A New Way of Thinking Utility in Pricing Mechanisms: A Neural Network Approach

Y. Hayel, G. Rubino, B. Tuffin [1]
IRISA, France

M. Varela [2]
SICS, Sweden

Abstract

Pricing is regarded as a solution to congestion control in telecommunication networks. Most mathematical models involve a so-called utility function accounting for the users' willingness to pay. However, this utility function is unknown in practice in terms of shape and important arguments. We propose here to limit this degree of uncertainty by aggregating all arguments in one quantity, the perceived quality of service, estimated using a Random Neural Network as a statistical learning tool according to the PSQA method. After arguing for this approach, we present a way of applying this tool to a model with two types of traffic and two classes of customers using strict priorities. We illustrate the proposal using a specific simple case.

Keywords: neural networks, pricing, queuing theory, telecommunications.

1 INTRODUCTION

Congestion control is an important issue in telecommunication networks, especially as applications become more and more demanding in terms of quality of service (QoS). Looking at the Internet, while congestion does not seem to be an issue anymore in the backbone thanks to the large capacity of core networks, the problem still persists in access networks and also in wireless networks, which are becoming ubiquitous.

Pricing has been seen as a valuable solution for controlling congestion [1]. A type of architecture that has received much attention is DiffServ (for Differentiated Services) which separates the network in classes treated differently thanks to a scheduling policy (say, strict priority). Several analytical studies of such pricing schemes can be found in the literature [4, 6]. Those mathematical models are based on a characterization of users' behaviour through a utility function representing their willingness to pay for a given value of performance.

For tractability reasons, and based on some (a priori relevant) heuristics, the shape of those functions is imposed, as well as the arguments they depend on, generally the mean delay, or the mean and/or peak throughput. However, very few studies exist on what should be the important arguments (the quality) and how they interact, as well as on how much users are willing to pay for a given quality. We propose here to base our analysis on a single quantity representing the perceived quality of service for each specific type of application. This value is determined by a Random Neural Network (RNN) used in a technique called Pseudo–Subjective Quality Assessment (PSQA), that learns from human input how to aggregate important arguments (such as delay, jitter, losses, consecutive losses, codec, etc.) into a real number Q which is close to the average quality perceived by human subjects.

The validity of the approach has been extensively investigated in [7, 9]. It presents the advantage of reducing the number of degrees of freedom of the model, which in our opinion constitutes a significant improvement over previous works. The paper is organized as follows. In Section 2, we

[1] IRISA/INRIA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France.
Emails: {yhayel,rubino,btuffin}@irisa.fr, mvarela@sics.se

[2] M. Varela's work was carried out during the tenure of an Ercim fellowship

introduce this quality assessment technique, its mathematical foundations and its practical validity. In Section 3, we present a pricing model for the DiffServ architecture that makes use of those RNN. This model is mathematically analysed in Section 4 in the case where the perceived quality of service is actually a function of the mean delay evaluated using a M/M/1 queue. Finally we conclude and give research perspectives in Section 5.

2 PSQA

When we need to determine the quality of a multimedia transmission over the Internet, the most accurate way to do it is to use a panel of humans and to show them a large enough amount of sequences. This is a standard procedure for which norms exist (e.g. ITU-T Recommendation P.800), and which gives the best results, but it is very costly. There are some methods to do an automatic quantitative assessment, that is, without using subjective tests, for audio flows, but they suffer from either the accuracy or efficiency points of view. As an alternative, the Pseudo-Subjective Quality Assessment (PSQA) technology has been recently developed. It allows to automatically quantify the quality of a video or audio communication over a packet network, as perceived by the user.

The PSQA technique is accurate, as it correlates well with the values given by panels of human observers, and it can work, if necessary, in real time. It has been tested on video [7] and audio [9] flows. It can be applied in many areas, for instance, for the analysis of the impact of different factors on quality, or for performance evaluation using standard models. For a global presentation of PSQA with a detailed description of the RNN tool, see [8] and the references therein. For the origins of RNN, see for instance [2], [3].

PSQA is based on learning, in a specific way, how human observers quantify the quality of a flow under standardized experimental conditions. The learning process consists of training a RNN to capture the relation between a set of factors having an *a priori* strong impact on the perceived quality and the latter. Let us briefly recall the structure of the RNN tool. As many other neural learning applications, PSQA is generally implemented with a 3-layer feedforward RNN. Such a network can be seen as a parametric function $\nu()$ mapping a vector of size $I + J$, denoted here $(\vec{C}, \vec{N}) = (C_1, \cdots, C_I, N_1, \cdots, N_J)$, into a real number. Let us denote by vector $\vec{W}$ the function's parameters. The input variables $C_1, \cdots, C_I$ are related to the network connection, or to the codec used (example: the bit rate), and the variables $N_1, \cdots, N_J$ correspond to the network state (example: the loss rate). The value of the function is the quality of the audio or video connection. The function's parameters are the weights in the neural network.

As a neural network, our RNN has three layers, the input one with $I + J$ variables, the hidden one with H units, and the output layer with a single node. The mapping can be explicitly written as

$$\nu(\vec{W}; \vec{C}, \vec{N}) = \frac{\sum_{h=1}^{H} \varrho_h w_{ho}^+}{r_o + \sum_{h=1}^{H} \varrho_h w_{ho}^-},$$

where

$$\varrho_h = \frac{\sum_{i=1}^{I} \frac{C_i}{r_i} w_{ih}^+ + \sum_{j=1}^{J} \frac{N_j}{r_j} w_{jh}^+}{r_h + \sum_{i=1}^{I} \frac{C_i}{r_i} w_{ih}^- + \sum_{j=1}^{J} \frac{N_j}{r_j} w_{jh}^-}$$

is the *activity rate* of hidden neuron h. The strictly positive numbers r_o, r_h for $h = 1..H$, r_i for $i = 1..I$ and r_j for $j = 1..J$ are fixed. They correspond to the firing rates of the neurons in the network (respectively, for the output one, the hidden nodes, and the $I + J$ input ones). The weights are the variables tuned during the learning process. We denote by w_{uv}^+ (resp. by w_{uv}^-) the weight corresponding to an exiting (resp. inhibiting) signal going from neuron u to neuron v (observe that both numbers w_{uv}^+ and w_{uv}^- are ≥ 0). For the interpretation and the dynamics of a RNN see the cited references above. For our purposes here, we can just see it as a rational parametric function. Learning will thus consist of finding appropriate values of the weights capturing the mapping from $(\vec{c}^{(k)}, \vec{n}^{(k)})$ to the real number $q^{(k)}$ where $q^{(k)}$ is the quality given by a panel of human observers to some audio or video sequence (depending on the application) when the source parameters had the values present in $\vec{c}^{(k)}$ and the parameters characterizing the network had the values in vector $\vec{n}^{(k)}$, for $k = 1..K$.

An important property of the PSQA metric is that it provides us with a closed-form expression for the perceived quality. Moreover, it has been shown that the perceived quality can be estimated reasonably well with a very simple RNN, for which the quality expression is also very simple. We can then fix some of the input variables and get a very simple (and quite accurate) expression of quality as a function of one or two of the most important ones.

3 PRICING MODEL

The pricing model we consider is taken from [4, 6]. Basically, we focus on two types of traffic: voice (indexed by v) and data (indexed by d), and priority queuing at a bottleneck modelled by a queue. Each voice (resp. data) user is assumed to send packets at rate λ_v (resp. λ_d). We assume that there are two classes of service, with class-1 being served with preemptive priority with respect to class-2, and such that the per-packet price p_1 of class-1 is larger than p_2, the price for class-2, that is $p_1 > p_2$. We investigate two strategies: the case of dedicated classes where voice users (resp. data) are forced to go to class-1 (resp. class-2), and the case of open classes where users can choose between service classes. We assume that there is an infinite population of potential customers that join the network as long as their utility exceeds the price for service.

As a consequence, there is a game played at the customer level on sending or not traffic, doing so if their utility is positive and leaving it otherwise, but analyzed at the class level since it may lead to an equilibrium over the number of customers of each type in each class. See [4, 5] for extensive discussions on this topic. In those papers, the utility function is chosen arbitrarily and depends on the mean delay D_i experienced in class-i $(= 1, 2)$ by $u_d(D_i) = 1/D_i^{\alpha_d}$ for data users and $u_v(D_i) = 1/D_i^{\alpha_v}$ for voice users. In order to express the preference of voice users for small delays, $\alpha_v > \alpha_d$. This presents a main drawback, namely the fact that we are using arbitrary functions $1/x_j^{\alpha}$, $j \in \{d, v\}$, for the utility.

In this paper, we aim at being both more general and more precise, using explicitly the approximation of perceived quality provided by the PSQA approach. Moreover, we do not use arbitrary utility functions $u_d()$ and $u_v()$, but rather root our choice on practical observations:

- voice users are interested in obtaining at least a basic quality level (in terms of voice clarity, absence of artefacts, etc.) which is defined mainly by the network conditions. Therefore, we consider a stair-step utility function of general form

$$f_v(Q) = \sum_{k=1}^{K} a_k 1_{[h_{k-1} \le Q < h_k]}$$

  where some quality level Q between h_{k-1} and h_k yields a willingness to pay a_k. We have $0 = h_0 < h_1 < \cdots < h_K$ and $a_1 < a_2 < \cdots < a_K$. The thresholds h_k can come from well-defined points in, say, MOS ranges, and the prices a_k are assumed to come from extensive testing with real users.

- Concerning data users, their willingness to pay for a given quality Q can be determined by tests through an RNN similarly to what is done for the quality with respect to performance parameters. Using the simplest 2-layers topology for the RNN, we obtain an utility function having the form

$$f_d(Q) = \frac{Q + \alpha_d}{\beta_d Q + \gamma_d}$$

  for some real numbers $\alpha_d, \beta_d, \gamma_d$.

The above setting allows for a numerical analysis of the equilibrium point $(N^*_{1,d}, N^*_{1,v}, N^*_{2,d}, N^*_{2,v})$ for a given set of prices, where $\forall i \in \{1,2\}$, $j \in \{d,v\}$, $N^*_{i,j}$ is the equilibrium number of type-j customers using class-i, as well as, in a second step, the prices optimizing the network revenue

$$R(p_1,p_2) = \sum_{i\in\{1,2\}} \sum_{j\in\{v,d\}} \lambda_j N^*_{i,j} p_i.$$

4 MATHEMATICAL ANALYSIS IN A SIMPLE CASE

Let us illustrate more in deep our model in a particular simple case. We consider the typical situation where the bottleneck is modelled by an M/M/1 queue with service rate μ. In order to derive analytical results while keeping somehow close to the previous published work, and for comparison purposes, we limit ourselves to the case where the perceived quality is a function of mean delay only. We thus have for a given delay D and $\forall j \in \{v,d\}$

$$Q_j = \frac{D+d_j}{b_j D + c_j}.$$

Note that for data users, combining the two rational functions, we still obtain a rational function of form $f_d(D) = (D+d'_d)/(b'_d D + c'_d)$ (we abusively use the same notation f_v and f_d in terms of D instead of Q).

4.1 Case of dedicated classes

Focus on the case of dedicated classes ($N_{1,d} = N_{2,v} = 0$), where voice packets, more sensitive to delay, are forced to class-1 (the higher priority class) and while data uses class-2. Define $N_v = N_{1,v}$ and $N_d = N_{2,d}$. If there are N_v voice customers in the queue, class-1 delay is given by

$$D_1 = \frac{1}{\mu - N_v\lambda_v}.$$

Voice users enter as long as $f_v(D_1) \geq p_1$. $f_v(D_1)$ is a decreasing function of N_v. Let k be the smallest integer such that $a_k \geq p_1$ (if $p_1 > a_K$ then no voice customer enters the network). The population cardinality N_v will increase up to the highest value N^*_v such that we still have $f_v(D_1) \geq p_1$. This gives $Q_v = h_{k-1}$, that is $D_1 = (\mu - \lambda_v N^*_v)^{-1} = Q_v^{-1}(h_{k-1})$. Therefore

$$N^*_v = \frac{1}{\lambda_v}\left(\mu - \frac{1}{Q_v^{-1}(h_{k-1})}\right) = \frac{1}{\lambda_v}\left(\mu - \frac{b_v h_{k-1} - 1}{d_v - c_v h_{k-1}}\right).$$

Similarly, data users enter class-2 traffic as long as $f_d(D_2) \geq p_2$. Using classical queuing results,

$$D_2 = \frac{\mu}{(\mu - \lambda_v N^*_v)(\mu - \lambda_v N^*_v - \lambda_d N_d)},$$

N^*_v being fixed from previous computation. Note that $f_d(D_2)$ is strictly decreasing in N_d and continuous. Thus, there is a unique equilibrium point N^*_d. If for $N_d = 0$, $f_d(D_2) \leq p_2$, i.e., if $\mu(\mu - \lambda_v N^*_v)^{-2} \geq f_d^{-1}(p_2)$, then no data user will join ($N^*_d = 0$). Otherwise, the unique N^*_d is such that $f_d(D_2) = p_2$, i.e., $D_2 = f_d^{-1}(p_2) = (c'_d p_2 - d'_d)/(1 - p_2 b'_d)$, or more explicitely:

- if $p_1 < a_K$, then

$$\begin{aligned} N^*_d &= \frac{1}{\lambda_d}\left(\frac{1}{Q_v^{-1}(h_{k-1})} - \frac{\mu Q_v^{-1}(h_{k-1})}{c'_d p_2 - d'_d}(1 - b'_d p_2)\right) \\ &= \frac{1}{\lambda_d}\left(\frac{b_v h_{k-1} - 1}{d_v - c_v h_{k-1}} - \mu\frac{1 - b'_d p_2}{c'_d p_2 - d'_d}\frac{d_v - c_v h_{k-1}}{b_v h_{k-1} - 1}\right). \end{aligned}$$

- if $p_1 > a_K$, i.e. if $N_v^* = 0$, then

$$N_d^* = \frac{1}{\lambda_d}\left(\mu - \frac{1-p_2 b_d'}{c_d' p_2 - d_d'}\right).$$

The network revenue is $R_D(p_1,p_2) = \lambda_v N_v^* p_1 + \lambda_d N_d^* p_2$. The subscript D denotes the dedicated classes situation. A numerical characterization of prices optimizing the revenue can be processed as follows. We first find the optimal value of low priority access price p_2, for a given value of the high priority price p_1, and consequently, for a fixed h_{k-1}. The revenue of the system depends on p_2 through

$$R_D(p_2) \quad = \quad p_1\left(\mu - \frac{1}{Q_v^{-1}(h_{k-1})}\right) + p_2\left(\frac{1}{Q_v^{-1}(h_{k-1})} - \frac{\mu Q_v^{-1}(h_{k-1})}{c_d' p_2 - d_d'}(1-b_d' p_2)\right).$$

Obtaining the optimal p_2 for a given p_1 is simple from a numerical point of view ($p_2 \geq 0$). The optimisation is then reduced to the parameter p_1. Though, $\forall k$, for each p_1 in the interval (a_{k-1}, a_k), demand is fixed. A discrete optimization can then be carried out over $p_1 \in \{0\} \cup \{a_1, \cdots, a_K\}$.

4.2 Case of open classes

The case of open classes can be analyzed in a similar way. Consider first class-1 independently of class-2 (since the former has a strict and preemptive priority over the latter). Let $N_{1,d}$ and $N_{1,v}$ be the number of voice and data users in competition for this class of traffic.

- Voice users enter the network as soon as $u_v(D_1) > p_1$ with $D_1 = (\mu - \lambda_v N_{1,v} + \lambda_d N_{1,d})^{-1}$. Let again k be the smallest integer such that $a_k \geq p_1$ (if $p_1 > a_K$, i.e. $N_v^* = 0$). For fixed $N_{1,d}$, $N_{1,v}$ will increase up to $D_1 = (\mu - \lambda_v N_{1,v} + \lambda_d N_{1,d})^{-1} = Q_v^{-1}(h_{k-1})$, i.e.

$$\lambda_v N_{1,v} + \lambda_d N_{1,d} = \mu - \frac{1}{Q_v^{-1}(h_{k-1})}.$$

- Similarly in the case of data users, for fixed $N_{1,v}$, $N_{1,d}$ will increase up to $u_d(D_1) = p_1$, leading to

$$\lambda_v N_{1,v} + \lambda_d N_{1,d} = \mu - \frac{1-p_1 b_d'}{c_d' p_1 - d_d'}.$$

Therefore, following the same principles as in [5], if $\mu - Q_v^{-1}(h_{k-1}) < \mu - (1-p_1 b_d')/(c_d' p_1 - d_d')$, the couple $(N_{1,v}, N_{1,d})$ will increase up to $u_v - p_1 = 0$, while $u_d - p_1$ will still be positive. Thus $N_{1,d}$ will increase, while $N_{1,v}$ will decrease down to 0 (because of a negative utility). $N_{1,d}$ will continue to increase up to the value such that $u_d - p_1 = 0$, where $u_v - p_1 < 0$, deterring voice users from entering. This leads to the following equilibrium point:

$$\left(N_{1,v}^* = 0, N_{1,d}^* = \frac{1}{\lambda_d}\left(\mu - \frac{1-b_d' p_1}{c_d' p_1 - d_d'}\right)\right).$$

In a symmetric way, if $\mu - Q_v^{-1}(h_{k-1}) > \mu - (1-p_1 b_d')/(c_d' p_1 - d_d')$, the equilibrium will be $(N_{1,v}^* = N_v^*, N_{1,d}^* = 0)$ with N_v^* the above value in the case of dedicated classes.

As a consequence, there will always be only one type of traffic in class-1, data if $\mu - Q_v^{-1}(h_{k-1}) < \mu - (1-p_1 b_d')/(c_d' p_1 - d_d')$, and voice otherwise.

The number of users being fixed for class-1, the analysis can be repeated for class-2. Again, the numbers $N_{2,v}$ and $N_{2,d}$ of customers in class-2 increase up to $u_v = 0$ i.e., $D_2 = Q_v^{-1}(h_{l-1})$ with l be the smallest integer such that $a_l \geq p_2$ (if $p_2 > a_{K'}$, with K' the quality level for the first infinitesimal class-2 user when class-1 fixed as above, then no voice customer uses class-2), or $u_d = 0$, i.e., $D_2 = (c'_d p_2 - d'_d)/(1 - b'_d p_2)$, that is respectively

$$\lambda_d N_{2,d} + \lambda_v N_{2,v} = \mu - \lambda_1 N_1^* - \frac{\mu}{(\mu - \lambda_1 N_1^*) Q_v^{-1}(h_{l-1})}$$

or

$$\lambda_d N_{2,d} + \lambda_v N_{2,v} = \mu - \lambda_1 N_1^* - \frac{\mu(1 - b'_d p_2)}{(\mu - \lambda_1 N_1^*)(c'_d p_2 - d'_d)},$$

where $\lambda_1 N_1^*$ is the total arrival rate in class-1 (depending on the value of p_1). So, following the same line of argument than for class-1, there will be only one type of traffic in class-2, data if $Q_v^{-1}(h_{l-1}) < (c'_d p_2 - d'_d)/(1 - b'_d p_2)$, with $N_{2,d}^* = \lambda_d^{-1}[\mu - \lambda_1 N_1^* - \mu(1 - b'_d p_2)]/(\mu - \lambda_1 N_1^*)/(c'_d p_2 - d'_d)$ and voice otherwise, with $N_{2,v}^* = \lambda_v^{-1}[\mu - \lambda_1 N_1^* - \mu/(\mu - \lambda_1 N_1^*)/Q_v^{-1}(h_{l-1})]$.

A numerical investigation of prices maximizing the revenue can be carried out similarly to the case of dedicated classes, but is not included here for sake of space.

5 CONCLUSIONS AND PERSPECTIVES

This paper aims at proposing the combination of pricing analysis with the PSQA technique which is able to automatically quantifying the perceived quality of a video, audio or multimedia communication through a packet network. The goal is to avoid the use of somehow arbitrary utility functions taking into account the way users see the benefit got from the transport of their packets. We included the PSQA evaluation of perceived quality into a model representing two typical and important types of traffic, voice and data, having very different quality constraints. The approach was illustrated using a simple model where packets are handled using two classes with priorities.

This work can be extended to more complex situation using numerical procedures, in particular to models where the quality function depends on more than one parameter. The methodology is the same, but the complexity of the models precludes any attempt of obtaining analytical results. Another aspect of this paper needing more development is the experimental one, and specifically the way the necessary input data will be effectively produced. This will be the object of future efforts.

References

[1] C. Courcoubetis and R. Weber. *Pricing Communication Networks—Economics, Technology and Modelling*. Wiley, 2003.

[2] E. Gelenbe. Random Neural Networks with negative and positive signals and product form solution. *Neural Computation*, 1(4):502–511, 1989.

[3] E. Gelenbe. Learning in the recurrent Random Neural Network. *Neural Computation*, 5(1):154–511, 1993.

[4] Y. Hayel, D. Ros, and B. Tuffin. Less-than-Best-Effort Services: Pricing and Scheduling. In *IEEE INFOCOM 2004*, Hong-Kong, China, March 2004.

[5] Y. Hayel and B. Tuffin. Pricing for heterogeneous services at a discriminatory processor sharing queue. In *4th IFIP-TC6 Networking Conference*, Waterloo, Canada, June 2005.

[6] M. Mandjes. Pricing Strategies under Heterogeneous Service Requirements. In *IEEE INFOCOM*, 2003.

[7] S Mohamed and G. Rubino. A study of real–time packet video quality using Random Neural Networks. *IEEE Transactions On Circuits and Systems for Video Technology*, 12(12):1071 –1083, December 2002.

[8] G. Rubino. Quantifying the Quality of Audio and Video Transmissions over the Internet: the PSQA Approach. In *Design and Operations of Communication Networks: A Review of Wired and Wireless Modelling and Management Challenges*, Edited by J. Barria. Imperial College Press, 2005.

[9] G. Rubino, M. Varela, and S. Mohamed. Performance evaluation of real-time speech through a packet network: a Random Neural Networks-based approach. *Performance Evaluation*, 57(2):141–162, May 2004.