. At arrival rate $\lambda \in [0, \Lambda]$, each user obtains a utility $V(\lambda)$. To model positive network effects we assume that V is of the form

$$V(\lambda) = w\lambda^{\beta} \quad (w > 0, \beta \in [0, 1]). \tag{1}$$

Note that the network effect becomes stronger as β increases, i.e., $\beta = 0$ corresponds to the case of no network effects, and $\beta = 1$ corresponds to the case of strong network effects.

We assume that each user perceives a non-negative latency cost $f(\lambda, C)$ that is a function of both the arrival rate of requests, λ, and the provisioned capacity of the provider, C. For simplicity, we model $f(\lambda, C)$ as the average (stationary) response time in an M/M/1 queue, assuming that each user request has a unit service requirement on average, i.e.,

$$f(\lambda, C) = \begin{cases} \frac{1}{C-\lambda} & \text{if } \lambda < C \\ +\infty & \text{otherwise} \end{cases}. \tag{2}$$

Thus, when the arrival rate of user-requests equals λ, each user derives a net payoff equal to $V(\lambda) - f(\lambda, C)$. We assume that the realized arrival rate, denoted by $\hat{\lambda}_{\Lambda}(C)$, is given by the Wardrop equilibrium between (infinitesimal) users:

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

Since the payoff function $V(\lambda) - f(\lambda, C)$ is concave in λ, and approaches $-\infty$ as $\lambda \uparrow C$, it is easy to see that the above equation describes the unique Wardrop equilibrium between users. Note that the above model assumes that each user selfishly seeks to maximize her own payoff.

Model of the firm. The final piece of the model is the strategic behavior of the firm, which seeks to choose capacity so as to maximize profit. We assume that the cloud service provider makes b dollars per user served from advertising and pays a dollar per unit cost for each unit of capacity. So, the firm's profit is given by $b\hat{\lambda}_{\Lambda}(C) - C$. Owing to the stability constraint, $\hat{\lambda}_{\Lambda}(C) < C$, and so we necessarily need $b > 1$ for the firm to consider offering the service. Thus, the firm provisions a capacity

$$C^*(\Lambda) = \max\{\arg\max_{C \geq 0} b\hat{\lambda}_{\Lambda}(C) - C\}.$$

For concreteness we choose the largest capacity if multiple solutions exist; it turns out that for large enough Λ, the above maximization has a unique solution (see [9]).

Result from [9]. The following theorem describes the behavior of the tuple $C^*(\Lambda)$ and $\lambda^*(\Lambda) := \hat{\lambda}_{\Lambda}(C^*(\Lambda))$ for large Λ, and the queueing regime that emerges from the interaction between the user base and the firm.

THEOREM 1. *Consider the case of a single service provider. For large enough Λ, $\lambda^*(\Lambda) = \Lambda$, and*

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

Note that for large enough Λ, it is beneficial for the firm to provision enough capacity to attract the maximum possible arrival rate. Moreover, the firm provisions the minimum capacity required to serve the realized arrival rate, plus a 'spare capacity' equal to $\frac{1}{V(\Lambda)}$.

3. TWO COMPETING FIRMS

We now move to the case of two competing firms. Our goal is to study the interplay of network effects, congestion, and competition in ad-supported services.

As in the case of a single firm, we assume that user requests are generated at a rate of at most Λ. Firm i, having service capacity C_i, sees an arrival rate $\hat{\lambda}_i(C_i, C_{-i})$, where for $i \in \{1, 2\}$, $-i := \{1, 2\} \setminus \{i\}$, and $\hat{\lambda}(C_1, C_2) := \hat{\lambda}_1(C_1, C_2) + \hat{\lambda}_2(C_2, C_1) \in [0, \Lambda]$. The precise definition of the traffic split $(\hat{\lambda}_1(C_1, C_2), \hat{\lambda}_2(C_2, C_1))$ depends on whether the network effects are industry-wide (see Section 3.1) or firm-specific (see Section 3.2).

With an arrival rate $\hat{\lambda}_i(C_i, C_{-i})$, Firm i makes a profit $b_i\hat{\lambda}_i(C_i, C_{-i}) - C_i$. We assume as before that $b_i > 1$. Since the profit of each firm is dependent on the capacity provisioned by the other, we study the interaction between the firms as a game, and analyze the resulting Nash equilibria.

To be precise, we define $(\lambda_1, \lambda_2, C_1, C_2)$ to be an equilibrium of our system if (λ_1, λ_2) is the response of the user base to the service capacities (C_1, C_2) and (C_1, C_2) is a Nash equilibrium between the providers, i.e., for $i = 1, 2$

$$\lambda_i = \hat{\lambda}_i(C_i, C_{-i})$$
$$C_i \in \arg\max_{c \geq 0} b_i\hat{\lambda}_i(c, C_{-i}) - c.$$

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

3.1 Industry-wide network effects

We first focus on industry-wide network effects. Under industry-wide network effects, given a split of arrival rates (λ_1, λ_2), each user obtains a payoff equal to $V(\lambda)$ where $\lambda = \lambda_1 + \lambda_2$. The key is that that the utility does not depend on which firm a user picks; it depends only on the net arrival rate into both services. So, network effects are not firm specific. Thus, a user of Firm i obtains a net payoff of $V(\lambda) - f(\lambda_i, C_i)$.

Given a net arrival rate λ, a split (λ_1, λ_2) satisfying $\lambda_1 + \lambda_2 = \lambda$ is called a Wardrop split if, for $i = 1, 2$,

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

It is easy to verify that for $\lambda < C_1 + C_2$, there is a unique Wardrop split $(\lambda_1(\lambda, C_1, C_2), \lambda_2(\lambda, C_2, C_1))$. The realized traffic split is then defined by

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

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

Results. The following theorem characterizes the equilibria that may emerge under industry-wide network effects.

THEOREM 2. *Consider the industry-wide network effects model. For large enough Λ, the following statements hold:*

1. *If $b_1, b_2 \in (1, 2]$, then a continuum of equilibria exist, including monopoly configurations. Moreover, any equilibrium is of one of the following forms:*
   - *(a) Monopoly for Firm 1: $\lambda_1 = \Lambda$, $C_1 = \Lambda + \frac{1}{V(\Lambda)}$, $\lambda_2 = C_2 = 0$;*
   - *(b) Monopoly for Firm 2: $\lambda_2 = \Lambda$, $C_2 = \Lambda + \frac{1}{V(\Lambda)}$, $\lambda_1 = C_1 = 0$;*
   - *(c) Firms 1 and 2 share the market such that $\lambda_1 + \lambda_2 = \Lambda$, and for $i = 1, 2$*
   $$\lambda_i \geq \frac{1}{(b_i - 1)V(\Lambda)}, \quad C_i = \lambda_i + \frac{1}{V(\Lambda)};$$
2. *If $b_1 > 2$, $b_2 \leq 2$, then the only equilibrium is a full monopoly of Provider 1:*
   $$\lambda_1 = \Lambda, \;\; C_1 = \Lambda + \frac{1}{V(\Lambda)}, \;\; \lambda_2 = C_2 = 0;$$
3. *If $b_1, b_2 > 2$, then no equilibrium exists.*

We highlight two key take-aways from Theorem 2. First, unless either firm has an extremely high advertising efficiency, there is a multitude of ways for the firms to divide the market. Second, in every equilibrium configuration demonstrated in Theorem 2, the user base sees exactly the same latency cost as in the case of a single provider (See Theorem 1). Thus, we see that competition in the marketplace does not improve the payoff experienced by the user base.

3.2 Firm-specific network effects

We now move to the case when the network effects are firm-specific. When the network effects are firm-specific, given a traffic split (λ_1, λ_2), users of Firm i obtain a net payoff equal to $V_i(\lambda_i) - f(\lambda_i, C_i)$. Here, $V_i(\lambda_i) = w_i \lambda_i^{\beta_i}$, with $w_i > 0$, $\beta_i \in [0, 1]$. The key point is that that there is a different utility function for each firm, and users only experience network effects corresponding to other users of the same firm.

In this paper, due to space constraints, we restrict ourselves to the case $\beta_1 = \beta_2 = 0$. Our results can be extended to the case where $\beta_i > 0$; however, this case requires a more sophisticated characterization of the user response. Also, note that if $w_1 = w_2$, we recover the industry-wide network effects model. We therefore focus here on the case $w_1 \neq w_2$. Let us assume, without loss of generality, that $w_1 > w_2$.

Given a net arrival rate λ, a split (λ_1, λ_2) satisfying $\lambda_1 + \lambda_2 = \lambda$ is called a Wardrop split if, for $i = 1, 2$,

$$\lambda_i > 0 \Rightarrow w_i - f(\lambda_i, C_i) = \max_{j=1,2} \{ w_j - f(\lambda_j, C_j) \}.$$

It is easy to show that for $\lambda < C_1 + C_2$, there is a unique Wardrop split $(\lambda_1(\lambda, C_1, C_2), \lambda_2(\lambda, C_2, C_1))$. We may now define the realized traffic split as follows.

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

Results. The following theorem shows that any equilibrium is necessarily a near-monopoly for Firm 1. That is, Firm 2 can never gather more than a bounded arrival rate, and thus a negligible fraction of the user population.

THEOREM 3. *Consider a non-cooperative user base with firm-specific network effects such that $\beta_1 = \beta_2 = 0$ and $w_1 > w_2 > 0$. For large enough Λ, any equilibrium $(\lambda_1, C_1, \lambda_2, C_2)$ must satisfy*

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

Theorems 2 and 3 reveal a fundamental difference in market structure between services with industry-wide and firm-specific network effects. If the network effects are industry-wide, then multiple firms can co-exist, sharing the market (unless there is considerable asymmetry in the advertising efficiency of the firms). In other words, in such situations it is hard for firms to grab market share from each other and the distinction between the firms disappears from user's point of view. However, if network effects are firm-specific, then near-monopolies tend to emerge. Moreover, the firm that obtains a near-monopoly is the one with the 'better' service, i.e., greater network effect. Surprisingly, advertising efficiency does not help.

4. REFERENCES

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

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

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

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

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

[6] IAB Internet Advertising Revenue Report, 2011.

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

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

[9] J. Nair, A. Wierman, and B. Zwart. Provisioning of large scale systems: The interplay between network effects and strategic behavior in the user base. In *Under submission*, 2013.

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

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

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