Framework of Applying a Non-Homogeneous Poisson Process to Model VoIP Traffic on Tandem Networks

IMAD AL AJARMEH
DePaul University
College of CDM
Chicago, USA
iajarmeh@cdm.depaul.edu

JAMES YU
DePaul University
College of CDM
Chicago, USA
jyu@cdm.depaul.edu

MOHAMED AMEZZIANE
DePaul University
Dept. of Mathematical Sciences
Chicago, USA
mamezzia@depaul.edu

Abstract: - This paper presents a new framework for Voice over IP (VoIP) traffic modeling based on a non-homogeneous Poisson process. The telecom industry is heading towards replacing the legacy TDM networks with an IP core network. The purpose of traffic engineering is to minimize call blocking probability and maximize resource utilization. A challenge of migrating to an IP core network is to develop an engineering model for VoIP traffic. We studied the call arrival process based on hundreds of millions of calls, and our analysis shows that the traditional traffic engineering approach based on the Poisson process fails to model the traffic behavior of modern telecommunication systems. We develop a new framework for modeling call arrivals as a non-homogeneous Poisson process, and propose a dynamic resource allocation procedure to maximize the bandwidth utilization for converged voice and data networks. The model is validated by real traffic data, and is also applied to predict the behavior of future data. We conducted statistical tests which demonstrate the validity of our model and the goodness-of-fit of predicted data and actual data

Key-Words: - Traffic engineering, VoIP, Erlang-B, Call arrival modeling, NHPP.

1. Introduction

The wide deployment of broadband, reliable, and cost-efficient IP networks is pushing towards a major paradigm shift in the telecommunication world. The wired as well as the wireless telecom industries are both heading towards an all-IP networks.

Traffic engineering is a fundamental requirement for designing and maintaining voice and data networks. It provides the tradeoff between service and cost. Traffic engineering of the Public Switched Telephone Network (PSTN) passed through many phases and matured enough to model the behavior of the legacy telecom process. IP networks are packet-switched rather than circuit-switched. The resources of IP networks are different from those of circuit-switched networks. For example the major resource in a circuit-switched network is the number of circuits (trunks), but the concept of trunks is not applicable to IP networks. The non-blocking nature of packet networks requires adding a Call Admission Control (CAC) component [1].

The introduction of IP-telephony and the wide spread of wireless technology have significantly affected the phone usage pattern, and change the traffic behavior. This effect results in inadequacy of the traditional traffic engineering approaches. The paper provides an in-depth analysis for call arrival patterns on an IP Tandem network. This study is based on hundreds of millions of calls. The data shows the deficiency in the traditional approach which is based on the Poisson process, and this deficiency causes significant underutilization of network resources. We propose a new approach to traffic engineering by applying a Non-Homogeneous Poisson Process (NHPP) for call arrivals. We then apply a generalized linear function to model call arrivals as a function of time. The proposed model supports dynamic allocation of network bandwidth based on predicted traffic, and modern network management system can easily support this dynamic bandwidth allocation procedure.

2. Call Arrival Process

The study of the call arrival process is to identify the key parameters to model the behavior of incoming call traffic into the system. The call arrival function, *p(k,t)*, is the probability of *k* calls arriving during the next *t* seconds

2.1 Traditional call arrival models

Traditionally, call arrival rate has been modeled using a *Poisson* distribution with a constant rate (λ)

$$p(k,t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad for\ k = 0,1,2,.. \qquad (1)$$

λ is the key parameter for the distribution. It indicates the average number of call arrivals in the given time interval (rate).

Under the Poisson assumption calls arrive independently from one another with a constant mean

arrival rate. Therefore, the observed call arrival process consists of the sum of a large number of independent call arrivals [2]. The Poisson assumption provides relative simplicity in the corresponding mathematical and analytical models. In practice, it is impossible to have a phone system with a constant call arrival rate over a long interval such as one day. Therefore, the model is designed for some maximum arrival rate, or the time of the day is divided into blocks and the arrival rate is assumed constant within each block. A separate queuing model is provided for each block.

Erlang-B model is used to estimate the telecom network resource requirements. This model is based on the Poisson arrival distribution where the rate (λ) is constant and is measured based on the Busy Season Busy Hour (BSBH) which is the busiest hour in the busiest week during the year. Networks are designed to handle traffic offered during this hour. Using a constant call arrival rate fails to adapt to the variation of traffic with respect to time, such as time of day, day of week, and day of year.

Under the BSBH approach, a considerable portion of network resources will remain idle for the majority of the year which results in poor resource utilization. Such problems can be justified in the PSTN world because of the difficulties associated with allocating and revoking network resources. For example, the typical limiting resource of a PSTN network is the number of trunks connecting central offices. Increasing or decreasing this number is a complicated and expensive process that involves the interaction of multiple parties. In the IP world resource allocation is flexible. Allocating more or less bandwidth for voice applications is a relatively simple process. Dynamic resource allocation for VoIP traffic can be useful especially for converged networks where voice and data share the same physical facilities. More bandwidth can be allocated to voice traffic during busy days, while providing non-used bandwidth for data applications during the reminder of the year.

2.2 Call arrivals as Non-Homogeneous Poisson Process (NHPP)

Recently, there has been a growing interest in using more flexible arrival models whose capabilities are not restricted by the traditional assumptions, for example: Irina et al [3] modeled the SS7 signaling traffic as exponential BCMP[1] queuing network, Brown et al [4] models the call arrivals at a call center as a time-inhomogeneous Poisson process with piecewise constant rates. This interest is driven by the need to solve the problems associated with the inadequacy of Poisson assumption to satisfy the modern engineering requirements. Finding explicit analytical equations for systems with complex arrival flows might be very difficult [5]. Research in this field tends towards simulations [6] or towards analyzing the system under the condition of heavy traffic (many calls in the system) [7] and low traffic (the system is mostly idle).

In the Poisson constant rate approach, the accuracy of the engineering process depends on the validity of the assumption that the arrival rate is constant within the given time blocks. In large systems with heavy traffic loads, one needs to have very small time blocks so that the rate can be assumed constant. Providing a separate queuing model for a large number of small time blocks is not a practical solution. A better approach is to model the call arrival process as a Non-Homogeneous Poisson Process (NHPP). The NHPP has independent call arrivals, the same as the Poisson Process; in addition it models non constant arrival rates [11]. The arrival rate is a function of time and can be captured using an appropriate time-dependent function. We adopt NHPP approach for modeling call arrivals of the IP tandem network, we go further steps in finding time-dependent function that models the variation of call arrivals.

3 Call Arrival Modeling Framework

This study is based on real VoIP traffic data obtained from one of the major IP-based Tandem carriers in the United States

3.1 IP tandem network: data collection

Tandem networks play critical roles in the telecommunications hierarchy. They interconnect different central offices together by means of toll switches. Central offices might belong to the same carrier or to different carriers, in the later case the tandem service provides interconnectivity and switching between different carriers (inter-carrier switching). As a result, tandem networks are expected to carry large amount of traffic and should be designed for high capacity, high availability, high scalability, and high cost-efficiency.

An IP-Based Tandem service utilizes IP core network instead of the legacy TDM as a transport for the voice traffic. Using a single converged IP network for data and voice provides substantial cost saving and enhanced network design and management. It is important to develop a traffic engineering scheme suitable for these

[1] BCMP network is a heterogeneous queuing network with multiple classes of customers having different distributions

converged networks. The goal is to increase resource utilization and hence decrease the cost and provide strong justification for migrating to VoIP networks. During this study we have collected several hundreds of millions of call detail records (CDR's) from an IP tandem network. We developed a simple yet useful library to collect raw data from the different sources and then to filter, aggregate, process and visualize the data according to our study needs. Fig. 1 illustrates a typical IP tandem network. The legacy PSTN is connected through TDM trunks. VoIP customers are connected via IP links. The network has an IP core which is used to interconnect different sites. The limiting resources on the network can be the IP backbone connections between different tandem offices, the IP connections to the VoIP customers, or the TDM connection to the legacy PSTN. The scope of our research is to optimize the first 2 IP-based resources.

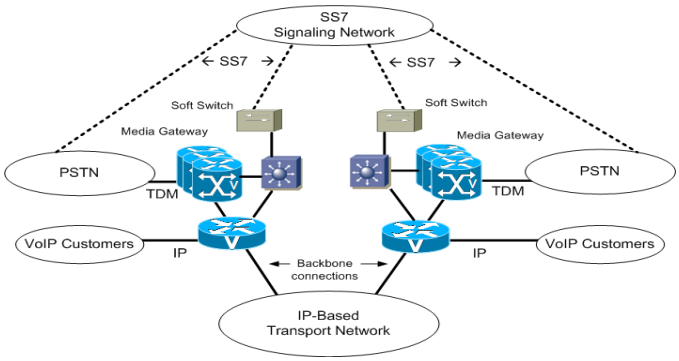

Fig. 1 Typical IP-based Tandem network

A Call Detail Record (CDR) is kept for every call on a local Billing Server located at each tandem office. Our scripts poll the distributed billing servers for CDR's every day. Once we have our copy of the CDR we divide the traffic into three categories: Wireless traffic, landline traffic, and VoIP traffic. Then we extract traffic information of our research interest. In order to satisfy the requirements of this research, we keep the raw time of arrival (TOA) for each call. We also generate aggregated forms of the data by dividing the day into time blocks and finding the mean of the call arrival rate over each time block. We generate 10, 100, 1200, 3600 seconds aggregated files.

3.2 Call arrival pattern

We study hundreds of millions of call arrivals collected over several months. We notice that the minimum call load occurs near 4 AM. Based on this finding we redefine the day from a traffic engineering point of view as the period between 4 AM and 4 AM of the next day. Furthermore, we notice that different days have different patterns. For example the difference between the call load on Fridays and that on Sundays is noticeable and should not be ignored. Fig. 2 shows call arrival patterns for a typical week. We notice call arrival difference of 70% between Friday and Sunday. Our proposed model takes the daily effect into consideration.

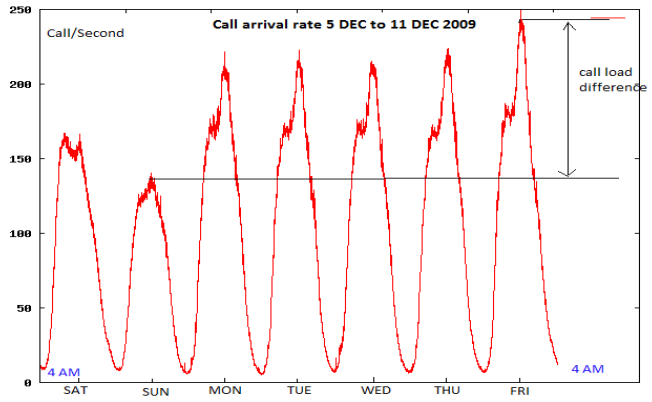

Fig. 2 Call arrival pattern for a typical week

3.3 Model formulation

After choosing NHPP to model the call arrival process, we need to find a time-dependent function that captures the variation of call arrivals. This step might vary from one system to the other depending on the nature of traffic (business, residential, enterprise, or mixed) and the scope and requirements of the engineering process. We show an example of using real data to derive a time function that takes the daily effect into consideration, we propose a model for the call arrival rate represented by a time-dependent intensity function $[\boldsymbol{\lambda(t)}]$

Given our data, we are inspired to construct a model that describes the variation of call arrival rates during a week. It is common in statistical analysis to model the logarithm of λ(t) instead of λ(t) itself for count data [8]. Such transformation would guarantee that the estimate of the intensity function is always non-negative. Our model takes into consideration the daily arrival patterns and has the time-dependent intensity function of:

$$\boldsymbol{log[\lambda(t)] = \mu + \sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t) + \beta_i \cos(i\omega_o t)] + \sum_{j=1}^{6} \gamma_j I_j(t)} \quad (2)$$

Where: $\boldsymbol{\lambda(t)}$ is a function of time (t).
$\boldsymbol{I_j(t)}$ is day Indicator function where $\boldsymbol{j}$ is the day of the week. The value of $\boldsymbol{I_j(t)}$ is 1 if the time $\boldsymbol{t \in j}$ and 0 otherwise. $\boldsymbol{K_o}$ is the number of harmonics in the model. $\boldsymbol{\mu}$ represents the model central tendency without daily

effects. γ_j is the effect of day *j* and represents the difference between μ and the mean number of calls for day *j*. α_i, and β_i are the contribution of the *ith* harmonic to the model.

Throughout the different days in a week, a similar pattern occurs regularly. In general, the rate of arrival during a given time of the day does not change significantly after a period of 24 hours, unless that day falls during the week-end or a special day. This pattern looks like a sinusoidal wave, and since the period is $T = 24 * 3600$ seconds, we consider the frequency $\omega_o = \frac{2\pi}{T}$. The model for $\lambda(t)$ belongs to a family of generalized linear models, known as Poisson regression models, which have been extensively studied and applied to analyze data from scientific sources [9].

3.4 Parameters estimation

We use Maximum likelihood estimation to fit our proposed λ(t) to the actual call arrivals. As explained in section 3.1, the processed call arrival data is aggregated into non-overlapping time intervals (δ) of 10, 100, 1200, and 3600 seconds. Thus we will use the total number of calls within time intervals rather than the exact call time of arrival. Let $n_1, n_2, .., n_{m-1}, n_m$ denote the number of calls arrived at the system in non-overlapping intervals $(a_1, a_2], (a_2, a_3], \ldots, (a_{m-2}, a_{m-1}] , (a_{m-1}, a_m]$
Therefore, the likelihood function L is given as

$$L = exp\left\{-\sum_{i=1}^{m}\int_{a_i}^{a_{i+1}} \lambda(t)dt\right\} \prod_{i=1}^{m} \frac{\left(\int_{a_i}^{a_{i+1}} \lambda(t)dt\right)^{n_i}}{n_i!} \qquad (3)$$

And the log-likelihood, apart from a given constant, is given as

$$l = \sum_{i=1}^{m} n_i log \int_{a_i}^{a_{i+1}} \lambda(t)dt \quad - \sum_{i=1}^{m}\int_{a_i}^{a_{i+1}} \lambda(t)dt \qquad (4)$$

Where: m is the number of intervals within each day.
Given that δ is the aggregation time interval, we can say that $a_{i+1} = a_i + \delta$ and $a_m - a_0 = m.\delta$
The value of δ is very small compared to the whole study duration. So practically, the integrals in (3) and (4) can be evaluated using the following approximation:

$$\int_{a_i}^{a_{i+1}} \lambda(t)dt \approx \delta\lambda(t_i)$$

Where $t_i = (a_{i+1} + a_i)/2$. The approximation error is of order $o(\delta^2)$. Hence Equation (4) becomes:

$$l = \sum_{i=1}^{m} n_i log[\delta\lambda(t_i)] \; - \sum_{i=1}^{m} \delta\lambda(t_i) \qquad (5)$$

Substituting the function $\lambda(t)$ given in (2) into the log-likelihood function, and excluding the constants that does not depend on the parameters, equation (5) becomes:

$$l = \sum_{k=1}^{m} n_k \left(\mu + \sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t_k) + \beta_i \cos(i\omega_o t_k)] + \sum_{j=1}^{6} \gamma_j I_j(t_k) \right) - \delta \sum_{k=1}^{m} exp\left(\mu + \sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t_k) + \beta_i \cos(i\omega_o t_k)] + \sum_{j=1}^{6} \gamma_j I_j(t_k) \right) \qquad (6)$$

Equation 6 can be rewritten as

$$l = n\mu + \sum_{i=1}^{k_o} [\alpha_i S_i + \beta_i C_i] + \sum_{j=1}^{6} \gamma_j F_j - \delta \sum_{k=1}^{m} G_k \qquad (7)$$

Where

$$S_i = \sum_{k=1}^{m} n_k sin(i\omega_o t_k), \qquad C_i = \sum_{k=1}^{m} n_k cos(i\omega_o t_k)$$

$$F_j = \sum_{k=1}^{m} n_k I_j(t_k) \qquad and$$

$$G_k = exp\left(\mu + \sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t_k) + \beta_i \cos(i\omega_o t_k)] + \sum_{j=1}^{6} \gamma_j I_j(t_k) \right) \qquad (8)$$

Notice that F_j is the total number of calls on day *j*. The terms S_i and C_i do not depend on the parameters while the terms G_k are exponential terms.

The ML estimators are obtained by taking the partial derivatives of the log-likelihood, with respect to the model parameters: $\mu, \alpha_i , \beta_i,$ and each of γ_j.
An implicit form of the solution to the first equation can be obtained easily as follows:

$$\hat{\mu} = log\left(\frac{\delta}{n} \sum_{k=1}^{m} G_k{}' \right) \qquad (9)$$

Where

$$G_k{}' = exp\left(\sum_{i=1}^{k_o} [\alpha_i \sin(i\omega_o t_k) + \beta_i \cos(i\omega_o t_k)] + \sum_{j=1}^{6} \gamma_j I_j(t_k)\right) \qquad (10)$$

The other score equations cannot be solved analytically; therefore we use Fisher scoring method to estimate the parameters as shown in the next section.

3.5 Inference about the model: significance, validation and prediction

To assess the performance of the estimators and hence of the model, we need to study the variability of these

estimators and test their significance. This task requires computing the covariance matrix of the estimators. ML estimation theory states that when the sample size is sufficiently large, as is the case of our call arrival data, the covariance matrix is equal to I^{-1} [10], where I is the information matrix, obtained by evaluating the negative expectation of the Hessian matrix of the log-likelihood function. We evaluated the variance terms for each parameter, then we used them to conduct Wald's significance test: H_o: $\theta = 0$ against H_1: $\theta \neq 0$, where θ is any parameter of interest (α_i, β_i, γ_j and μ). Table 1 shows the values of the estimated parameters, their standard errors and p-values of Wald's test.

Table 1 Estimated parameters for $\lambda(t)$

Parameter	Estimated value	Std. Error	p-value
μ	12.4851183	0.0002360	< 2e-16
α_1	0.6244975	0.0002387	< 2e-16
α_2	0.3730669	0.0002675	< 2e-16
α_3	0.1122494	0.0002224	< 2e-16
β_1	-1.2787258	0.0003443	< 2e-16
β_2	-0.4221888	0.0002767	< 2e-16
β_3	-0.1487193	0.0002205	< 2e-16
γ_1	-0.2266414	0.0003971	< 2e-16
γ_2	-0.5476155	0.0004534	< 2e-16
γ_3	0.0833744	0.0003486	< 2e-16

The small magnitude of the parameters' p-values suggests that the considered parameters are significant. Parameters with p-values larger than 0.05 were removed since their presence would be a nuisance to the model and might contribute to variance inflation. The significance of the model's parameters reflect the significant of the model itself and that it explains the variability in the data as can be seen in the plot in Fig. 3.

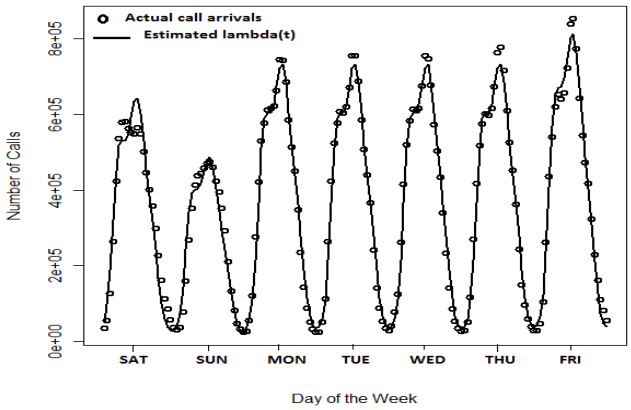

Fig. 3 fitting actual call arrivals to the suggested model

The model significance can also be evaluated by conducting the likelihood ratio test:

$$H_o: The\ process\ is\ HPP \Rightarrow \alpha_i = \beta_i = \gamma_j = 0$$
$$\Rightarrow (\lambda(t) = e^{\mu}\ (constant))$$
$$H_1: The\ process\ is\ a\ NHPP\ (\lambda(t) is\ time\ dependent)$$

If the null hypothesis is true, all the parameters in the model are non-significant and the process becomes a traditional HPP. If the null hypothesis is rejected then our model is significant and hence the call arrival process is generated by the NHPP with $\lambda(t)$ as described in equation (2). The likelihood ratio test statistic used here is evaluated as the ratio of the likelihood function for the restricted model (HPP under H_o), and of likelihood of the full model (NHPP with $\lambda(t)$) [8]. The null distribution of the test statistic is a chi square whose number of degrees of freedom equals the number of parameters minus 1. For our model and data, this value is equal to 34,676,131 with 8 degrees of freedom, corresponding to a p-value that is practically 0. Such very small p-value confirms our earlier results that the considered model is a very good fit to the data. Therefore, we can conclude that the observed call arrivals are definitely, not generated by the traditional homogeneous Poisson process and hence the need for modeling a time-dependent arrival is well justified.

In this example, the model is written in terms of 3 day indicators and 3 harmonics ($j=1,2,3$ and $K_o=3$), however in some cases we might consider including more indicators and harmonics and even relevant predictors (covariates) if necessary. The abundance of significant parameters in our model is due to the fact that the data set is very large, which results in many of the estimated terms being significant. Another way of looking at the proposed $\boldsymbol{\lambda(t)}$, is that it resembles a semi-parametric model in which the nonparametric component consists of a Fourier series expansion while the parametric component is a linear combination of indicator functions.

In general, if a model consists of more than a few parameters, one might argue against its usability. First of all, a model with many parameters is more flexible than a model with fewer parameters. Secondly, we developed this model based on real data and proved the high significance of its parameters. Using traditional models which are based on HPP involves unjustifiable approximations that lead to improperly engineered systems. The traditional traffic engineering process involves collecting sample call information and estimating a constant call arrival rate which is fed to a system based on Erlang-B model in order to calculate the required resources. In our approach we follow the same sample collection process, and then we feed the

collected sample (actual call information) to a system that will construct the model, estimate its parameters and compute the required resources accurately.

The importance of using a significant model lies in the capability of such model to predict future data. In this section we use our proposed framework to construct a model and estimate its parameters based on data collected in week 1 and then we use the model to predict data for other weeks. We compare the predicted data to the actual data that we already have for these weeks. Fig. 4 shows a plot of the predicted data against the actual data of two random weeks. The figure shows clearly that the actual observations fall very close to the curve of the estimated model.

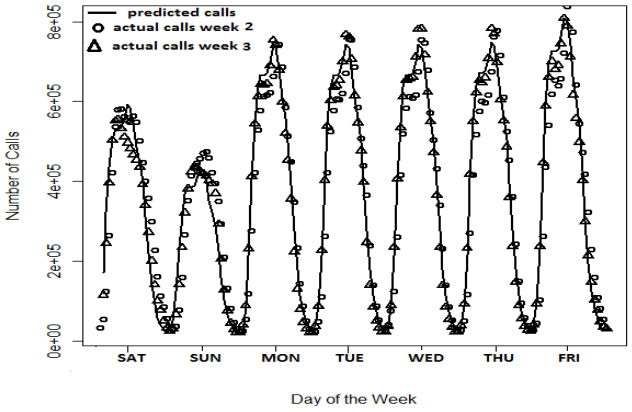

Fig. 4 Predicted against actual call arrivals for two random weeks

This approach of modeling call arrival rate allows us to build different time-dependent models based on the exact engineering requirements. For example, one might consider the variation of call arrivals from one week to another, or from one month to another or even from one hour to another and hence construct a different $\boldsymbol{\lambda(t)}$. Holidays, and special occasions can easily be dealt with by giving them indicator functions.

4. Conclusion

This research presents a new framework to model VoIP traffic based on real data collected from an IP tandem network. The data shows that the traditional Poisson process is not appropriate to model the VoIP traffic, while a non-homogeneous Poisson Process is able to capture the traffic behavior. A major contribution of this research is modeling the call arrival rate as a function of calendar time. We validate the model behavior with real traffic data over several months. The statistical analysis of predicted data and actual data shows strong model validity and goodness-of-fit. This traffic engineering model could support network management systems to develop a dynamic bandwidth allocation procedure. During the peak time of voice traffic, more bandwidth is allocated to the voice application. When voice traffic is low, more bandwidth is allocated for data services. Our next step in this research is to study the call holding time and find a model that captures its variability. Once we have models for call arrival rate and call holding time we will provide a queuing system to calculate the network resources

References

[1] James Yu and Imad Al-Ajarmeh, "Design and Traffic Engineering of VoIP for Enterprise and Carrier Networks," IARIA Journal, Volume-1, January 09

[2] P. Abry, P. Borgnat, F. Ricciato, A. Scherrer, and D. Veitch ,"*Revisiting an old friend: On the observability of the relation between Long Range Dependence and Heavy Tail*," Telecommunication Systems Journal, Springer Vol. 43, Numbers 3-4, April 2010

[3] I Buzyukova, Y Gaidamaka, G. Yanovsky , "*Estimation of GoS Parameters in Intelligent Network*", 9th International Conference, NEW2AN 2009, St. Petersburg, Russia, September 2009

[4] L Brown, N Gans, A Mandelbaum, A Sakov, H Shen, Sergey Zeltyn, and Linda Zhao, "*Statistical Analysis of a Telephone Call Center A Queueing-Science Perspective*", Journal of the American Statistical Association, March 2005, Vol. 100, No. 469, Applications and Case Studies

[5] K Kim, and H Choi, "*Mobility Model and Performance Analysis in Wireless Cellular Network with General Distribution and Multi-Cell Mode*," Wireless Personal Communications Journal, Springer Netherland, Volume 53, Number 2 / April, 2010

[6] J Bylina, B Bylina, A Zoła, T Skaraczyński, "*A Markovian Model of a Call Center with Time Varying Arrival Rate and Skill Based Routing*", Computer Networks: 16th Conference, Wisla, Poland, June 2009

[7] L. Afanas'eva and E. Bashtova, "*Limit theorems for queuing systems with doubly stochastic poisson arrivals (Heavy traffic conditions)*", Problems of Information Transmission, Vol. 44, No. 4. pp. 352-369, Dec 2008

[8] Dobson, Barnett, "*An Introduction to Generalized Linear Models*," Third Edition, Chapman & Hall/CRC Textx in statistical Science. 2008

[9] S. L. Rathbun, and S. Fei, "*A Spatial Zero-Inflated Poisson Regression Model for Oak Regeneration*," Environmental and Ecological Statistics Volume 13, Number 4 / December, 2006

[10] Deng X. and Yuan M, "*Large Gaussian covariance matrix estimation with Markov structures*," Journal of Computational and Graphical Statistics, 18(3), 640-657, (2009)

[11] H Wong, S Hsieh, Y Tu, "*Application of Non-Homogeneous Poisson Process Modeling to Containership Arrival Rate*," icicic, pp.849-854, 2009 Fourth International Conference on Innovative Computing, Information and Control, 2009