Journal of Network and Computer Applications 57 (2015) 12–20

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

A PCA based optimization approach for IP traffic matrix estimation

Erdun Zhao, Liansheng Tan *

School of Computer Science, Central China Normal University, Wuhan 430079, PR China

ARTICLE INFO

Article history:
Received 18 October 2014
Received in revised form
11 May 2015
Accepted 8 July 2015
Available online 22 July 2015

Keywords:
Traffic matrix
Mahalanobis distance
Moore–Penrose inverse
Principal component analysis (PCA)
Prior distribution
On-line estimation

ABSTRACT

Inferring traffic matrix (TM) from link measurements and routing information has important applications including capacity planning, traffic engineering and network reliability analysis. The challenge comes from that there are more unknowns than data. To face this challenge, this paper describes the inference problem as an optimization problem, where the objective is to minimize the Mahalanobis distance between the solution and a certain prior distribution, subject to the routing and link measurement constraints. This optimization problem is then solved by the Moore–Penrose inverse of the routing matrix. To reduce the computing complexity, a principal component analysis (PCA) approach is further applied in solving the optimization problem. We obtain the explicit formulas by using the Moore–Penrose inverse and the PCA theory. On the basis of the generalized inverse of routing matrix and the PCA theory, we propose an interesting generalized Tomogravity approach, which is subsequently termed as PCAOM. We present the complete mathematical solution and the algorithm of the described TM estimation problem. By introducing a weight parameter, a generalized algorithm is presented, which can be applied flexibly by adjusting the importance of the prior according to the accuracy of the prior or even no prior is required when the prior is unavailable. Numerical results are provided to demonstrate the accuracy of our method with the dataset of Abilene network through the comparison with the famous Tomogravity method. Given that we have proposed two algorithms for the optimization problem of TM estimation, we also provide a guideline on how to choose the proper algorithm according to the availability of the prior information.

1. Introduction

As an important branch of Internet engineering, traffic engineering plays an increasingly significant role in the next generation Internet architecture. Traffic matrix (TM), which records the traffic volume of flows between origin–destination (OD) Internet node pairs, is important for many Internet engineering tasks, such as traffic engineering, network planning and dimensioning, load balancing and fault diagnosis and so on (Roughan et al., 2003). Recently, TM has also been applied to anomaly detection (Lakhina et al., 2004a; Wang et al., 2012; Tian et al., 2014). However, it is very difficult, if it is impossible, to obtain an accurate TM directly by using the existing Internet hardware or software. One reason is the bursty nature of Internet traffic, and the other insurmountable difficulty is due to the large number of OD pairs.

Although it is difficult to record the OD pair traffic flow volumes directly, the number of links is usually relatively small, and subsequently the link traffic counts can be obtained without difficulty by using the simple network management protocol (SNMP) or the information in the router which connects to that link. With consideration to the above fact, many research efforts have focused on TM estimation by utilizing the information of measurements on link loads, using linear programming (LP) or statistical inference techniques (see, e.g. Medina et al., 2002 and the references therein).

Being organized into an n-dimension vector for convenience, let X denote the unknown TM. By denoting the m-dimension vector of link counts as Y and the $m \times n$ routing matrix as A, we formulate the linear equations $Y=AX$ to relate the three components X, A and Y. Because the number of OD pairs is almost always much larger than the number of the links, to find an unique solution from the equations $Y=AX$ is obviously an ill-posed problem. For this reason, most methods on TM estimation use the TMs prior distribution or a given theoretical distribution as the extra information.

Recent attempts to the above TM estimation problem include the following. Goldschmidt (2000) modeled the TM estimation problem as a linear programming (LP) optimization problem with the assumption that the link counts Y are accurate. The aim therein is to find a

* Corresponding author.
E-mail addresses: erdunz@mail.ccnu.edu.cn (E. Zhao), L.Tan@mail.ccnu.edu.cn (L. Tan).

http://dx.doi.org/10.1016/j.jnca.2015.07.006
1084-8045/

vector as the TM estimation from the feasible solution space of equations $Y=AX$ which has the maximum sum of all elements. Cao et al. (2000) utilize a time-varying statistical approach to estimate TM by using link counts at router interfaces. A statistical algorithm called the expectation-maximization (EM) method is applied by Medina et al. (2002) to compute TM under the assumption that X obeys Gaussian distribution. Under the assumption that X is Poisson distributed, Vardi (1996) infers the TMs by a moment estimator which used the estimation of the first moment and the second moment (covariance matrix) of link counts. The Bayesian method by Tebaldi and West (1998) models the TM X as a Poisson distribution and then X is estimated by the conditional expectation $E(X/Y)$. With the OD flows independence assumption, a method called the gravity model by Roughan et al. (2002) was applied to estimate TMs in IP network, in which the traffic volume of OD pairs nodes (origin or destination) is proportional to the total traffic volume of that node. Through the analysis of various objective functions in the references, a new objective function with information entropy has been proposed by Zhang et al. (2003a, 2005), which is also based on the similar OD independence assumption and is a generalized version of the gravity models by Roughan et al. (2002). By resorting to the special structure of the route matrix in an IP-based virtual private networks (VPN), Shioda and Ohtani (2006) proposed a computation-efficient TM estimator without computing matrix inverse. It should be pointed out that the efficiency comes from the simple structures of routing matrix and the simple assumption of TM covariance matrix, which unfortunately limits its application scope. Some TM estimation algorithms for new applications have also been explored recently (Luigi Conti et al., 2010; Jiang et al., 2011b; Hua et al., 2015; Nie et al., 2015). By considering the long-range dependent nature of Internet traffic, Luigi Conti et al. (2010) have explored an expectation-maximization (EM) based algorithm to estimate the TM and its Hurst parameter which is related to the long-range dependent model. Jiang et al. (2011b) divide traffic matrix into tendency terms and fluctuation terms, and explored a time–frequency model to characterize the traffic matrix time–frequency nature. By analyzing the two terms carefully and using different algorithms respectively, a comprehensive algorithm has been designed. This method is originated from analysis of the structure of the TM process, and can deal with more complex situation. Apart from the methods which take the IP OD flows as study targets, the authors of the most recent references (Hua et al., 2015) and Nie et al. (2015) applied the TM estimation techniques in data center networks (DCNs) and IP-over-WDM backbone network respectively. Both references have designed new algorithms to adapt their special environments.

Most of the afore-mentioned methods are based on the assumption that the TM obeys a certain distribution, which unfortunately limits their application scopes. To overcome this drawback, recent progress has been achieved by Zhang et al. (2003b) and Tan and Wang (2007a,b), with new objective functions and subsequently new models in nonlinear optimal problem for TM estimation being proposed, which finds the desired solution (a vector) from the feasible space. In particular, the Tomogravity method proposed by Zhang et al. (2003b) is the early efficient approach, in which the prior is set to be the output of gravity method and TM is solved using a weighted quadratic programming method. The method gained the efficiency for its simplicity. But its accuracy is heavily dependent on the accuracy of the prior information and the weights used in the optimization. A novel method based on Lagrangian multiplier is proposed by Tan and Wang (2007a) and an algorithm based on a kind of generalized inverse of matrix, namely {1}-inverse is given by Tan and Wang (2007b). Furthermore, by discussing the advantages and disadvantages of the Fanout model deeply, we recently presented an improved Fanout estimator called Tomofanout by Tan and Zhou (2015), which can grasp more information than the original Fanout model and then results in a more accurate TM estimation performance.

Recently, a Mahalanobis distance based regressive inference (MDRI) method is used to investigate TM estimation problem by Jiang et al. (2010). The method includes two steps. It firstly resorts to the autoregressive moving average (ARMA) model to compute an initial TM X_0. Secondly, an optimal model based on the Mahalanobis distance is established, which aims to minimize an objective function:

$$(Y-AX)^T V_Y^{-1}(Y-AX)+\lambda(X-X_0)^T V_X^{-1}(X-X_0)$$

subject to the constraint $Y-AX=0$, where V_Y^{-1}, V_X^{-1} are the covariance matrices of Y and X respectively, λ is a weight parameter. The model was then solved by an iterative algorithm. By considering the variance matrices of Y and X simultaneously, the paper has obtained a good TM estimator. However, the iterative algorithm has to deal with the high-dimension vector X at every step and so made the computation complexity unexpectedly. Another problem is that the constraint $Y-AX=0$ would make the first part $(Y-AX)^T V_Y^{-1}(Y-AX)$ of the objective function useless. The principal component analysis (PCA) can reduce the dimensionality of possibly correlated variables by using orthogonal transformation into a set of values of linearly uncorrelated variables called principal components. The PCA method can be used to TM estimation if the elements in TM are correlated. Lakhina et al. (2004b) applied the PCA method to reduce the dimensionality of the set of all OD flows into a set of low dimension flow: eigenflows. By this way, a new approach for traffic matrix estimation is developed by Soule et al. (2005), by only estimating the k most important eigenflows. When $k \ll N$, the problem of estimating the eigenflows from link traffic is a well-posed problem and the OD flow can be recovered from these estimated eigenflows. However it is argued that the PCA-based method can obtain good performance but behaves unstable since the top k principal components that best describe the underlying dimension can change with time. More importantly, it would lose too much information to make the problem well posed.

In this paper, we first describe the inference problem as two optimization problems, where the objective is either to minimize the Euclidean distance or the Mahalanobis distance between the solution and a given prior distribution, subject to the routing and link measurement constraints. Both optimization problems are then solved by applying the Moore–Penrose inverse. Then, we formulate the problem to be a PCA based one which gives rise to a method termed PCA based optimization method (PCAOM). By the method, we reduce the problem's dimensions and obtain the PCA based solution explicitly, on which two algorithms are subsequently developed. At last, we evaluate our algorithms by the Abilene dataset (Zhang). Being different from the method by Jiang et al. (2010), our methods have the capacity to combine the PCA into the optimal model in order to reduce the problem complexity by working with the explicit formula of the solution. Secondly, our method does not need an iterative algorithm which may result unpredicted computation complexity. Moreover, we do not need special prior estimation procedure or even no need for prior, though accurate prior can be helpful. Compared to the PCA method by Soule et al. (2005), our method can benefit from the prior information to gain a more accurate and computation-efficient solution.

The rest of the paper is organized as follows. In Section 2, we model the TM estimation problem as two optimization problems, present the solutions for the optimization problems, and discuss their computational complexity. In Section 3, we firstly give a PCA based model with its solution, and then provide an algorithm to calculate the TM. We then extend the algorithm to a generalized case, where the weight on the prior is considered. In Section 4, we compare the PCAOM approach with the state-of-the-art method