Automatica 50 (2014) 756–767

Contents lists available at ScienceDirect

Automatica

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

Bertrand equilibria and efficiency in markets for congestible network services ☆

Parijat Dube [a], Rahul Jain [b,1]

[a] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

[b] EE & ISE Departments, University of Southern California, Los Angeles, CA 90089, USA

ARTICLE INFO

Article history:
Received 31 December 2011
Received in revised form
10 September 2013
Accepted 28 November 2013
Available online 7 January 2014

Keywords:
Queueing games
Differentiated services
Bertrand competition
Price of anarchy

ABSTRACT

This paper is motivated by study of the economics of Quality of Service (QoS) of congestible services. We introduce a queueing game framework to study such problems. We consider multiple competing providers, each offering a queued service. Users are sensitive to both access price and expected delay, and pick providers with the smallest price plus delay cost. We study equilibrium of the pricing (Bertrand) game between the congestible network service providers. We establish the existence of a Nash equilibrium under some natural assumptions. We then consider a setting with multiple classes of differentiated service. Differentiated Services (DiffServ) technologies of the Internet that can provide QoS guarantees have failed to catch on, primarily due to economic impediments. Each provider is now modeled as operating a multi-class queue. We provide sufficient conditions for the existence of a Nash equilibrium in the Bertrand (pricing) game between the providers. We characterize the inefficiency (price of anarchy) due to strategic pricing to be 2/3. Surprisingly, the price of anarchy for the multi-class setting is the same as for the single-class setting.

1. Introduction

Communication and computation are primarily *congestible services*. Communication utility of customers from Internet data speeds, wireless quality of service (QoS), etc. all depend on how many other users are using it, i.e., how much congestion or interference there is. Similarly, new technology services such as cloud computing are also congestible services since the delay in serving a computational request by a compute cloud or a data-center depends on the congestion level. To study service provisioning over a congestible network service, we will model it as a queueing system. Our concern is the impact strategic (oligopolistic) competition between service providers has on social welfare. A goal of this study is to characterize the inefficiency due to strategic competition among the providers, and provide a framework wherein efficacy of various market regulatory mechanisms can be investigated that enable social welfare optimizing outcomes. To that end, we will first consider a simple single-class *queueing game model*.

We will ignore a network's internal topology, and model autonomous network of each network service provider (NSP) as a single-server queue. Let there be N providers each of whom operates a single-class GI/G/1 queue, i.e., a queue with independent arrivals and service times with general distributions. Traffic is elastic and arriving customers cannot renege upon arrival. Upon arrival, each customer decides which queue/provider to join. If it joins a particular queue, it has to pay a service price to that provider. Each customer is delay sensitive and his expected service utility is a non-increasing function of the expected delay he suffers (exact form specified later). Thus, he joins a queue that offers the maximum expected net utility, i.e., expected service utility minus expected delay cost and price.[2] We will assume that the queues are *unobservable* in the sense that the customers cannot observe the instantaneous queue lengths but the expected waiting time is known to them. Each provider's objective is to maximize his total revenue. The players are competitive and strategic. The question at hand is what kind of outcomes may one expect with competition between the providers? How close are they to social welfare maximizing outcomes?

We then study multi-class queueing game models. Now, each of the providers operates a GI/G/1-PS queue with, say, two classes with processor-sharing between them. Traffic itself is now of multiple types with say, type 1 traffic more delay-sensitive than

☆ The second author's research was supported by the NSF grant IIS-0917410 and CAREER award CNS-0954116. The material in this paper was partially presented at the 46th Annual Allerton Conference on Communication, Control, and Computing, Monticello, Illinois, USA; International Conference on Game Theory in Networks (GameNets '09), May 13–15, 2009, Istanbul, Turkey; and the 48th IEEE Conference on Decision and Control (CDC) and 28th Chinese Control Conference (CCC), December 16–18, 2009, Shanghai, China. This paper was recommended for publication in revised form by Editor Berç Rüstem.

E-mail addresses: pdube@us.ibm.com (P. Dube), rahul.jain@usc.edu (R. Jain).

1 Tel.: +1 213 740 2246; fax: +1 213 740 4447.

2 Throughout this paper, we will deal only with expected utilities, expected delay and expected revenue. Thus, we will drop "expected" for expositional convenience.

0005-1098/$ – see front matter
http://dx.doi.org/10.1016/j.automatica.2013.12.020

type 2. Users of each type arrive according to a renewal process. Providers could charge a higher service price for class 1 than for class 2. Upon arrival, each user decides which provider/queue to join (type i users go to class i). If it joins a particular queue, it has to pay a service price to the provider. And each user joins a queue that offers it the maximum net utility, i.e., service utility minus delay cost and price. The objective of each provider is to pick service prices for the two classes as well as the *processor-sharing ratio* that maximizes his total payoff taking into account the competition provided by the other providers. Again, the question at hand is what kind of outcomes may one expect from the competition between the providers? Are the outcomes optimal or efficient in some sense? And, how does this compare with when each queue is operated as a single-class queue, i.e., are there any efficiency gains to be achieved for providing "differentiated services". We study these questions in the fluid regime for the said queueing models. This provides a framework for studying differentiated services in networks under strategic competition.

It is well-known that introduction of service differentiation in networks has failed not due to inadequacies of technological solutions (e.g., DiffServ) but primarily due to economic impediments. This became part of the *network neutrality* debate (Musacchio, Schwartz, & Walrand, 2009) wherein network service providers (NSPs) (e.g., AT&T, Verizon, etc.) argued in favor of service and price differentiation while content providers (such as Google, and Yahoo) argued in opposition. Both sides have made various arguments for and against network neutrality and service differentiation whose veracity has been hard to judge. We provide a simple modeling framework to study this issue. We caution that the larger policy issue is a lot more complicated and we cannot hope to have all the answers by studying such simple models. Nevertheless, the framework can be used to address some basic questions: Is service differentiation better for 'network utility maximization'? How much is lost in network utilization due to competition between providers when they offer differentiated services?

In this paper, we first consider a pricing game between providers that offer congestible services. Each provider is modeled as a single-class queue. Traffic is *elastic*,[3] and each arriving user joins a queue with the least full price (price plus delay cost), thus leading to a *Wardrop equilibrium*. Each provider now picks a price to maximize his revenue. Thus, there is pricing competition between the providers. We establish that under very mild conditions, a Nash equilibrium exists in this game. We note that the same model has been studied before (Musacchio, 2009; Ozdagler, 2008). Therein, a bound on the price of anarchy, i.e., the maximum relative loss in social welfare due to strategic pricing competition between the providers was shown to be 1/3. However, the existence of a Nash equilibrium was not established.

We, then, consider a two-class queueing setting since in a network, various traffic/user classes have different sensitivity to congestion/delay. Now, each provider operates a processor-sharing two-class queue, and may set different prices for different classes. We consider traffic of class i only going into class i of any queue. The question of interest now is whether one can obtain better social welfare due to service differentiation. We first give sufficient conditions for the existence of a Nash equilibrium in the multi-class pricing game. We then establish that the price of anarchy is still 2/3, i.e., there are no gains in (worst-case) equilibrium social welfare due to service differentiation.

Related literature

A pricing mechanism leading to queueing stability was first considered by Naor in Naor (1969). The effect of tolls on queueing behavior was studied with Poisson arrivals with an exponential server queue. It was shown that if balking is allowed, the socially optimal price is greater than the revenue maximizing price. The implication is that the revenue maximizing expected delay is larger than the socially maximizing one. It was shown in Edelson and Hildebrand (1975) that the two prices are the same when balking is not allowed. In Harrison (1975), a scheduling policy was derived for a multi-class queue that maximizes the expected discounted net value (service value minus delay cost). Mendelson and Whang (1990) introduced a stylized model for a single queueing service provider that has multiple priority classes. They introduced an incentive-compatible[4] priority pricing rule for the M/M/1 queue which is efficient as well (maximizes the social welfare). Many other variants of pricing by a single provider of service in a single or multi-class queue are studied in Borkar and Manjunath (2004), Hassin and Haviv (1995), Haviv (2001), Masuda and Whang (1999), and Van Mieghem (2000). A detailed survey of such work is provided in the book by Hassin and Haviv (2003).

While single provider models have been well-studied, models with multiple queueing service providers are less well understood. The earliest work on this is Luski (1976) which showed that the equilibrium prices are always different from the social welfare maximizing ones. A variation of the Mendelson–Whang model for two identical servers (with unobservable queues) was considered in Loch (1991). It was established that the total arrival rate at equilibrium in the Cournot (capacity) game is smaller than in the Bertrand (pricing) equilibrium, which is smaller than the socially optimal total arrival rate. In Lederer and Li (1997), a multi-server multi-class extension of the Mendelson–Whang model was considered. However, a competitive setting was considered where there are enough servers in the market, and the influence of each is negligible, i.e., the servers act as price takers. Thus, the existence of competitive equilibrium was established. The existence of a stochastic (competitive) equilibrium in such models was studied in Stahl and Whinston (1994).

More recently, Hayrapteyan, Tardos, and Wexler (2005) considered the Bertrand (pricing) game between multiple NSPs with linear affine delay functions (a very rough approximation for delay in queueing networks) and showed the game to have a price of anarchy (PoA) of 1/3.125. This result was sharpened by Musacchio (2009), Ozdagler (2008) to 2/3 for general concave delay functions. While a bound on PoA was obtained, it was conjectured that a Nash equilibrium may not exist in general. Thus, sufficient conditions for its existence were provided in Dube and Jain (2008), Johari, Weintraub, and Van Roy (2010). All these works focused on single class queues only. Our recent works Jain and Dube (2009) and Dube and Jain (2009) did study multi-class queues but there is only one type of traffic and allowed for priority scheduling between the classes (whereas here we consider processor-sharing). In Dube and Jain (2008), we also considered single-class queueing games where providers can choose both prices as well as service capacities, and provided sufficient conditions for the existence of a Nash equilibrium in such settings. In Dube, Touati, and Wynter (2007), the authors considered the case of inelastic traffic where each arriving user must choose among two M/G/1 queues and cannot balk. There, a Stackelberg game was considered and it was shown that an equilibrium does not exist, but an oscillatory "price wars" behavior around a limit point was noticed.

3 Traffic is elastic means that the actual demand depends on prices and delay costs.

4 Each user puts a different weight on cost due to delay. And the users do not have an incentive to lie about these weights.