2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Estimation of Network Configuration based on the Genetic Programming and Delay Tomography

Yoshikazu Ikeda† , Shozo Tokinaga‡ and Jianjung Lu‡

†Faculty of Economics
Shinshu University
3-1-1 Asahi, Matsumoto-shi, Nagano 390-8621 Japan
‡Graduate School of Economics
Kyushu University
6-19-1 Hakozaki, Higashi-ku, Fukuoka 812-8581 Japan
Email: ikeda@econ.shinshu-u.ac.jp, tokinaga@en.kyushu-u.ac.jp

Abstract—The estimations of the performance parameters have been proposed using measurement made at a limited subset of computers is defined as the network delay tomography. Under the assumption the network topology and the routing table is fixed and the distribution of each link delay is independent, the problem is reduced to the parameter estimation using the likelihood method. However, we must also note that the network topology is not necessarily known or fixed throughout the observation. In this paper, we propose a method of network delay tomography for the network where the network topology is also not known and to be estimated as well as the density distribution of link delay. The representation of network topology is the same as the arithmetic expression, then we can assign the fitness for each individual by calculating the link delay estimation by using the pseudo likelihood method. The simulation studies are given for the artificially generated data.

1. Introduction

According to the fast growth of the Internet and related services, network monitoring and inference need to deal with a large number of network performance parameters such as packet delay on links. Since it is hard to directly measure network traffic characteristics by observing individual routers and servers, the estimation of the performance parameters can only be based on measurement made at a limited subset of computers.

Under the assumption the network topology and the routing table is fixed and the distribution of each link delay is independent, the problem is reduced to the parameter estimation using the likelihood method. Since the parameter estimation based on the full likelihood method is computationally infeasible and time consuming, several alternatives were proposed by modifying the full likelihood function to estimate parameters such as pseudo likelihood method.

However, we must also note that the network topology or the routing table is not necessarily fixed throughout the observation. Even more, if the network includes a closed form facilities such as the platform and the ASP (Application Service Provider), then the internal structure of the portion of network is not known to the users.

In this paper, we propose a method of network delay tomography for the network where the network topology and the routing table are also not known and to be estimated as well as the density distribution of link delay. The network delay topology is estimated by extending the Genetic Programming (GP) which has been successfully applied to the approximation of functions. By regarding the end-node as the terminal symbols and the arithmetic operation as the node of network connecting links, the same representation of the network as the arithmetic expression of GP is available.

Each individual is the representation of network topology, and then we can assign the fitness for each individual by calculating the link delay estimation by using the pseudo likelihood method proposed by Liang and Yu.

The proposed method of the paper is applied at first to the artificially generated data for a certain network structure. The result shows that after about 200 generations of the GP procedure, we have good estimation for the network topology and the delay distribution of link.

2. Estimation in Delay Tomography and Network Topology

2.1. Basic Model and Framework

We assume that in a general network topology a node represents a computer or a subnet (a connection of computers). A connection between any two nodes in the network is called a path, which may consist of several links. A packet is a unit of data of bits. Information is exchanged by sending packets along a path from a source nodes to destination nodes.[1][3]

Let $X = (X_1, X_2, ..., X_J)'$ be a J dimensional random vector, which reflects the network dynamics of interest, namely, the link delay or traffic flow counts at a particular time interval. Let $Y = (Y_1, Y_2, ..., Y_I)'$ be an I-dimensional

measurement vector. The goal of the network tomography is to estimate X from the observed Y.

$$Y = AX. \tag{1}$$

where A is $I \times J$ routing matrix.

Usually, the matrix A is determined by the network topology and the routing table at each router in the network, and is also restricted to be a fixed routing (without dynamic routing). Thus, the matrix is a fixed 0-1 matrix. However, in the paper, we assume that the matrix A is also estimated by using the GP procedure.

We assume that all components of X are independent each other, even though such an assumption does not hold strictly. But it give us a good approximation for the first step of analysis. Furthermore, we assume that the variable Xj obeys a density function f_j such as

$$X_j \sim f_j(\theta_j). \tag{2}$$

where is the parameter to be estimated. Then, the whole model includes the parameters $\theta = (\theta_1, \theta_2, ..., \theta_J)$.

Let $Y1, Y_2, ..., Y_T$ (whose elements are Y_{ti} be the observed data vectors at T consecutive time points, and $X_1, X_2, ..., X_I$ (whose elements are X_{tj}) be the corresponding unobserved network performance quantities of interest.

2.2. Subproblem with Pseudo Likelihood Estimation

Since the full likelihood method is still computationally infeasible or time consuming for most network tomography, a unified pseudo likelihood method is proposed. In the following, we follow the pseudo likelihood method by taking the product of subproblems with similar but simpler dependence structures.

Subproblems are formed by selecting some pair of rows from the routing matrix A. Namely, we select all possible pairs, but a subset can be judiciously chosen to reduce the computation.

Let S denote the set of subproblems by selecting all possible pairs of rows from the routing matrix A. Then, for each subproblem

$$Y_s = A_S X_S, \tag{3}$$

where X_S is the vector of network dynamics involved in the subproblem, and A_S is the corresponding routing matrix.

A discretizaton scheme is imposed on link-level delay in such a way that X_j takes finite possible values in $(0, q, 2q, \cdots, mq, \infty)$, where q is the bin width and m is a constant. Assume q is known so that each X_j is a independent multinominal variables with $\theta_j = (\theta_{j0}, .\theta_{j1}, ..., \theta_{jm}, \theta_{j\infty})$.

$$\theta_{jl} = Prob(X_j = lq). \tag{4}$$

If the delay is infinite, it implies that the packet is lost during the transmission.

Let n_{jl}^S be the number of packets whose length is lq. For a given subproblem s, each component of X_S is an independent multinominal random variable so that the log-likelihood function gives the complete data$X_1, X_2, ..., X_T$ is

$$l^s = \sum_{j \in J^S} \sum_l n_{jl}^S \log \theta_{jl}. \tag{5}$$

Where, J^S is the jth element of X is involved in s.Let $\theta^{(k)}$ be the parameter estimate obtained in the kth step of pseudo likelihood EM. Then, the Expectation Maximization algorithm is given as follows.

A uniform distribution for initial values of θ is used for all possible j and l, then, $\theta_{jl}^{(0)} = 1/(m+2)$.

(E-step) Calculate next value

$$\hat{n}_{jl} = \sum_{s \in S} E_{\theta^{s(k)}}(\sum_{t=1}^{T} 1\{X_{tj}^s = lq\} | Y_t^s). \tag{6}$$

(M-step) Update as follows

$$\theta_{jl}^{(k+1)} = \frac{\hat{n}_{jl}}{\sum_{r \in R} \hat{n}_{jr}}, R = [0, 1, ..., m, \infty]. \tag{7}$$

3. Estimation of network topology by the GP

In the basic method of the network tomography, we assume that the network topology is known and fixed. Then, we extend the model to the case where the network topology is not known as well as the delay distribution. In our method, the network topology is represented as a tree structure, and is regarded as an individual to be improved by the genetic operations.

If a network topology is compared with the tree structure of GP, a middle node in the tree structure can be considered as a node in the network(a network router or a hub), and a terminal node of the tree is considered as an end-node(a terminal computer).

Since each individuals(tree) in GP corresponds to a realization of one network topology, the network topology which an individual expresses can be changed by carrying out genetic operation of GP. And, more appropriate estimates of the network structure depending on the fitness of individuals are obtained.

Once the network topology is determined by interpreting the individual of GP, then the delay distribution on the end-nodes are calculated by the PLE. The fitness of an individual is defined as the inverse of estimation error for the delay distribution on the end-nodes.

The algorithm of estimation of the network topology and delay density functions by the network tomography and the GP is summarized as follows.

(Step 1)an initialization of individuals of GP

The population(pool) of first individuals is generated based on a random number. Generate an initial population

of random composition of possible nodes and terminals expressing the network. In this case, it is necessary to count each number of nodes of middle and terminal so that the tree structure which makes sense as topology may be obtained.

(Step 2) Calculation of the fitness of each individuals

Compute routing matrix A of each individuals in the population. And fitness value of individual are defined as the inverse of estimation error of joint density function of delay at the end-nodes of the estimated network and those of the observed value in receivers. Then, sort the individuals according to the fitness S_i.

(Step 3)Selection

Select a pair of individuals chosen with a probability p_i based on the fitness. The probability p_i is defined for i-th individual as follows.

$$p_i = (S_i - S_{min})/\sum^{N}(S_i - S_{min}). \tag{8}$$

where S_{min} is the minimum value of S_i, and N is the population size.

(Step 4)Crossover operation

Then, create new individuals (offsprings) from the selected pair by genetically recombining randomly chosen parts of two existing individuals using the crossover operation applied at a randomly chosen crossover point.

Iterate the step 3 and 4 several times to replace individuals with lower fitness.

(Step 5)Mutation operation

Apply the mutation operations at a certain probability. If the result designation is obtained by the GP (the maximum value of the fitness become larger than the prescribed value), then terminate the algorithm, otherwise go to Step 2.

Although the tree which does not express the network structure appropriately may be generated in initial individual generation, crossover operation and mutation operation. The problem is resolved by imposing some restriction in the generation of initial set of individuals and the offsprings generated in the crossover operations. Give the number of end-nodes N_e in the observation, then we select only initial set of individuals in which the number of nodes corresponding to the terminal nodes in the tree structure is the same as N_e. Moreover, at the crossover operation and the mutation operations to generate offsprings, it is claimed that the number of terminal nodes is the same as N_e.

3.1. The estimation of the artificially generated delay

As an application of the GP method to identify and estimate the delay tomography, we carry out model simulation on multicast tree depicted in Fig.1. Here, it is assumed that the network topology shown in Fig. 1 is unknown and delays at the terminal node are only observed.

Followings are used for the simulation study.

- Discretization parameters of a delay density function, described in section 2.2: $q = 1, m = 20$

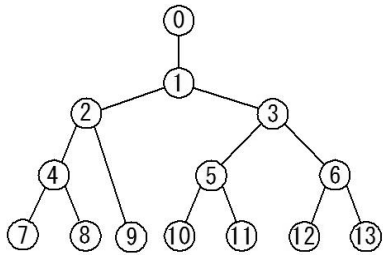

Figure 1: The network used for estimation of GP

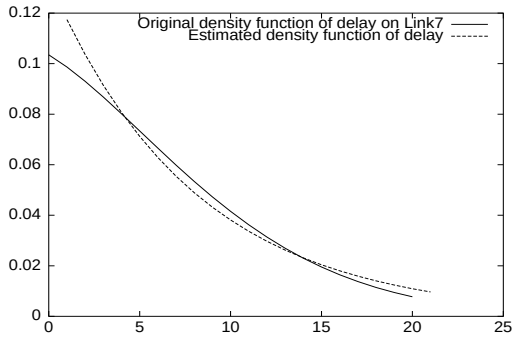

Figure 2: delay distribution on link7

- delay density functions of nodes: Exponential distribution (Averages are from 3 to 8)
- the number of multicast delay measurement:200

Since the topology of this network is simple, GP and EM method can estimate the same topology at 26th step of GP.

Figure 2 shows estimated density functions of delay compared with true density functions at link7. The x-axis of this figure shows the discretization parameter m. Although there is a divergence of original and the estimated density function where m is small, it can be said that the almost same density function as the original density function is estimated by this method. This is because the topology estimated by GP and the original network topology is the

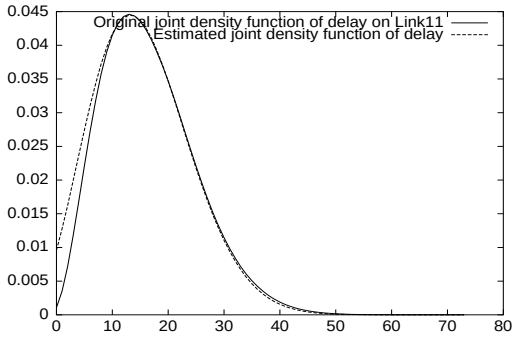

Figure 3: delay distribution on link11

same, if the estimated topology is different, the estimated error of density functions will become large.

Figure 3 shows estimated join density function of delay compared with true joint density functions at link11. The x-axis of this figure shows the discretization parameter m. If the estimated network topology differs from the original one, the passed number of links may be different and an estimated error becomes large.

The results show that MPLE with GP procedure provide us almost the same estimation for delay distribution. Furthermore, the network topology is also estimated and identified by the GP procedure given in the paper.

4. Conclusion

We proposed a method of network delay tomography for the network where the network topology and the routing table are also not known and to be estimated as well as the density distribution of link delay. Each individual in the GP was the representation of network topology, and then assigned the fitness by calculating the link delay estimation by using the pseudo likelihood method. The simulation studies were shown for the artificially generated data and real world data.

The problems remain to be solved are the extension for various real world data and estimation of density function in a functional form, and the further research will be done by the authors.

References

[1] Y.Vardi,"Network tomography : Estimating source-destination traffic intensities from link data", J. Amer. Statist. Assoc., vol.91, pp.365–377, 1996.

[2] M.Coates, A.ero, R.Nowak and B.Yu, "Internet tomography", IEEE Signal Processing Mag., vol.19,pp.47–65, 2002.

[3] G.Liang and B.Yu,"Maximum pseudo likelihood estimation in networks tomography", IEEE Trans.Signal Processing, vol.51,no.8, pp.2043–2053, 2003.

[4] Y.Tsang,M.Coates and R.D.Nowak, "Network delay tomography",IEEE Trans.Signal Processing, vol.51,no.8, pp.2125–2136, 2003.

[5] C.Ellis, et.al."Dynamic change within workflow systems", Proc. Conf. on Organizational Computing Systems (COOCS'95), pp.10–21,ACM, 1995.

[6] J.R.Koza, Genatic Programming,MIT Press, 1992.

[7] J.Koza,"Genetic programming:A paradigm for genetically breeding populations of computer programs to solve problems", Report No.STAN-CS-90-1314, Dept.of.Computer Science Stanford University, 1990.

[8] Y.Ikeda and S.Tokinaga,"Approximation of Chaotic Dynamics by using smaller number of data based upon the Genetic Programming and its Applications", Trans. IEICE, vol.E83-A.no.8, pp.1599–1607, 2000.

[9] Y.Ikeda and S.Tokinaga,"Controlling the chaotic dynamics by using approximated system equations obtained by the genetic programming", Trans. IEICE ,vol.E84-A,no.9,pp.2118–2127, 2001.

[10] Y. Ikeda,"Estimation of the chaotic ordinary differential equations by co-evolutionary genetic programming", Trans. IEICE, vol.J85-A,no.4, pp.424–433, 2002.

[11] X.Chen and S.Tokinaga,"Synthesis of multi-agent systems based on the co-evolutionary genetic Programming and its applications to the analysis of artificial markets", Trans. IEICE, vol.J86-A, vol.10, pp.1038–1048,2003.

[12] X Chen and S.Tokinaga,"Approximation of chaotic dynamics for input pricing at service facilities based on the GP and the control of chaos", Trans. IEICE, vol.E85-A, no.9, pp.2107–2117, 2002.