Centre
Tecnològic
de Telecomunicacions
de Catalunya

Methodology for Traffic Load Estimation in WLANs Based on Real Traces

Andrey Krendzel, Marc Portoles-Comeras, Josep Mangues-Bafalluy

Publication:	Proc. 14th European Wireless Conference 2008
Vol:	
No.:	
Date:	June 2008

Methodology for Traffic Load Estimation in WLANs Based on Real Traces

Andrey Krendzel, Marc Portoles-Comeras, Josep Mangues-Bafalluy
IP Technologies Area
Centre Tecnològic de Telecomunicacions de Catalunya Barcelona (CTTC), Barcelona, Spain
{andrey.krendzel, marc.portoles, josep.mangues}@cttc.es

Abstract—**This paper presents a methodology to estimate traffic load parameters in a WLAN network consisting of a large number of access points (APs). The methodology takes into account the high variability of data traffic throughout the network that APs have to handle and process. It is based on using a low number of initial data formed from known statistical data about behavior of the network. The methodology exploits the inequality in AP popularity along the wireless network to estimate traffic load parameters. It is validated using real WLAN traces of a popular SNMP data collection of Dartmouth College. The methodology provides the traffic load estimations that coincide with the results of actual load measurements when initial input are extracted from everyday real WLAN traces. It also provides appropriate results when some of the initial data for the model are formed by averaging over a certain arbitrary long time period.**

Keywords-Access Points, Inequality, Gini coefficient, Lorenz curves, traffic load estimation, WLAN networks

I. INTRODUCTION

WLAN networks are emerging as an attractive solution for users to access a large variety of services at any time from locations of the most diverse nature. In accordance with [1] the services may use different kinds of media (audio, visual, audiovisual), be real-time and non-real-time, delay-sensitive and delay-tolerant, guaranteed or not, and have different bandwidth demands. Besides, in currently deployed networks, a high variance is observed in the activity of the users in buildings (depending on their type: residential, academic, library, administrative, etc.), and usage of access points (APs) and cards over both time and space [2]. These issues lead to finding *a large diversity of data traffic handled and processed by APs* throughout a WLAN network consisting of a large number of APs. Due to all this, it is becoming more and more complex to make an appropriate estimation of the offered traffic load generated by thousands of users and transferred to/from hundreds of APs in the WLAN network.

There have been some publications in the literature concerning traffic load issues in wireless networks, in particular, statistical load measurement observations in wireless LAN [2-4], load balancing in dense wireless multi-hop wireless networks [5],[6], spatio-temporal modeling of traffic workload in a Campus WLAN [7], and others. However, there have not been any publications (to the best of our knowledge) about a way to estimate the offered load generated by users in a WLAN network consisting of a large number of APs.

Let us imagine that detailed statistical information about a WLAN network is known. However, how can we apply the statistical data to estimate traffic load on the network per day/week/month and so on? In particular, what initial data based on these statistics are needed to model traffic load parameters? How can we take into consideration the huge variability of traffic transferred to/from WLAN when traffic load parameters are estimated? How to reduce the number of computations concerning traffic load estimation?

Devising a methodology that takes into account these questions and is suitable to be used to model traffic load characteristics of WLAN networks is gaining importance as the number of large wireless networks is increased.

The main motivation is that on the basis of such methodology, a network planner should be able to calculate an upper bound of the traffic load in the network, the most typical load in the network, etc. Besides, he/she should have the possibility to estimate network workload parameters in the view of an increase/decrease of the number of users, a change in the amount of Access Points (APs), or a change in a demand for different types of information content. Such estimations may be useful in different aspects of WLAN network planning and deployment, e.g. to adjust with them performance measures of IP core network nodes.

The study presented in this paper is, to the best of our knowledge, a first effort in devising such methodology taking into account the above-mentioned aspects to model the offered traffic workload in large WLAN networks based on a low number of initial data known from existing statistical information adequately describing the behavior of the network.

The rest of the paper is organized as follows. Section II provides a description of the methodology to model traffic load parameters in general. Section III presents an approach to determine the unknown parameters for the methodology that exploits the unequal distribution of AP popularity throughout the wireless network. Section IV is devoted to practical validation of the methodology using real collection of WLAN traces of a popular SNMP data collection of Dartmouth College. Section V concludes the paper.

II. GENERAL DESCRIPTION OF METHODOLOGY

WLAN networks consist of a number of APs that spread all over the network and are interconnected through a common backbone (wired or wireless).

A common practice when gathering information from deployed WLAN networks is using periodic SNMP polls of the APs (e.g. [2]). The SNMP trace consists in a set of data records coming from the periodic polling (every 5 minutes) of each one of the APs deployed in the network. For each one of the SNMP polls of the trace, a data log showing an AP number and the amount of traffic that such AP has handled in a period of 5 minutes is obtained [2,8].

Generally, the traffic load transferred to/from a WLAN network for any observation time interval is calculated as:

$$U(t) = \sum_{l=1}^{M} S_l(t), \tag{1}$$

where M is the number of APs, $S_l(t)$ is the amount of data handled by the l-th AP for the time interval (t).

The amount of data $S_l(t)$ may be expressed as summation of bits of all logs that are generated by the l-th AP for the time interval (t):

$$U(t) = \sum_{l=1}^{M} \sum_{r=1}^{N} L_{lr}(t), \tag{2}$$

where N is the number of logs for the time interval (t), L_{lr} is the amount of bits in the r-th log handled by l-th AP.

Definitely, such traditional approach for the load calculation $U(t)$ is impracticable when an amount of APs (M) in the network and the time interval (t) under consideration are becoming very large since it deals with enormous amount of computations. Besides, such approach uses all collected data and does not select a representative short time period characterizing behavior of the network to estimate, for instance, the most typical load or the maximum load (depending on needs of an operator) in the network for a long term period.

We develop the methodology taking into account these lacks. It is based on using a low number of the initial input data to model the offered traffic load $U(t)$.

Let us assume that detailed statistical information about the behavior of the WLAN network is known.

Then, the decomposition of all SNMP logs (N) into some conditional levels (i = 1,2,…,n) based on an amount of handled data per a log is fulfilled in the methodology. As a rule for decomposition, each level includes just those logs that are characterized approximately equal values of amount of handled data. So, the logs with a low amount of information belong to the first level (i=1), the logs with a big amount of transferred data correspond to the top level (i=n), the rest logs belong to the intermediate level (or, to some intermediate levels). The decomposition of all logs into levels is illustrated in the left part of Fig. 1.

So, the second sum in the expression (2) is changed as follows:

$$U(t) = \sum_{l=1}^{M} \sum_{i=1}^{n} \lambda_{il}(t) w_i(t), \tag{3}$$

where $\lambda_{il}(t)$ is the rate of logs belonging to the i-th level (i = 1,2,…,n) generated by the l-th AP for the time interval *(t)*, $w_i(t)$ is the average amount of data in logs belonging to the i-th level.

After that, one can make the division of all APs into some groups (j = 1,2,…,m) in accordance with the following rule (see the right part of Fig. 1). The APs belonging to one group should be characterized by similar behavior in the sense of the certain level of traffic handled. In other words, the group j includes those APs that handle the most part of logs belonging to the level i (i = j) although sometimes they handle logs belonging to other levels. So, the first sum in the expression (2) is changed as well

$$U(t) = \sum_{j=1}^{m} \sum_{i=1}^{n} M_j w_i(t) \lambda_{ij}(t), \tag{4}$$

where M_j is the number of APs in the group j, $\lambda_{ij}(t)$ is the specific (per an AP) rate of logs belonging to the level i (i = 1,2,…,n) generated by APs from the group j (j = 1,2,…,m) for the time interval (t).

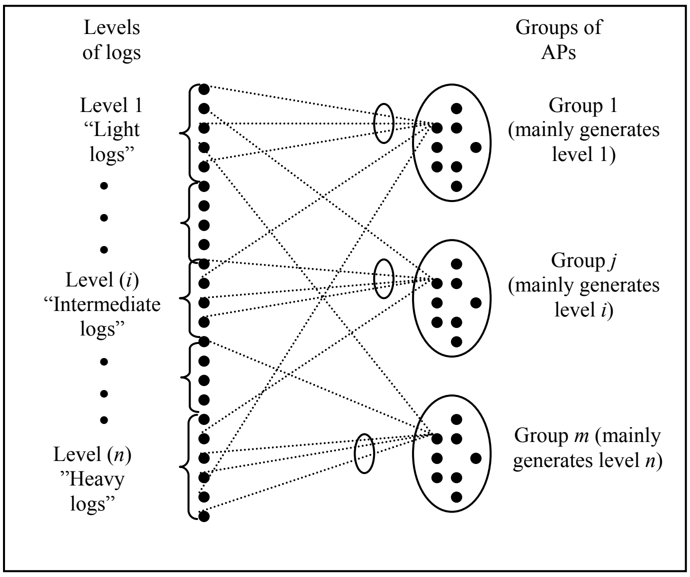

Figure 1. The common approach

So, normalizing the expression (4) by the number of APs in the network (M) we have

$$U(t) = M \sum_{j=1}^{m} \sum_{i=1}^{n} F_j w_i(t) \lambda_{ij}(t), \tag{5}$$

where F_j (j = 1,2,…,m, $F_1 + F_2 + \ldots + F_m = 1$) is the share (proportion) of APs in each group.

If there is statistical information describing adequately behavior of a network then it is possible to determine the most representative short time period (hour/day/weak), to estimate traffic load parameters in it, and to extend them for a long time period. For instance, the busiest hour (BH) as the most typical short time period may be used. In this case, the daily load (U) handled by a WLAN network may be estimated as

$$U = \frac{M}{k} \sum_{j=1}^{m} \sum_{i=1}^{n} F_j w_i \lambda_{ij}, \tag{6}$$

where k is the concentration factor showing the portion of daily transferred amount of traffic in the BH (e.g. the value k = 0.1 means that 10% of daily traffic were generated during the busiest hour). The parameters λ_{ij} and w_i in the expression (6) are considered in the BH as well.

Thus, the main steps of the methodology are to define on a basis of known statistical information the most representative short time period (e.g. the BH), to decompose both logs into levels (i = 1,2,…,n) and APs (j = 1,2,…,m) into groups, to calculate traffic load parameters in the BH within each of the defined mxn segments on a basis of the formed initial date, to find the total offered traffic load in the BH, and after that, to estimate the traffic load for a long term period, e.g. day/week/month and so on.

Such methodology *allows reducing* significantly a number of computations since the estimation of load parameters in a network is based on a low number of the initial data formed from statistical known information for the most representative short time period.

Besides, the methodology based on the decomposition both logs and APs into levels and groups respectively enables segregating mxn segments from the common set of traffic initiated in the BH. *The variance of traffic load parameter values within each of the mxn random segments is less than the one relating to the common data traffic that has a huge data scattering as mentioned above.*

In the expression (6) the parameters M, k, w_i are considered as the initial data that are known from statistical information about the network. The estimation of the common traffic load (U) then depends on the two parameters: the share of APs in each of the groups F_j, and the specific (per an AP) rate of logs in each of the segments λ_{ij} in the BH. So, it is needed to develop an approach to express the parameters F_j and λ_{ij}. Besides, a number of levels of logs (n) and a number of groups of APs (m) should be defined. The approach for estimating the above-mentioned parameters is considered in the next Section.

III. Approach to Estimate Unknown Parameters for Methodology

A. Estimating the share of Access Points (F_j) in each group

It was observed analyzing a SNMP trace [9] that the usage of APs highly varies at different points of the network. The same observations were made in other WLAN networks. The authors of [4] report variations from close to 0 up to 21 GB in the mean AP usage throughout their network. Even more, the measurement in [2] show that while the median APs handle an average of only 39 MB per day the busiest APs handle an average of over 2 GB per day. In fact, since there is the inequality of users associated with different APs from a poll to a poll and inequality of interests of users for different information content then, as a result, there is also inequality of distribution of APs throughout a network from the viewpoint of their usage (commonly called popularity of the AP).

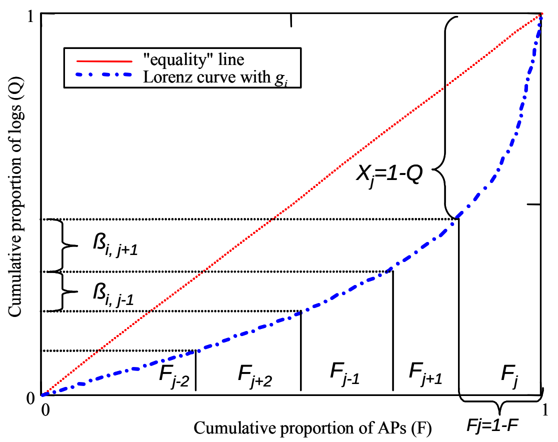

Figure 2. Lorenz curve using for the i-th level of logs

We exploit this specific feature of WLANs consisting of a large number of APs in the approach to determine such unknown parameters in (6) as the share of APs in each of groups F_j (j = 1,2,…,m, $F_1 + F_2 + \ldots + F_m = 1$) and the specific (per an AP) rate of logs λ_{ij} (i = 1,2,…,n; j = 1,2,…,m) in each of the segments mxn in the BH.

In order to find share of APs in each group F_j (j = 1,2,…,m) Lorenz curves illustrating inequality are built for each level of logs (i) as cumulative proportion of logs vs. cumulative proportion of APs on a basis of the collected statistical information as it is shown in Fig. 2. After that, values of the Gini coefficient characterizing deviation of a Lorenz curve from the equal distribution line are determined for each level of logs. Step-by-step procedures to build the Lorenz curve and to determine the Gini coefficient are given in [10] and [11] correspondingly.

A Lorenz curve may be presented in mathematical view in accordance with [12] as follows

$$Q(g) = 1 - (1 - F)^{\frac{1-g}{1+g}}, \tag{7}$$

where Q in our case is cumulative proportion of logs for a given proportion of APs, F is cumulative proportion of APs, g is the Gini coefficient.

In accordance with definition given in Section II for the group decomposition, the group j includes those APs that generate the most amount of logs (e.g. 60%) corresponding to the certain level i ($i = j$) of logs. This share of the amount of logs (X_j=1-Q) is known from statistical collected data. Then, the Lorenz curve is built for the corresponding level i to get Gini coefficient g_i. After that, the share of all APs (F_j) in the group j may be found modifying the expression (7) as

$$F_j = X_j^{\frac{1+g_i}{1-g_i}}, \quad i = 1,2,\ldots n, j = 1,2,\ldots m, j = i. \tag{8}$$

In such way all values of the relative number of users in each group F_j, j = 1,2,…,m are determined.

B. Estimating the rate of logs per an AP (λ_{ij}) in each traffic segment in the BH

The specific (per an AP) rate of logs λ_{ij} (i = 1,2,…,n; j = 1,2,…,m) in each of the traffic segments mxn in the BH may be defined taking into account that each AP handles 12 logs per hour (polling is every 5 minutes [2]) as

$$\lambda_{ij} = \frac{12\beta_{ij}\gamma_i}{F_j}, \tag{9}$$

where γ_i (i = 1,2,…,n; $\gamma_1+\gamma_2+\ldots+\gamma_n$=1) is the specific share (per each level of logs) of the total amount of logs in the BH, β_{ij} is the share (proportion) of logs of the j-th group of APs corresponding to the i-th level of logs in the BH.

Values of the parameter γ_i are considered as the initial data that may be formed on a basis of the collected statistical data. Values of the parameter β_{ij} may be found from the same Lorenz curves that were built earlier for determination values F_j (j = 1,2,…,m). In fact, for the Lorenz curve (i) proportions β_{ij} are located along the vertical axis Q. Groups of APs (F_j) corresponding to these proportions are located along the horizontal axis F. In accordance with observations that were made analyzing the SNMP trace collection [9], the horizontal axis F (see Fig. 2) records the groups of APs ranked by $F_1,F_2,\ldots,F_j$ when $j = m$, $i = n$, $j = i$ only. In case when $j \neq m$, $j = i$ the groups of APs are ranked along the axis F by $\ldots F_{j-2},F_{j+2},F_{j-1},F_{j+1},F_j$. For instance, if there are three levels (i=1,2,3) of logs, three groups (j=1,2,3), and the Lorenz curve is built for the third level (i=3) then the horizontal axis F records the groups of APs ranked by F_1,F_2,F_3 and the vertical axis Q records the proportions of logs ranked by $\beta_{31},\beta_{32},\beta_{33}$. If the Lorenz curve is considered for the second level of logs (i=2) then the horizontal axis F records the groups of APs ranked by F_1,F_3,F_2 and the vertical axis Q records the proportions of logs ranked by $\beta_{21},\beta_{23},\beta_{22}$ correspondingly, and so on. Applying the expression (7) each of parameters β_{ij} are determined.

Thus, using the approach on a basis inequality of popularity of APs, the share of APs in each of the groups (F_j), and the specific (per an AP) rate of logs (λ_{ij}) in each of the traffic segments in the BH may be determined.

However, there is one more important aspect that influences on the estimation of the traffic load (U) in (6). This is a decision of number of levels of logs (n) and a number of groups of APs (m). Recommendations about it are given in the next Section.

C. A choice of a number of levels and a number of groups

The decision about a number of levels (n) and a number of groups (m) is based initially on analyzing the representative statistical information about the behavior of a WLAN network and extracting the explicitly expressed levels of logs and groups of APs in accordance with rules given in Section II. As a criterion for the final decision, a notion of propagation of uncertainty (or propagation of error) may be used [13].

In fact, the final propagation error (ΔU) of the function (U) is the effect of the uncertainties of variables $N, k, F_j, w_i, \lambda_{ij}, \gamma_i, X_j$. The most part of these uncertainties ΔF_j, Δw_i, $\Delta\lambda_{ij}$, $\Delta\gamma_i$, ΔX_j depend on the selection of a number of levels (n) and a number of groups (m). When values n and m are changed some of the uncertainties are increased and, at the same time the other uncertainties are decreased. Thus, a number of levels and a number of groups should be selected as minimum in such way that the final uncertainty propagation (ΔU) cannot exceed the given threshold ε defined by a network manager:

$$(\Delta U)^2 \leq (\Delta U(N,k,F_j,w_i,\lambda_{ij},\gamma_i,X_j, \Delta N,\Delta k,\Delta F_j,\Delta w_i,\Delta\lambda_{ij},\Delta\gamma_i,\Delta X_j))^2 \leq \varepsilon^2 \quad (10)$$

The detailed approach how to determine uncertainties is presented in [13].

Thus, in the expression (6) all parameters (F_j, λ_{ij}, n, m) are defined and the daily (weekly, monthly,…) load (U) handled by a WLAN network may be estimated. Values of parameters k, N, w_i, X_j, γ_i are formed on a basis of the representative statistical information and are considered in the methodology as the initial data. Practical validation of the methodology is considered below.

IV. PRACTICAL VALIDATION OF METHODOLOGY

There is no (to the best of our knowledge) public available general statistics of traces for such type of networks. However, public repositories of traces resulting from periodic SNMP polls of WLAN networks are arising (see [8], [14]) that allow researchers to use real measurement data to test their research.

For practical validation of the methodology, a SNMP metadata collected at the Dartmouth College is used [9]. This is because of the fact that this is the largest WLAN network (more than 450 APs and several thousand users) with public available repository of traces. The short description of the SNMP trace is given below.

A. Analysis of SNMP trace

The data records [9] collected between November 24, 2003 and December 16, 2003 have been analyzed. Once obtained the set of data logs of the SNMP trace, we got the following characteristics. The number of APs identified (and numbered) in the whole trace is 557 though not all of them correctly respond to SNMP polls. Whenever a poll for an AP is missing, we consider that this AP handled no traffic. The whole trace contains total of 1.8 *million* data logs (i.e. polls where APs have shown some activity). The overall traffic load of the trace during the period is 4.4 TB and the mean traffic handled per an AP is 2.4 MB. Observing the whole trace of data it was found that all logs may be explicitly divided into three conditional levels: up to 10 kB, from 10 kB to 1MB, and above 1 MB. Besides, it was found that there are also three groups such that the group 1 generates the most logs of the level 1, the group 2 generates the most logs of the level 2, and the group 3 generates the most logs of the level 3. The histogram of the levels of logs for the whole trace of data is presented in Fig. 3.

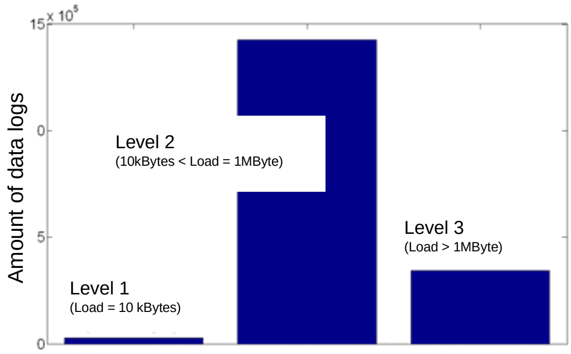

Figure 3. Histogram of levels of logs

B. Load estimation when initial data are formed on a daily basis

Firstly, the following procedure to validate the methodology is used. For a day, the initial data in the BH are formed using data set collected for the day. Then, the developed approach to get the traffic load for the day (U) in accordance with the expression (6) is applied using the formed initial data in the BH. The obtained estimation is compared with actual measured load for the same day obtained by using the expression (2). The procedure was fulfilled for all days of the whole trace records.

For this procedure we selected initially three levels (up to 10 kB, from 10 kB to 1MB, and above 1 MB) and three groups taking into account the above-mentioned analysis of the whole trace of data.

The results of the comparison show that the estimated load (U) coincides almost exactly (the average estimation error is about 10^{-7}) with the actual measured load for each of the days. It means that the methodology based on using such feature as inequality in AP popularity throughout the WLAN network is able to give very accurate values with the estimation error tends to zero. Note that we have not revised the number of levels and groups in accordance with (10) since initially it was chosen correctly and very exact results were obtained.

As an example of utilization of the methodology, let us imagine that a network manager/planer knows from long term observations the most representative day in the network, the busiest hour in this day, and the portion of daily traffic in the BH (k). Then, the network manager just needs to take a set of data logs collected in the BH, to form the initial input data on a basis of the set of logs in accordance with the methodology, to apply the methodology to get the load estimation for the day, to compare it with the actual load for the day, and if the estimation error is satisfied to calculate the typical traffic load in the network for a week/month/year, etc. (multiplying the result from (6) by 7/30/365 respectively). If the estimation error is not acceptable, a decision about the number of levels and groups should be revised in accordance with the technique described in Section III c.

Similarly, if the busiest day is known then the network manager using the methodology can estimate the upper bound of the traffic load that the network should handle for a week/month/year, etc.

C. Load estimation when the parameter k is formed by averaging over long time period

Note that the most representative day, the busiest hour in the day and the portion of daily traffic in the BH (k) are usually determined analyzing a set of the most typical days from long term observations of a network. However, what happens if for some reasons the BH for the most representative day and as a result, the portion of daily traffic in the BH (k) are not determined accurate by the network manager from long term statistical collection? In this section we check "robustness" of the proposed methodology using the average value of the parameter k obtained for some arbitrary selected long time period as input data.

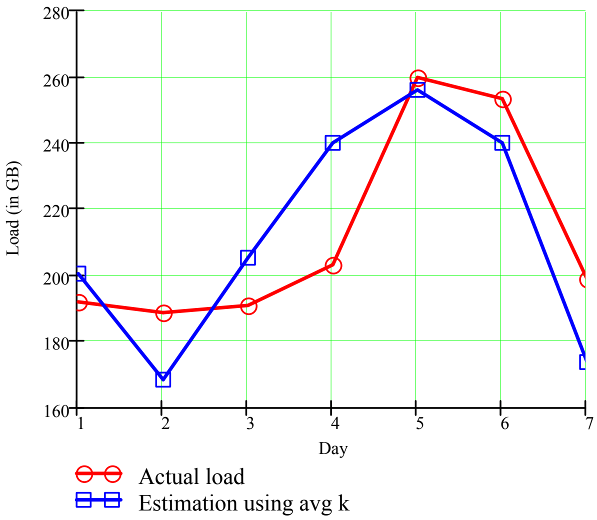

Figure 4. Real vs. estimation of traffic load

We take 7 arbitrary days (between December 03, 2003 and December 09, 2003) from whole period of data records and fix the hour from 17 to 18 as the busiest one for all these 7 days. Then, we get the average value of the concentration factor (k) showing a portion of daily traffic in the BH and use the value as initial data for the methodology to get estimations.

As it is seen from Fig. 4, the curve of the load estimation (U) when the average value k taken as initial data is slightly different from the curve of the actual load, although their behavior is the same. The plot of the load estimation error for this case is presented in Fig. 5.

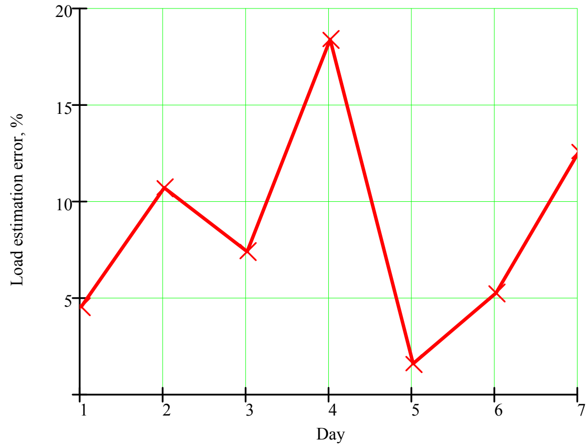

Figure 5. Estimation error of the methodology with av. value k as initial data

One can see that the load estimation error does not exceed 19%. The estimations may be considered as acceptable especially since arbitrary days were used to form the value of the parameters k for initial data instead of the most representative days that are used in such cases when statistical information adequately describing behavior of the network is available.

Thus, the methodology provides good estimations of the traffic load even if the busiest hour and the portion of daily traffic in the BH for the most representative day (or the busiest day) are not determined accurate from long term statistical collection.

V. Conclusion

In this paper the methodology for traffic load estimation in a WLAN network consisting of a large number of Access Points have been considered. It exploits the fact that usage of APs popularity distribution throughout the network is unequal. A popular SNMP metadata collection was used to validate the adequacy of the proposed methodology since there is no available statistic information generalized for such type of network. The obtained results show that the methodology gives a very exact approximation of the estimated traffic load to the actual values obtained from measurements. Additionally, the paper focuses on some practical issues related to utility and usage of the methodology for the offered traffic load estimation in WLAN networks. It is planned to continue verification of the methodology when general statistics of SNMP traces characterising large (hundreds of APs) WLAN networks are public available.

Acknowledgment

This work was partially supported by Generalitat de Catalunya under grant number SGR2005-00690 (Grup de Recerca Singular).

REFERENCES

[1] ITU-T Y.2001 (12/2004), General overview of NGN.

[2] D. Kotz, K. Essien, Analysis of a Campus-wide Wireless Network. In Proc. MOBICOM'02, Atlanta, USA, September 23–26, 2002.I. F.

[3] A. Balachandran, G. M. Voelker, Paramvir Bahl, P. V. Rangan. Characterizing User Behavior and Network performance in a Public Wireless LAN. In Proc. ACM SIGEMTRICS'02, June 2002.

[4] M. Balazinka, P. Castro, Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network. In Proc. MobiSys 2003.

[5] E. Hyytiä, J. Virtamo. On Load Balancing in a Dense Wireless Multihop Network. In Proc. of the 2nd Conference on Next Generation Internet Design and Engineering, 2006.

[6] E. Hyytiä, P. Lassila, A. Penttinen, J. Roznik. Traffic Load in a Dense Wireless Multuhop Network. In Proc. 2nd ACM International Workshop on PE-WASUN, Montreal, Canada, 2005.

[7] Felix Campos, Merkourios Karaliopoulos, Maria Papadopouli, Haipeng Shen. Spatio-Temporal Modeling of Traffic Workload in a Campus WLAN. Second annual international WIreless internet CONference (WICON06), Boston, USA, August 2-5, 2006

[8] Yeo, J., Kotz, D., and Henderson, T. 2006. CRAWDAD: a community resource for archiving wireless data at Dartmouth. SIGCOMM Comput. Commun. Rev. 36, 2 (Apr. 2006)

[9] http://crawdad.cs.dartmouth.edu/meta.php?name=dartmouth/campus

[10] G. Bellù, "Charting Income Inequality. The Lorenz Curve" University of Urbino, Institute of Economics, Urbino, Italy

[11] B. Slack and J.-P. Rodrigue, "Gini coefficient", located in http://people.hofstra.edu/geotrans/eng/ch4en/meth4en/ch4m1en.html

[12] L.E. Varakin. The Pareto law and the rule 20/80: the distribution of incomes and telecommunication services.,MAC proceedings, pp.3-10, no.1, 1997.

[13] http://en.wikipedia.org/wiki/Propagation_of_uncertainty

[14] http://netserver.ics.forth.gr/datatraces/