Traffic uncertainty models in network planning

Yang, S; Kuipers, Fernando

DOI
10.1109/MCOM.2014.6736759

Publication date
2014

Document Version
Accepted author manuscript

Published in
IEEE Communications Magazine

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable). Please check the document version above.
Traffic Uncertainty Models in Network Planning

Song Yang, Student Member, IEEE and Fernando A. Kuipers, Senior Member, IEEE

Abstract—In network planning one is architecting the network and its routing policies based on a “guesstimate” of the amount of traffic that might be routed through the network in the future. Clearly, not being able to foretell the future means that one of the biggest challenges in network planning is how to deal with traffic uncertainty. If not solved appropriately, the network operator has to resort to overprovisioning, which is costly. On a smaller time scale, traffic that is being routed through the network is generally stochastic in nature, which is also reflected in dynamically varying network parameters like available bandwidth and delay of links. In this article, we provide an overview of several traffic uncertainty models useful for network planning. Our focus is on (1) off-line approaches for network planning and (2) on-line routing in stochastic networks. We conclude with some possible directions for future work.

Index Terms—Network planning, robust optimization, stochastic networks, traffic uncertainty.

I. INTRODUCTION

With many new communications applications emerging, the growth of content sharing, and the wide-spread use of cloud computing, network planning is no sinecure. On the contrary, the aforementioned trends are merely a few examples to indicate that predicting and planning for future traffic scenarios is a daunting task. Consequently, one of the main challenges in network planning is how to deal with demand and traffic uncertainty. In this article, we consider network planning as the collective of

1. designing and reconfiguring a (virtual) network based on a long-term traffic scenario (we refer to this as off-line planning), and
2. the choice of what kind of algorithms are going to be used for dynamically allocating paths for incoming connection requests (we refer to this as on-line planning).

In the past, network providers have often resorted to overprovisioning (and under-utilization) to cope with traffic uncertainty. Needless to say, overprovisioning is costly, and any mechanism that can lead to higher utilization without sacrificing (too much) the reliability and quality of service will result in higher revenues. Currently, one of the approaches toward better network utilization is introducing flexibility into the network, e.g. by more fine-grained optical spectrum usage, or by dynamic software-defined networking (SDN). However, regardless of the level of network flexibility, the network and its traffic are still likely to behave in a stochastic manner. For example, especially in large networks, it is difficult to obtain an accurate view on the link characteristics like bandwidth utilization or latency, because their dynamics are usually of the same order as the time it would take to distribute information on the link state throughout the network. We believe that adopting stochastic models can aid in network planning for higher traffic utilization. Clearly, if we do not know anything at all about future traffic scenarios, then planning is most challenging, while if the future would be known, a deterministic tailor-made solution would be most profitable. In general, between the aforementioned two extremes, we do have some ideas on the order of magnitude of future traffic and expect only some fluctuation around it, which falls in the category of stochastic traffic models.

We will use the following notation throughout the article. Let $G(N,L)$ denote a network, where $N$ represents a set of $N$ nodes and $L$ denotes a set of $L$ links. A Traffic Matrix (TM) is represented by $T$, and reflects the traffic demand between each possible source-destination pair. For any node pair $(a,b)$ and link $(i,j) \in L$, let $f_{ab}(i,j)$ denote the fraction of flow between $a$ and $b$ that traverses link $(i,j)$. $t_{ab}$ represents the total demand from $a$ to $b$ for a certain traffic matrix $T$, and $C_{ij}$ stands for the capacity of link $(i,j)$.

In the remainder of this article, we provide a brief overview of planning for networks subject to traffic uncertainty. We first highlight, in Section II, some papers on modeling traffic that are used to decide how much capacity should be reserved per link. The goal here is to use the least amount of resources to accommodate all traffic at an acceptable reliability level. The second part (Section III) relates to more dynamic traffic requests that should be routed over a network where the link weights are reflected in some stochastic model. We will conclude in Section IV by pointing out some challenges for future work in network planning. Fig. 1 schematically presents the traffic uncertainty models discussed in this article.

II. OFF-LINE APPROACHES TO TRAFFIC UNCERTAINTY

Contrary to the often used deterministic traffic model, where the volume of traffic between each source-destination pair is “known,” traffic uncertainty refers to traffic volumes that are not or only partially known in advance. The volume of traffic could be a random value that falls within a certain range or it could follow a known probability distribution. Based on a certain traffic model, a corresponding planning algorithm is suggested. Table 1 provides an overview of off-line approaches to dealing with traffic uncertainty.

### A. Oblivious Routing

Oblivious routing refers to a routing scheme that does not use any information on the current allocated traffic in the network. Routing decisions are made solely based on knowledge of the source and destination nodes of the arriving requests. In this sense, routing traffic along shortest-hop paths could be considered to be an oblivious routing scheme. However, in routing schemes bearing the name oblivious, usually there is guaranteed oblivious ratio $\rho$ that states that the performance in some metric (like utilization) of the routing scheme is never more than a factor $\rho$ away from a utopian exact algorithm that would know the entire traffic matrix in advance.

Applegate and Cohen [1] base their oblivious ratio on the maximum link utilization, where the link utilization of a link for a TM and routing $f$ is defined as the total flow of traffic on that link divided by its capacity. The oblivious ratio reflects the worst-case ratio over all possible traffic matrices between the maximum link utilization of the oblivious routing and the maximum link utilization of an exact algorithm (for a TM). Applegate and Cohen provide a Linear Program (LP) of polynomial size that is able to compute a routing $f$ that minimizes the oblivious ratio for a given network. More specifically, the LP returns $f_{ab}(i,j)$ that gives for a node pair $(a,b)$ the fraction of flow that should traverse link $(i,j)$. Applegate and Cohen also propose an LP for when the traffic matrix is restricted to containing values in the range $[t_{ab}/w, t_{ab}w]$, where $t_{ab}$ can be interpreted as the mean traffic between node pair $(a,b)$ and $w\geq 1$ is an error margin. Clearly, the smaller $w$ is, the smaller the oblivious ratio will be.

<table>
<thead>
<tr>
<th>Traffic Model</th>
<th>Objective</th>
<th>Hardness</th>
<th>Solution</th>
<th>Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range of $[t_{ab}/w, t_{ab}w]$ [1]</td>
<td>Minimize maximum link utilization</td>
<td>Polynomially solvable</td>
<td>LP</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Hose Model [2]</td>
<td>Maximize the throughput</td>
<td>Polynomially solvable</td>
<td>LP</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Polytope [4]</td>
<td>Minimize the cost</td>
<td>Polynomially solvable</td>
<td>LP</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Polyhedron [5]</td>
<td>Minimize the cost</td>
<td>NP-hard</td>
<td>ILP</td>
<td>Exponential time</td>
</tr>
<tr>
<td>Random Variable [6]</td>
<td>Maximize the revenue</td>
<td>Polynomially solvable</td>
<td>Convex programming</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Truncated Gaussian Distribution [7]</td>
<td>Minimize the maximum link capacity</td>
<td>Polynomially solvable</td>
<td>Convex programming</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Multivariate Gaussian [8]</td>
<td>Minimize the cost</td>
<td>NP-hard</td>
<td>MILP</td>
<td>Exponential time</td>
</tr>
</tbody>
</table>

### B. Hose Traffic

As discussed already for the work of Applegate and Cohen [1], the more we know about the range in which traffic matrices could appear, the better routing strategies could be devised. We will briefly introduce such a constrained traffic matrix model, called the hose traffic model.

Contrary to knowing the complete traffic matrix (“pipe model”) where customers – for instance in the context of VPNs – specify for each destination the bandwidth to be reserved, in the hose model every node (customer) $m \in \mathcal{N}$ specifies two threshold values $u_m \geq 0$ and $v_m \geq 0$, for the maximum traffic emanating, respectively terminating at it. The bounds correspond to bounds on the row and column sums of possible TMs, where there is a natural bound based on the link capacities.

Kodialam et al. (e.g., see [2]) have proposed a two-phase oblivious routing scheme for traffic obeying the hose model, where in the first phase fractions of the traffic from a node are sent to intermediate nodes, while in the second phase those intermediate nodes forward the traffic to the intended destinations. The fractions are predetermined via a polynomial-sized LP.

Frêchet et al. [3] introduce a capped hose model, which is a hybrid form of the pipe model where peak demands between node pairs are constrained and the hose model where nodal ingress/egress traffic is constrained. As such, the capped hose model can span the entire range from deterministic point-to-point demands to the space of uncapped hose traffic. For relatively large hose node constraints, the pipe model is dominant and routing can be done based on the traffic matrix obtained from the point-to-point bounds. Vice versa, relatively large point-to-point bounds, means that a hose approach – like the single-hub two-phase approach from Kodialam et al. [2] – could be used. Frêchet et al. [3] consider the case when neither the pipe model nor the hose model clearly dominates and for that case propose a hierarchical multi-hub routing approach.

### C. Robust Optimization

Ben-Ameur and Kerivin [4] assume that traffic matrices are confined to a given polytope $\mathcal{T} = \{t \in \mathbb{R}^{N(N-1)}_+ : At \leq b\}$,
where \( \tilde{t} = (t_{ij})_{N \times N} \) is a vector with the traffic demands (instead of a representation via a traffic matrix), \( A \) is a real-valued matrix and \( b \) is a real-valued vector. For instance, Fig. 2 plots the traffic polytope corresponding to the two traffic demands \( t_1 \) and \( t_2 \) with \( A = \begin{bmatrix} -7 & -1 \\ 1 & 1 \\ -1 & -3 \end{bmatrix} \) and \( b = \begin{bmatrix} -7 \\ 5 \\ -6 \end{bmatrix} \).

![Fig. 2 Traffic polytope.](image)

The authors demonstrate that the polyhedral traffic model includes the hose model as a special case. The authors further develop a polynomial-size LP for finding a minimum-cost multi-path routing for traffic captured by a given polytope.

The approach taken by Ben-Ameur and Kerivin [4] is considered to be part of the field of robust optimization, where robust refers to robustness against a certain degree of traffic (or any other relevant problem parameter) uncertainty. Koster et al. [5] consider robust network design under the condition that the traffic matrix is uncertain but confined by a polyhedron. Their article also provides many relevant references to the field of robust optimization. Koster et al. [5] study a robust network design problem in which the objective is to find a minimum-cost routing, such that no link capacities are exceeded and where the used capacity (i.e., flow) must be integral. Even in the absence of traffic uncertainty, such problems are known to be NP-hard (i.e., computationally difficult to solve). Koster et al. [5] propose an Integer Linear Programming (ILP) approach to solve the problem.

**D. Probability Distribution**

Another approach to model uncertainty in traffic demand is assuming that the traffic follows a certain probability distribution. Mitra and Wang [6] consider the case where traffic between different node pairs is reflected in a random variable with known probability distribution. The objective function in the proposed model is to maximize \( \mu(W_T) - \delta \sigma(W_T) \), where \( \mu(W_T) \) and \( \sigma(W_T) \) represent the mean and the standard deviation of the revenue \( W_T \) for TM \( T \), and \( \delta \) is a non-negative control parameter reflecting the tolerance to risk due to uncertainty in traffic forecasts. This maximization objective leads to a concave maximization problem when \( k_\alpha \gg \delta \), where \( 1/k_\alpha \) is the ratio of the maximum variance of the revenue from node pair \( v \) to the minimum variance of the revenue from all other node pairs. Johnston et al. [7] assume traffic demands follow a truncated Gaussian distribution and aim to design a logical topology to accommodate the demands on top of an existing network. They define a non-linear formulation, its linearization and an iterative scheme to minimize the maximum link capacity needed for that logical topology, subject to the link overflow probability not exceeding a given threshold. Aparicio-Pardo et al. [8] consider a similar problem as in Johnston et al. [7] where they assume a multivariate Gaussian traffic distribution together with a correlation (between the demands) matrix. Under their multivariate model, the region containing a certain desired percentage of all possible matrices is defined by an ellipsoidal region. The authors develop a Mixed Integer Linear Program (MILP) to minimize the cost in accommodating any traffic matrix captured in the ellipsoid region.

A general optimization framework for dealing with uncertainty is that of stochastic programming, e.g. see [9]. This branch of optimization theory is similar to robust optimization, but instead of having (only) constrained traffic/parameters, it is assumed that the probability distributions of the relevant parameters are known or can be estimated based on historical data.

**E. Discussion of off-line approaches**

The more we can reduce the uncertainty in future traffic demands, the better we can tailor our network and algorithms to accommodate those demands. The network architect has a choice to adopt an agnostic approach that makes decisions completely oblivious to the current or predicted traffic demands or to use a fixed predicted traffic matrix on which network planning is based. Both cases display a level of uncertainty, either reflected by the oblivious ratio in oblivious routing approaches or in the uncertainty in predicting a future traffic matrix. There is insufficient data on which of these approaches work better in practice (although there is some work, e.g. by Kronberger et al. [9], on the impact of uncertainty on a pre-planned result), but we believe that planning based on constrained stochastic traffic matrices (possibly via robust optimization [5] or stochastic programming [9]) serves as a good compromise.

### III. On-line Approaches to Traffic Uncertainty

In this section, we focus on on-line routing algorithms for stochastic networks. Table 2 provides an overview of the stochastic link models, corresponding routing problems and associated routing algorithms.

#### A. Discrete Markovian Delay

In the discrete Markov delay model [10], time is divided into discrete unit slots and per slot the delay is assumed to be fixed. Each link’s delay is modeled by a discrete Markov chain with finite states, known transition probability matrix, and non-negative delay values. Only upon arriving at a node \( a \), one can acquire knowledge of the state and transition probability matrix plus corresponding delay values of \( a \’ s \) adjacent links.
Waiting, for the duration of a number of time slots, at an intermediate node is allowed. For example, in Fig. 3(a) we assume that link (1,2) can be in (two) different states, while the other links permanently reside in one state (for which the delay value is given). For link (1,2), the transition probabilities \( p_1 \) are given and the delay values for states 1 and 2 are 10 and 1, respectively. Assume we want to find a path from node 1 to node 4 that has smallest expected delay. Upon arriving at node 1 it is disclosed that link (1,2) resides in state 1. By directly traversing the link (1,2) the delay will be 10, but by waiting for 1 time unit the expected delay would be 1(waiting time)+0.10(transition time at state 1)+0.90(transition time at state 2) =9.9. Since going along link (1,3) will consume 8 time units, the optimum choice is to wait for 1 time unit and then check the state on link (1,2). When the state on link (1,2) after 1 time unit is still 1, the same approach is used to decide whether to wait or directly traverse link (1,2) or link (1,3).

Based on the Markovian link delay model, Orda et al. [10] consider four problems in which a routing policy is sought that minimizes the expected delay. The four variants differ in whether knowledge of the (past) states of previous visited nodes can be accumulated or not and whether the permitted number of hops is bounded or not. These problems are typically NP-hard, but \( \varepsilon \)-approximation solutions are presented. When nodal Markovian delays are assumed together with deterministic link delays, a polynomial-time algorithm can be devised.

![Fig. 3 Stochastic link delay models.](image)

<table>
<thead>
<tr>
<th>Link Model</th>
<th>Problem</th>
<th>Hardness</th>
<th>Solution</th>
<th>Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete Markov Nodal Delay Model [10]</td>
<td>DM</td>
<td>Polynomially solvable</td>
<td>Exact</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Discrete Markov Link Delay Model [10]</td>
<td>QDM</td>
<td>NP-hard</td>
<td>( \varepsilon )-approximation</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Time-Dependent Stochastic Delays [11]</td>
<td>LET</td>
<td>NP-hard</td>
<td>Exact</td>
<td>Exponential time</td>
</tr>
<tr>
<td>Probability Density Function [12]</td>
<td>SOTA</td>
<td>NP-hard</td>
<td>Exact</td>
<td>Exponential time</td>
</tr>
<tr>
<td>Probability Density Function [13]</td>
<td>MPDCP</td>
<td>NP-hard</td>
<td>FPTAS and ( \varepsilon )-approximation</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Link Delay Probability Function [14]</td>
<td>DBR</td>
<td>NP-hard</td>
<td>FPTAS</td>
<td>Polynomial time</td>
</tr>
<tr>
<td>Rate-Based Probability Function [15]</td>
<td>RBR</td>
<td>NP-hard</td>
<td>Exact and ( \varepsilon )-approximation</td>
<td>(Pseudo-)Polynomial time</td>
</tr>
</tbody>
</table>

C. Probability Distribution

Like discussed for off-line approaches to traffic uncertainty, there are also on-line approaches that assume that the delay on link \( (i,j) \) follows a known probability distribution function. Fan et al. [12] study the Stochastic On-Time Arrival (SOTA) problem, where given a delay constraint \( D \), the goal is to find a path from source to destination that has the biggest probability to have a delay no more than \( D \). Xiao et al. [13] consider only the Gaussian distribution: link delays are non-negative Gaussian random variables. Given a delay constraint \( D \), the probability \( \pi_D(p) \) that the delay of a path is no larger than \( D \) is:

\[
\pi_D(p) = \Phi \left( \frac{D - \mu(p)}{\sigma(p)} \right)
\]

where \( \Phi(x) \) is the probability density function of a Gaussian distribution, and \( \mu(p) \) and \( \sigma(p) \) denote the mean and standard deviation of the delay in path \( p \). Since \( \Phi(x) \) is an increasing function, to maximize \( \pi_D(p) \) is to maximize \( \frac{D - \mu(p)}{\sigma(p)} \). This maximization objective is referred to as the Most Probable Delay Constrained Path (MPDCP) problem. Xiao et al. [13] prove that the MPDCP problem is NP-hard and develop a Fully Polynomial Time Approximation Scheme\(^2\) (FPTAS) for when there exists a path whose mean delay is no more than \( D \) and they propose an approximation scheme for when no such path exists.

D. Link Delay Probability Function

Lorenz and Orda [14] assume that each link \( (i,j) \) has a function \( \pi_D(D_{ij}) \) that represents the probability that link \( (i,j) \) introduces a delay of no more than \( D_{ij} \) time units. The Delay-Based Routing (DBR) problem is to find a path that has the biggest probability of not exceeding \( D \). The DBR problem is proved to be NP-hard, but by decomposing the end-to-end delay constraint \( D \) into local delay constraints, an FPTAS can be obtained.

\(^2\) An FPTAS has a time complexity that is polynomial in both the problem size and \( \frac{1}{\varepsilon} \) and produces a solution that is within a factor \((1 + \varepsilon)\) of the optimal solution (or \((1 - \varepsilon)\) for maximization problems).
E. Rate-Based Probability Function

Guérin and Orda [15], instead of delay, use \( \pi_f(r) \) to represent the probability that link \((i,j)\) is able to allocate a rate \(r\). They study the Rate-Based Routing (RBR) problem, which is to find a path from source to destination that has the biggest probability to not exceed a given delay constraint \(D\). The delay of an \(n\)-hop path \(P\) is represented as

\[
d(P) = \frac{c + \sigma}{r} + \sum_{(i,j) \in P} d_{ij}
\]

where \(\sigma\) is the burst size of the requested flow, \(c\) is the flow's maximum packet length, \(d_{ij}\) is a static delay value corresponding to the link propagation delay and \(r\) is the minimal rate that can be guaranteed to the flow at each link along the path. Except for \(r\), all the other parameters in this model are fixed. Guérin and Orda [15] prove that the RBR problem is NP-hard and propose an \(\varepsilon\)-approximation algorithm.

F. Discussion of on-line approaches

On-line routing of dynamically arriving requests could be done by trying to obtain a good view of the state of the network and its resources (e.g., via an SDN controller) and based on that to compute appropriate paths. Generally, increasing the accuracy of link-state update policies comes at the expense of more signaling overhead. This overhead could be reduced by operating with stochastic link weights, although unfortunately most stochastic network routing problems are NP-hard and need heuristic algorithms to keep pace with the frequency of arriving requests.

IV. CONCLUSION

The aims of this article are two-fold:

On the one hand we have provided an overview of various models to deal with traffic uncertainty in network planning, both from an off-line as well as an on-line routing perspective. Our aim was not to provide an extensive survey, but rather to highlight some relevant models and to point to algorithms operating within the context of those models. These algorithms on their turn may be used as building blocks or inspiration for dealing with traffic uncertainty in one’s own network environment.

On the other hand, there is still much “uncertainty” in how to plan for networks with traffic uncertainty, and our aim is therefore also to stimulate new work on the topic. Possible directions for future work include:

1) Testing the validity of several traffic uncertainty models in practice. Which models are best applicable to which cases or do other models that better reflect reality need to be developed? Can we do much better in predicting future traffic?
2) Extending the work to also consider issues like quality of service and resiliency. For instance, how to deal with traffic diversity in off-line approaches or how to compute link-disjoint paths in stochastic networks.
3) Development of on-line routing algorithms that take the global network performance perspective into account.
4) Most of the work has been directed to the planning of logical topologies or routing schemes. Designing a physical network from scratch or adding/removing links with respect to uncertain traffic volumes has received less attention.

While uncertainty (either from not knowing traffic, or due to other elements like availability or possibly reduced power consumption) will remain a feature in network planning, technological advances are making the network more flexible in responding to unforeseen scenarios. We therefore believe that planning for the longer term via off-line approaches should go hand-in-hand with on-line routing schemes that react quickly to unforeseen or differently classified traffic.

REFERENCES

Song Yang (S) received the B.S. degree in software engineering and the M.S. degree in computer science from the Dalian University of Technology, Dalian, Liaoning, China, in 2008 and 2010, respectively. He is currently pursuing a Ph.D. degree in the Network Architectures and Services (NAS) group at Delft University of Technology, the Netherlands. His Ph.D. work mainly focuses on algorithmic aspects of routing in optical networks.

Fernando Kuipers (SM) is associate professor in the Network Architectures and Services group at Delft University of Technology (TUDelft). He received the M.Sc. degree in Electrical Engineering at TUDelft in June 2000 and subsequently obtained his Ph.D. degree (cum laude) in 2004 at the same faculty. Currently, he is chairman of the board of examiners for the two M.Sc. programs Computer Engineering and Embedded Systems at TUDelft, is on the steering board of the TUDelft PowerWeb consortium, and is member of the executive committee of the IEEE Benelux chapter on communications and vehicular technology.

His research interests mainly revolve around network algorithms and cover Routing, Quality of Service, Network Survivability, Optical Networks, and Content Distribution. His work on these subjects has resulted in over 50 publications in refereed journals and conferences and include distinguished papers at IEEE INFOCOM 2003, Chinacom 2006, IFIP Networking 2008, IEEE FMN 2008, IEEE ISM 2008, and ITC 2009.