An-MDP based Dynamic Pricing and Revenue Maximization in Wireless Networks

Asim Rasheed, Sung-Guk Yoon, and Saewoong Bahk
INMC, School of EECS, Seoul National University, Seoul, Korea
{asim, sgyoon}@netlab.snu.ac.kr, sbahk@snu.ac.kr

Abstract

In this paper, we propose a user centric based general framework for revenue maximization in highly competitive wireless networks. By applying dynamic pricing, each service provider operates an optimal policy that aims at maximizing its revenue. The problem is formulated using the Markov Decision Process (MDP) framework and Q-learning is applied to determine an optimal policy which maximizes the long term revenue of the service provider. Simulation results show that the service provider can maximize its revenue by applying dynamic pricing while supporting an appropriate level of user satisfaction in terms of price and call level QoS.

I. Introduction

Network heterogeneity is a common feature of 4G wireless networks. It is expected that in a near future, the users would subscribe to service providers on a short term basis, i.e., per connection duration only. In highly competitive wireless environment service providers have the potential to increase their revenue using customized price which is designed according to user's satisfactions and personal benefit. So, dynamic pricing plays a very important role in achieving the goal of revenue maximization. Our work focus on the issues related with pricing scheme raised in previous studies [1], [3].

We establish system model and assumptions in Section II. In section III, the problem is formulated and Q-learning is applied. Numerical results which show the performance are provided in Section IV. Section V concludes our paper.

II. System Model and Assumptions

We assume a common user pool that is under common service of multiple BSs. These BSs can belong to either the same service provider or different service providers. Under such environment, network users usually have freedom to associate and act "selfishly" to maximize their price aware utility. Users are uniformly distributed in this service area. As the user attachment is per connection base so connection initiations and establishment follow Poisson distribution with an average rate λ. Each connection's holding time is exponentially distributed with mean $1/\mu$. The price offered by service provider at any time greatly impacts the decision of a user to establish a connection. So, we have user connection rate as

$$\overline{\lambda} = \lambda.(\exp(-(\frac{p_b}{P_0^c}-1)))^2, \qquad (1)$$

where p_b is the price announced by service provider $b \in B$ and P_0^c is the normal price (budget) of class-*c* users.

III. Problem Definition

We formulate the dynamic pricing and revenue maximization problem under framework of Markov Decision Process (MDP) [5]. The detail problem formulation is as follows.

State-space: The network is represented by finite number of discrete time Markovian states indentified by

$$S = \sum_c n_c(t)\delta_c \le C, \qquad (2)$$

where n_c is the number of connections of service type c at time t, δ_c is bandwidth requirement of class c user, and the capacity of system is C.

Action space: The feasible offered price range is modeled in action space as $A^s = P = \left[p_1, p_2, p_3, ..., p_N\right]$ and we assume that $p_1 < p_2 ... < p_N$. At each decision epoch, the state of the wireless network changes and according to its new state the network announces a new price among feasible offered price range.

Reward Function: Because our proposed scheme is user centric, we define a global reward as the sum of all user satisfaction functions. Let $r(s,a)$ denote the immediate reward which the system get if $a \in P$ is chosen in some system state s. We design the reward function as

$$r(s,a)_{a\in P} = [y_1 u_1^+(p) - (y_1 u_1^-(p) + y_2 u_2^-(d))], \qquad (3)$$

where y_1 and y_2 are the weight associated with each user satisfaction function. We assume that system is stable under a scheduling and call admission control. As the system behavior is the same for all ongoing connections, any unstable situation means all the connections experience the same situation and that turn the immediate reward into zero. For example, a threshold based QoS user satisfaction function contributes negative towards global reward. When the QoS offered by the network exceeds the predefined threshold, all the users will churn from that network and the resulting reward will be zero. The sigmoid function has been widely used to capture user satisfaction [1], [2], [3] and the references there in. The first user satisfaction as a function of offered price p is given by [3], that is,

$$u_1^+(p) = \begin{cases} \dfrac{1}{1+e^{-L_c(P_0^c-p)}}, & p < P_0^c \\ 0 \quad , & \text{otherwise} \end{cases}. \qquad (4)$$

$$u_1^-(p) = \begin{cases} \dfrac{1}{1+e^{L_c(P_0^c-p)}}, & p > P_0^c \\ 0 \quad , & \text{otherwise} \end{cases}. \qquad (5)$$

This work is supported by "Higher Education Commission (HEC), Government of Pakistan" and the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency). (NIPA-2009-(C1090-0902-0006))

In the function, L_c represents the steepness of these functions for class-c users. To guarantee packet level QoS for ongoing connections, we model user churning behavior through packet delay. We assume that each ongoing call has certain QoS threshold requirement represented by $d_{\max}$ for class-c user. If packet delay is higher than this threshold, churning rate is likely to be higher. The packet delay d is calculated simply by M/M/1 queuing analysis [6]. So, we have the second user satisfaction function as

$$u_2^-(delay) = \frac{1}{1+e^{K_c(d^c_{\max}-d)}}, \qquad (6)$$

where K_c represents the steepness of this function for class-c users. Q-Learning [4] is a reinforcement learning technique for solving MDP problem when the state transition probabilities are not known. This technique works by directly learning MDP's action value function by interacting with control environment. If the value function is learned, the optimal policy is simply the set of actions which maximizes the function at each state. The optimal policy is given by

$$Q_t^*(s_t,a) = \max_{a\in P} Q_t(s_t,a). \qquad (7)$$

Here a policy is determined in which the action with the best Q-value is chosen. An action in each state is selected from the feasible action set using an exploitation and exploration policy.

IV. Numerical Results

We consider a single hotspot cell with a transmission range of 100 m. The mobile users are uniformly distributed and the new connection arrival rate follows Poisson distribution and connection duration is exponentially distributed with mean 1 min. Arrival rate of users is varied from 1-18 /min. The users generate persistent data traffic and always have data to send. Only single class real time traffic with a delay constraint $d_{\max}$ = 5 sec is considered. The other parameters which used in our simulation are followed: learning rate α =0.6; the discount rate γ = 0.9; P_0^c = 1.3 and p = 1.0-2.0 units/min. The constant $L_c = K_c$ = 10, the utility weight y_1 = 10 and y_2 = 5. The simulation is under varied connection arrival rate and the results are averaged over 20 runs. For the purpose of comparison, we consider a greedy scheme that always considers the maximum possible reward at any decision epoch, so it can not consider the long term effect of the policy for revenue maximization.

Fig. 1 and Fig. 2 show the mean price and the total revenue of service provider according to the arrival rates, respectively. Our proposed scheme dramatically improves both of them compared to the greedy scheme. Since the service provider chooses a higher price for higher arrival rate, our proposed scheme results in total increased revenue. The comparison is based on the average revenue generated from all connections admitted into the network.

V. Conclusion

We consider a highly competitive environment and users who are interested in the service price. Because the environment is highly competitive, current static pricing scheme causes the revenue loss. The problem is formulated under framework of MDP and Q-learning algorithm is used to get an optimal policy which maximizes the total network revenue under dynamic pricing. Numerical results show that how the mean price charged by the service provider vary at varied arrival rate. As a result, total network revenue earned by service provider is increased under our proposed dynamic pricing scheme.

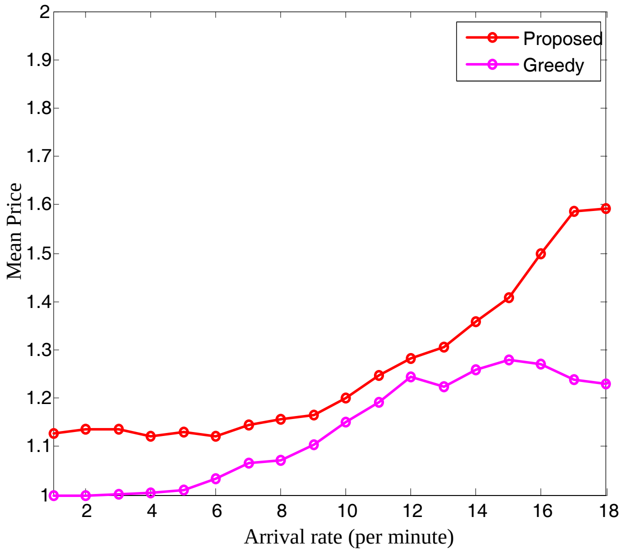

Figure 1. Average offered price vs arrival rate

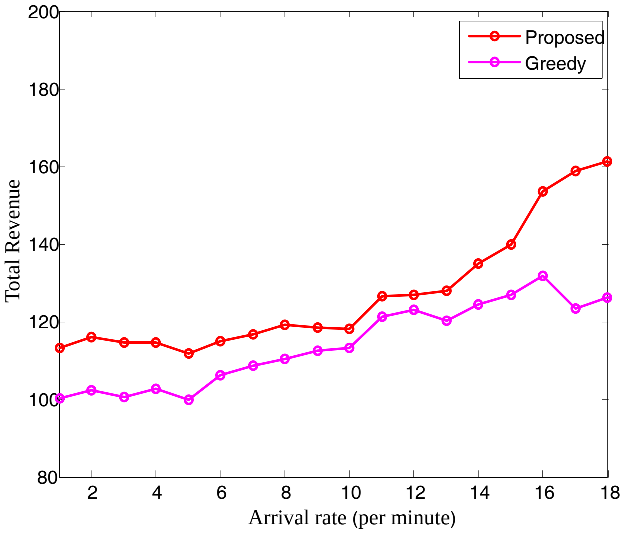

Figure 2. Total network revenue vs arrival rate

References

[1] H. Lin, M. Chatterjee, S. K. Das, and K. Basu, "ARC: An Integrated Admission and Rate Control Framework for Competitive Wireless CDMA Data Networks Using Noncooperative Games," *IEEE Transactions on Mobile Computing*, vol. 4, no. 3, pp. 243-258, May/June 2005.

[2] S. Sengupta, S. Anand, M. Chatterjee, and R. Chandramouli, "Dynamic Pricing for Service Provisioning and Network Selection in Heterogeneous Networks," *Elsevier Physical Communication*, vol. 2, no. 1-2, pp. 138-150, March/June 2009.

[3] A. N. Rouskas, A. A. Kikilis, and S. S. Ratsiatos, "A Game Theoretical Formulation of Integrated Admission and Pricing in Wireless Networks," *European Journal of Operational Research*, vol 191, no. 3, pp. 1175—1188, March 2008.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, USA: The MIT Press, 1999.

[5] M. L. Putterman, *Markove Decision Process: Discrete Stochastic Dynamic Programming*, New York: Wiley, 1994.

[6] L. Kleinrock, *Queueing Systems*, New York: Wiley, 1975, vol. 1.