Throughput prediction in wireless networks using statistical learning

Claudina Rattaro , Pablo Belzarena
*Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay.

***Abstract*—The focus of this work is on the estimation of throughput in wireless networks, more specificaly on IEEE 802.11. Our proposal is based on active measurements and statistical learning tools. We present a methodology where the system is trained during short periods with application flows and probe packets bursts. We learn the relation between throughput obtained by the application and the state of the network, which is inferred from the interarrival times of the probe packets bursts. As a result we obtain a continuous non intrusive methodology that allows to determine the maximum throughput of a wireless connection only knowing some characteristics of the network. We use Support Vector Machines (SVM) for regression and we show results obtained by simulations.**

Key words: Statistical learning, Support Vector Machines, Wireless networks

I. Introduction

In recent years, Wireless Local Area Networks (WLANs) have become increasingly popular. WLANs play a key role in providing internet connections. The family of IEEE 802.11 [7] protocols has become the most popular access method for WLANs. In 802.11 protocols, the fundamental medium access method is DCF (Distributed Coordinated Function), a form of carrier sense multiple access with collision avoidance (CSMA/CA).

Up to now, most research works in this area have been developed to model IEEE 802.11 DCF and evaluate its performance analytically (examples can be found in [3], [5]) but there are few works that focus on the problem of estimating quality of service (QoS) parameters by measurements in WLAN.

We are working on a monitoring system for wireless networks, in particular, for Wireless Mesh Networks (WMN,[1]). This paper focus on one issue of such monitoring system: the prediction of the throughput of a new connection in a 802.11 wireless link. The prediction methodology that we propose can be used by different applications like: admission control, QoS estimation, load balancing algorithms, etc..

There are other works that use measurements in wireless networks for different applications like admission control (see for example [6] and the reference therein). However, most of these works do not predict the maximum throughput of a new connection. The main drawback of the works that predict the maximum throughput, is that the measurement procedures used for estimating are strongly intrusive, affecting the QoS of the incumbent connections. However, in order to develop a monitoring system, one important feature is measuring the different parameters in a non-intrusive way. Therefore, we propose a methodology that predict the maximum throughput using only light probe packets that do not overload the wireless link. There are other works that use the same technique but in wired networks (an example is [2]).

In this paper, we focus on predicting the throughput, but the methodology proposed can be also used to estimate other QoS parameters seen by the application traffic (like delay or jitter for example).

We organize this paper in the following way. In section II we introduce the methodology. In section III we describe the network's state estimator. In section IV we show results from different simulations and finally, we conclude and discuss future work in section V.

II. Methodology

We propose a methodology that consists in learning the relation between the probe packets interarrival times and the user's throughput. Once the relation has been learned, we may predict the throughput just by sending light probe packets.

We consider the regression model: $Y = \Phi(X) + \epsilon$, where X, Y and ϵ are random variables. In this work the variable Y represents the new connection throughput, X is an estimation of the state of the wireless link and ϵ is the error. The estimator of the state of the wireless link X is obtained from the probe packets interarrival times as is explained in section III.

To estimate the function Φ we use Support Vector Machines (SVM), in particular, we use SVM for the regression (SVR) [9].

We consider an environment consisting of a reference node (RN) and a number of fixed nodes distributed over the network area. The nodes in the network will contend to send data packets using the DCF protocol.

In order to estimate Φ we divide the experiment into two phases. The first phase is called the learning phase. During the learning phase, when a new client (NodeX) connects to the reference node RN and starts sending traffic we measure the new connection throughput (Y_i) of NodeX. Afterwards, when this transmission finishes, NodeX sends a burst of probe packets with fixed size and interdeparture times. We build the variable X_i by measuring in each experiment i the interarrival times of the probe packets burst. Therefore, in each experiment, in the learning phase we have a pair (X_i, Y_i). After we have collected a set of samples (X_i, Y_i), we estimate

the function Φ using SVM. We call $\widehat{\Phi}$ the estimation of the function Φ.

After the system was trained in the learning phase, the second phase, called the monitoring phase, starts. During the monitoring phase NodeX only sends probe packets. We build the variable X in the same way as in the learning phase. The throughput of a new connection $\widehat{Y}$ is estimated using the function $\widehat{\Phi}$ built in the learning phase by $\widehat{Y} = \widehat{\Phi}(X)$. We remark that this procedure does not load the network during the monitoring phase because it does not send the packets to measure the maximum throughput. The model is robust to spatial/temporal variation of the wireless medium, it only needs piecewise stationarity.

III. The estimator of the state of the wireless link, X

In this section we will describe the estimator X, that is the estimator of the state of the wireless link. This state will be estimated from the probability distribution of the variable delay seen by the probe packets.

In 802.11 links, there are many factors that influence the state of the wireless link, but the most important of them are the collision probability (p) and the channel interference (I).

We consider a probe packet n that arrives to the queue of the wireless link at time t_n^i and leaves the link at time t_n^o. If the latency of the link is D, the free capacity is $C_n(p, I)$ (we consider the case of adaptive multirate where the link capacity varies with p and I), P is the packet's size and $V_n(p, I)$ represents the delay caused by retransmissions. The difference between t_n^o and t_n^i is such that:

$$t_n^o - t_n^i = \frac{P}{C_n(p, I)} + D + V_n(p, I) \qquad (1)$$

Some factors in (1) are constant: example: $Cmax$, min delay, etc, so we can express the equation (1) as: $t_n^o - t_n^i = K + K_n(p, I)$

Therefore, we can consider instead the following equation:

$$K_n = K_{n-1} + (t_n^o - t_{n-1}^o) - (t_n^i - t_{n-1}^i) \qquad (2)$$

Equation (2) is obtained by considering two consecutive packets, $n-1$ and n, and the difference between the following equations: $t_n^o - t_n^i = K + K_n(p, I)$ and $t_{n-1}^o - t_{n-1}^i = K + K_{n-1}(p, I)$.

Applying equation (2) recursively we have:

$$K_n = K_0 + \sum_{j=1}^{n} [(t_j^o - t_{j-1}^o) - (t_j^i - t_{j-1}^i)] \qquad (3)$$

Equation (3) allows us to estimate the probability distribution of the variable component of the delay using only the arrival times and departure times.

Remark 1: K_0 can be set to zero starting the sequence where some consecutive packets have interarrival times equal to the interdeparture time.

In figures 1 and 2 we show the empirical distribution of K_n for different states of the wireless link.

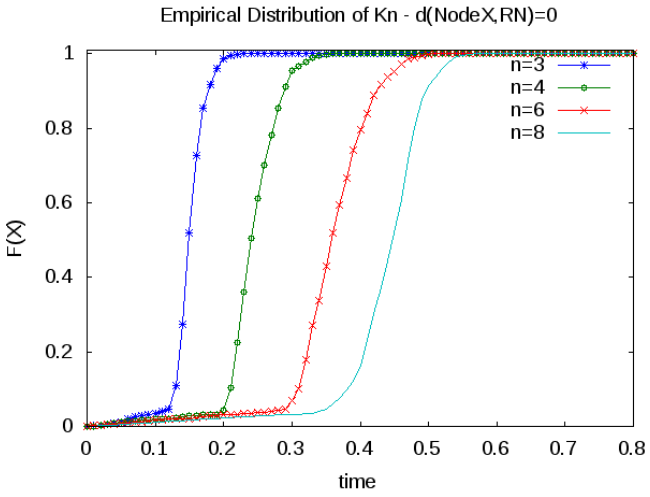

Fig. 1. Empirical distribution of K_n for different states of the wireless link. Changing the number of fixed nodes (n) in the wireless network. The distance between RN and NodeX is the same in all the simulations

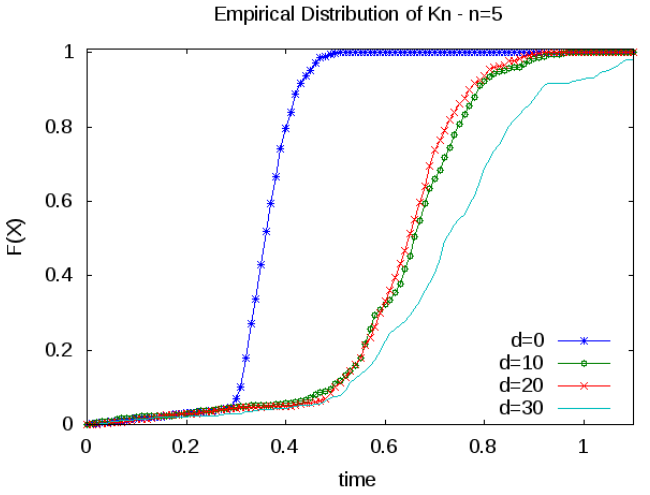

Fig. 2. Empirical distribution of K_n for different states of the wireless link. Changing the distance (d) between NodeX and RN. The number of fixed nodes is the same in all the simulations

We will use as variable X (estimator of the wireless link's state) some statistics of K_n like the expected value, variance, etc..

IV. Simulations

We will consider an environment that consists of a reference node (RN) and a number of fixed nodes randomly distributed over the network area (we use Poisson's distribution). We will call *constellation* the set of fixed nodes. We will use IEEE 802.11g in all the simulations and ARF (Auto Rate Fallback) as the algorithm of Rate Adaptation. The methodology used in order to train the system and to predict the throughput of the new connection is the explained in section II. All simulations were done using the Ns-2 simulator [8]. The training and the prediction using SVM were done with the libsvm library [4].

Following, two scenarios are presented: saturated and unsaturated traffic. In each simulation, we change the distance between the reference node and the NodeX, also we change the number of nodes in the constellation and their positions.

A. Saturated traffic conditions

We want to estimate the maximum throughput of a wireless connection in a situation where the network is saturated, in this case, we train with nodes that always have packets to send. We modeled the traffic of all the nodes as CBR/UDP with rate=20Mbps.

General characteristics:

- Topology: RN + n nodes + NodeX
- Traffic Type: All nodes (also the NodeX) generate CBR/UDP uplink (Rate= $20Mbps$ Packet Size= $1500bytes$)
- Probe Packets: UDP, Packet Size= $10bytes$, Interval= $1ms$
- The positions of the n nodes are randomly choosen with Poisson distribution.

Remark 2: There is a trade-off between the size and interarrival time of the test packets, since the objective is not to affect the network performance. We did several tests and we concluded that the values: Packet Size= $10bytes$ and Interval= $1ms$ are sufficient to achieve a good estimate without affecting the system.

Remark 3: The simulation throughput is calculated considering the bytes received and the total transfer time (in saturated conditions).

In order to get many values to train and to verify the model, we made several simulations, increasing n from 1 to 10 and changing the distance between Node X and RF (d).

Each simulation is divided into three well defined parts, being t the the simulation's time:

- Part 1 ($t = 0$, $t = 50$): The n nodes of the constellation start sending packets. They continue injecting traffic to the network in all the simulation. We wait until the system reaches stationary state.
- Part 2 ($t = 50$, $t = 100$): The NodeX starts sending packets (it finishes at $t = 100$).
- Part 3 ($t = 130$, $t = 140$): The NodeX sends the test packets.

From each simulation we obtain as an output:

- The throughput of NodeX
- X, the estimator of the wireless link state.

We use the mean value of K_n as the estimator X. Figure 3 shows the measured throughput of the new connection and its prediction using as X only the mean value of K_n for each sample (X_i). As we can see in this figure the estimation is accurate. In this case, the results of SVM are:

Mean squared error	0.282229
Squared correlation coefficient	0.868246

In the example of figure 3, we made four sets of simulations: One of them was used to train and the others to verify the model. The four sets of simulations were done with different topologies. In each one, $n = 1 : 1 : 10$ and $d = 0 : 1 : 40$, n is the number of nodes in the constellation and d is the distance between RN and NodeX. This explains the shape of the graph.

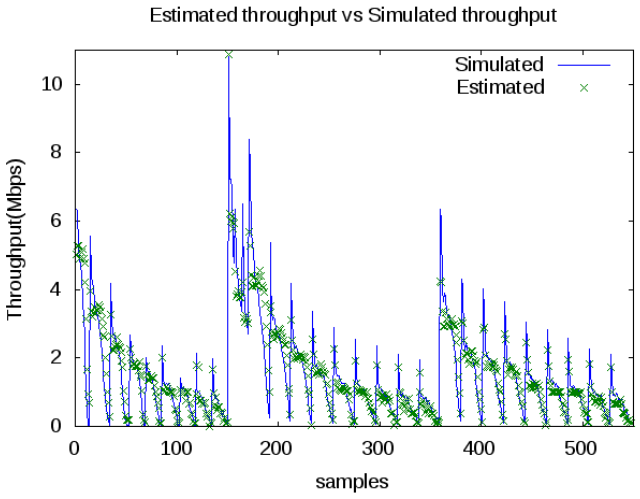

Fig. 3. Measured and predicted Throughput using only the mean value of K_n

We remark that training with about 100 samples good results are obtained: Squared correlation coefficient > 0.80. This accuracy is enough for taking some decisions in wireless networks.

B. Nonsaturated traffic conditions

In this section, the objective is the estimation of the maximum throughput of a wireless connection in a situation where the network is not saturated. Therefore, we train with nodes that are making small file downloads or Internet browsing. We model this traffic as an exponential on-off process.

In the present work, we focus on the case of fixed rate in physical layer, which means that, we disable the ARF mechanism.

General characteristics:

- Topology: RN + n nodes + NodeX
- Traffic Type: The nodes of the constellation generate Exponential On-Off over UDP (burst time= $0.5s$, idle time=$0.5s$, mean rate=$500kbps$, Packet Size= $1500bytes$)
- Probe Packets: UDP, Packet Size= $10bytes$, Interval= $1ms$
- The positions of the n nodes are randomly choosen off with Poisson distribution.
- The NodeX generates CBR/UDP uplink traffic (Rate=$20Mbps$ Packet Size= $1500bytes$)
- Physical Layer Rate= $12Mbps$

We proceed the same way as above, making several simulations, increasing n and changing the distance between Node X and RF (d).

The simulation can be decomposed into three steps. Following we present each of them.

- Step 1 ($t = 0$, $t = 50$): The n nodes of the constellation start sending packets. They continue injecting traffic to the network during all the simulation. We wait until the system reaches stationary state.
- Step 2 ($t = 50$, $t = 100$): The NodeX starts sending packets with CBR/UDP traffic, and finishes at $t = 100$.
- Step 3 ($t = 130$, $t = 170$): The NodeX sends the test packets.

In figures 4 we show the empirical distribution of K_n for different states of the wireless link.

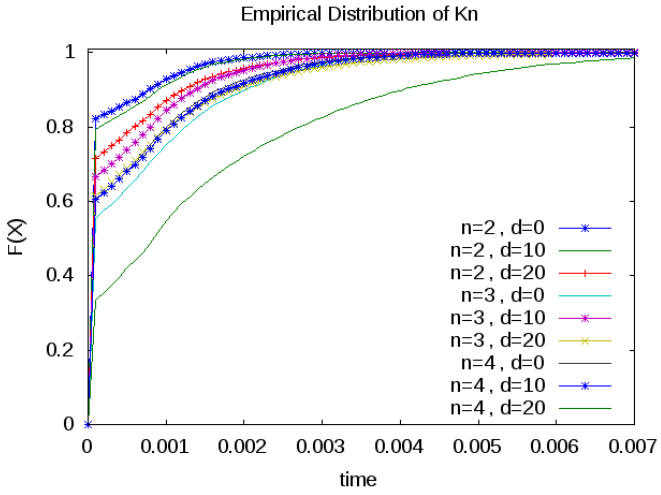

Fig. 4. Empirical distribution of K_n for different states of the wireless link. We changed the distance (d) between NodeX and RF and the number of fixed nodes (n)

After several tests, we obtained the best accuracy of the throughput using the following three parameters as X:

- Mean and variance of the estimator K_n
- Value of the empirical distribution of K_n in 0 (minimum delay achieved under the wireless link conditions).

Considering these three parameters, we used SVM and obtained encouraging results. They are in the following table and in figure 5. In this situation, we made one set of simulations to train and another to verify the model.

Mean squared error	0.371387
Squared correlation coefficient	0.897626

V. Conclusion

This work estimates the throughput seen by applications in a wireless link. The main contributions of this paper are the proposed estimator of the wireless link state and the methodology presented that uses SVM and probe packets in order to predict the maximum throughput of a new connection. In addition, the proposed methodology is a non intrusive procedure.

This statistical learning approach gives accurate throughput prediction in both situations: saturated and non-saturated traffic conditions.

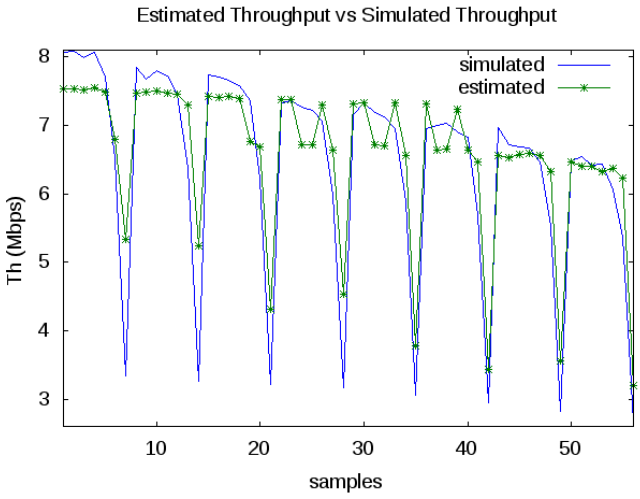

Fig. 5. Simulated and estimated throughput for the non saturated case, using mean, variance, and minimum delay.

This work has many future lines of research like: extending and applying this procedure to develop an admission control algorithm for multi-hop wireless networks, predicting other QoS parameters seen by applications in wireless networks (like delay, jitter, etc.) and applying the previous methodology to 802.11e wireless networks.

We are working on improving the estimation for the case of unsaturated traffic conditions. We are interested on investigating other features of K_n that help to raise the accuracy of the estimation. In addition we want to extend the analysis to the case of variable rate.

Acknowledgment.

This work was supported by the project *CSIC-UdelaR: Modelado y evaluación del desempeño de redes inalámbricas estructuradas y mesh.*

References

[1] I.F. Akyildiz and Xudong Wang. A survey on wireless mesh networks. Communications Magazine, IEEE, 43(9):S23S30, September 2005.
[2] P. Belzarena, L. Aspirot, End-to-end quality of service seen by applications: a statistical learning approach Computer Networks (March 2010)
[3] G. Bianchi. Performance Analysis of IEEE 802.11 Distributed Coordination Function. IEEE Journal on Selected Areas in Communications, 18(3):535-547, Mar. 2000.
[4] C-C. Chang, C-J. Lin, LIBSVM: a library for support vector machines, url:http://www.csie.ntu.edu.tw/ cjlin/libsvm, (2001).
[5] F. Daneshgaran , M. Laddomada , F. Mesiti , M. Mondin, On the Behavior of the Distributed Coordination Function of IEEE 802.11 with Multirate Capability under General Transmission Conditions, IEEE Transactions on Wireless Communications, October 21, 2007.
[6] A. Herms and G. Lukas. Preventing admission failures of bandwidth reservation in wireless mesh networks. Computer Systems and Applications, ACS/IEEE International Conference on, 0:10941099, 2008.
[7] IEEE 802 Standard Working Group, Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: further higher data rate extension in 2.4 GHZ Band, IEEE 802.11g Standard, 2003.
[8] McCanne S., Floyd S., ns network simulator, url:http://www.isi.edu/nsnam/ns/.
[9] V.N. Vapnik, The nature of statistical learning theory, Springer NY, 1995.