

# Call Routing with Continuous Uncertainties

Aniket Gune and Jussi Keppo

Department of Industrial and Operations Engineering, University of Michigan

1205 Beal Avenue, Ann Arbor, MI 48109-2117, USA

Email: aniket@umich.edu, keppo@umich.edu

## **Abstract**

This paper considers call routing in a telecommunications network. The network demands are assumed to be distributed according to lognormal distributions and call routing takes place either directly between the start and end nodes or alternatively via an intermediate node. The call routing processes and blocking probabilities consider the demand uncertainty levels and demand correlations and they are solved in terms of trivariate normal distributions. The blocking probability can be viewed as an extension to the Erlang's loss probability formula. We illustrate our model with a numerical example and also show how our model can be used with a usual call routing method.

Key words: Telecommunications, network, optimal routing, Brownian motion.

## 1. Introduction

Call routing is an integral part of telecommunications networks and is of primal importance in network design and operation. Efficient call routing procedures have resulted in significant improvements in network connection availability and network robustness while simultaneously reducing network costs.

Telecommunications networks are in general modeled as graphs, which have nodes and edges. Graphs are a natural choice for telecommunications networks, because the networks are not fully meshed (i.e. all nodes are not connected to each other). In the graph, nodes act as endpoints for point-to-point connections. A direct point-to-point connection constitutes an edge also known as a link. Other connections are modeled as a sequence of links, often referred as routes, and correspond to paths in the graph.

Routing of calls takes place along various routes. Most of the telecommunications companies have extended traditional static routing methods to dynamic strategies. These strategies route calls depending on the given network load and, therefore, they guarantee better quality. Some of the best-known strategies employed by major telecommunications networks are DNHR (Dynamic Nonhierarchical Routing) and more recently RTNR (Real-time Network Routing) by AT&T and DAR (Dynamic Alternate Routing) by British Telecom. AT&T's DNHR is a decentralized non-hierarchical routing strategy (see e.g. Ash and Oberer [3]). DNHR increases network efficiency by taking advantage of the non-coincidence of the busy hours in the network. RTNR makes routing decision for each call (see e.g. Ash, Chen, Frey, and Huang [4]

and Ash [2]). If a direct point-to-point connection is blocked, the call is routed along an alternative path with the least load. Alternate path routing has a long and successful history within the telephone network (see e.g. Mitra and Seery [14], and Mitra, Gibbens, and Huang [15]) where the shortest path between connections is used until its capacity is reached, and then alternate paths are tried. In this paper we use this kind of routing.

In order to see how calls are routed in real-world telecommunications networks, as an example we discuss British Telecom's DAR method (see e.g. Gibbens [8], Stacey and Songhurst [24], and Kelly [11]) and later in Section 3 we connect our model to this routing method. DAR is based on the principles of Dynamic Alternate Routing and Random Sticky Principle (see e.g. Mees [13]) as follows. Suppose that a call is to be connected between London and Glasgow. If all the direct point-to-point connections between these cities are currently busy then London remembers an intermediate city, say Edinburgh, for this kind of overflow situation. Therefore, the network tries to route the call from London to Glasgow via Edinburgh. If the connection goes through then that two-link route is used for this call and is memorized for future incoming calls between London and Glasgow. If the two-link connection is rejected because it is currently too busy then this call cannot be routed (engaged tone is heard). The intermediate city is now reset randomly to a new city, say Manchester. The next overflow call on the direct routing between London and Glasgow will then attempt to use the new path London-Manchester-Glasgow.

Various assumptions on call demand are made to simplify the call routing models. For instance, Kelly [10, 11] has assumed an independent Poisson process for actual call arrival process and RTNR is used with gamma distributions (see e.g. Ash [2]). Norros [18] has modeled traffic demand with Fractional Brownian Motion and Addie, Zukermann, and Neame [1] with the discrete-time analog, Fractional Gaussian Noise. In our model we do not use network traffic directly. Instead, we model point-to-point demands, because their processes are independent of the network structure and, therefore, they can be used as network risk factors. We assume that these point-to-point demands are distributed according to lognormal distributions and we consider the correlations between the demands. Similar demand model is used, e.g., in Ryan [20]. We test the lognormal assumption with dial-up data and show that usually this assumption holds. We utilize the model in Keppo [12], where network routers are modeled as real options, and this way solve the expected routing processes and the blocking probabilities analytically in terms of trivariate normal distributions. Due to this blocking probability formula, we can explicitly consider the effect of demand correlations and uncertainty levels on the call routing. This is important since, for instance, according to Paxson and Floyd [19] the traffic is often highly correlated. Our blocking probability equation can be seen as an extension of Erlang's loss probability formula to the case of lognormal demands and the DAR type of call routing. We also link our model to trunk reservation, which is an important factor in the routing of calls since it has been shown to stabilize telecommunications networks (see e.g. Gibbens, Hunt, and Kelly [9]).

The remainder of this paper is organized as follows. Section 2 introduces the stochastic processes of the paper and then by using these processes in Section 3 we solve the expected call routing and blocking probabilities. Section 4 tests our stochastic process assumption with telecommunications data. Section 5 illustrates the model with the estimated model parameters and finally Section 6 concludes.

## 2. Model

We consider a telecommunications network where calls can be routed continuously within a time horizon  $[0, \tau]$ . Throughout our model, we analyze a simple network of three nodes because this is the basis in a DAR type of routing. Thus, as in Gibbens [8], Stacey and Songhurst [24], and Kelly [11] we consider the simple network structure that is used in the British Telecom's DAR method.

The network structure and the corresponding capacities and demands are illustrated in Figure 1.

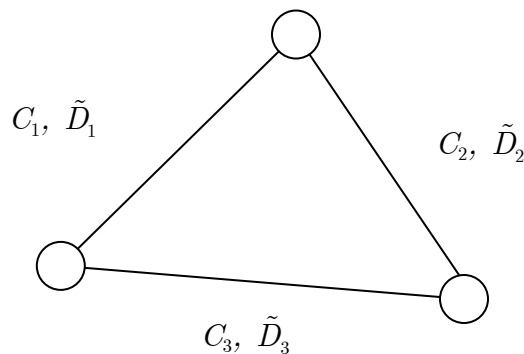


Figure 1. Network structure.  $C_i$  and  $\tilde{D}_i$  are the  $i$ 'th connection's capacity level and demand.

We assume that each time we know the current point-to-point demands and therefore we can model these demand processes. The demands of Figure 1 are independent of network routing, i.e., they are capacity demands between the start and end points over all the possible routings. The capacities are fixed and, for instance, the  $i$ 'th direct point-to-point connection could consist of  $C_i$  circuits where each circuit could route one call.

Let us next consider the stochastic process for the demands in Figure 1. Since the point-to-point demands are always positive let us model log-expected demands and define  $S_i^m(k) = \log(\tilde{D}_i(k \frac{T}{m}, T)) - \log(\tilde{D}_i((k-1) \frac{T}{m}, T))$  for all  $k \in \{1, \dots, m\}$  and  $m \in \{1, 2, \dots\}$ , where  $\tilde{D}_i(t, T) = E[\tilde{D}_i(T) | F_i]$  is the expected total demand of the  $i$ 'th point-to-point connection at time  $T$  calculated based on the information at time  $t$ ,  $m$  is the number of discrete time intervals on  $[0, T]$ , and  $k$  is the index for discrete times. This gives  $\log(\tilde{D}_i^m(T)) = \log(\tilde{D}_i(0, T)) + \sum_{k=1}^m S_i^m(k)$ , because  $\tilde{D}_i(T) = \tilde{D}_i(T, T)$ . Let us assume that  $\{S_i^m(k)\}$  are mutually independent and identically distributed random variables with mean  $\beta_i \frac{T}{m}$  and standard deviation  $\sigma_i \sqrt{\frac{T}{m}}$ . Then we get from the central limit theorem  $\lim_{m \rightarrow \infty} \sum_{k=1}^m (S_i^m(k) - \beta_i \frac{T}{m}) = \sigma_i B_i(T)$ , i.e.,  $\lim_{m \rightarrow \infty} \log(\tilde{D}_i^m(T)) = \log(\tilde{D}_i(0, T)) + \beta_i T + \sigma_i B_i(T)$ . Thus, if  $\beta_i = -\frac{1}{2} \sigma_i^2$  and if we assume that in the discrete-time the differences of the log-expected demands are independent and identically distributed then when we speed up the arrivals of the discrete-time events we get the following assumption that is empirically justified in Section 4.

*ASSUMPTION 2.1 The process of the expected  $i$ 'th demand is given by the following Itô stochastic differential equation*

$$d\tilde{D}_i(t, T) = \tilde{D}_i(t, T)\sigma_i(T)dB_i(t) \quad \text{for all } i \in \{1, 2, 3\}, \quad t \in [0, T], \quad T \in [0, \tau], \quad (2.1)$$

where  $\tilde{D}_i(t, T) = E[\tilde{D}_i(T)|F_t]$  is the expected total demand of the  $i$ 'th point-to-point connection at time  $T$  calculated based on the information at time  $t$ ,  $\sigma_i(T)$  is a deterministic and bounded volatility function,  $B_i(\cdot)$  is the Brownian motion corresponding to the  $i$ 'th point-to-point connection on the probability space  $(\Omega, F, P)$  along with the standard filtration  $\{F_t; t \in [0, \tau]\}$ , and we denote by  $\rho_{i,z}$  the correlation between the  $i$ 'th and  $z$ 'th Brownian motions.

According to Assumption 2.1 we model conditional expectations. Note that in equation (2.1) time  $T$  is fixed and, therefore,  $\tilde{D}_i(T) = \tilde{D}_i(t, T) + \sigma_i(T) \int_t^T \tilde{D}_i(y, T)dB_i(y)$  since  $\tilde{D}_i(T) = \tilde{D}_i(T, T)$ . Equation (2.1) implies that the stochastic process for the expected demand  $\tilde{D}_i(t, T) = E[\tilde{D}_i(T)|F_t]$  follows an exponential process where  $\tilde{D}_i^2(t, T)\sigma_i^2(T)$  is the rate of change of the conditional variance at time  $t$ . The boundedness of the volatility parameter guarantees the existence and uniqueness of the solution to (2.1). According to (2.1),  $\tilde{D}_i(T)$  follows a lognormal distribution with mean  $\tilde{D}_i(t, T)$  and variance  $\tilde{D}_i^2(t, T)[\exp(\sigma_i^2(T)(T-t)) - 1]$ . Similar demand model is used, e.g., in Zhao and Kockelman [25] and Ryan [20]. Since we model the expected value  $\tilde{D}_i(t, T)$ , the demand process  $\tilde{D}_i(t)$  can be e.g. a geometric Brownian motion or mean-reverting (see, for instance, Schwartz [23]). Also note that there can be cycles in the expected point-to-point demands. For instance, in our model we can have  $\tilde{D}_i(0, 1) = 100$  and  $\tilde{D}_i(0, 1.1) = 1$ . Due to call duration the demand process in (2.1) might be path depended and we model this dependency in the expected demand  $\tilde{D}_i(t, T)$ .

### 3. Optimal Call Routing

In this section we derive a call routing model based on the model in Section 2. We use a DAR type of routing (see e.g. Kelly [10]) and assume that the call is routed either directly or alternatively via an intermediate node using the Random Sticky Principle. Specifically, we assume:

*ASSUMPTION 3.1 Call routing depends only on the network capacities and the current point-to-point demand levels. The alternative routing is selected if there is no free capacity on the direct routing.*

Assumption 3.1 implies that there is no direct effect from call duration on the routing, only an indirect effect from the demand process. We use Assumption 3.1 in order to solve the routing problem analytically and similar assumption is made, for instance, in Mitra, Morrison, and Ramakrishnan [16, 17].

From Assumption 3.1 we first solve the expected routing processes in this section. Then in the next section we calculate the corresponding blocking probabilities and by using these blocking probabilities we show how our framework can be linked to DAR and to the alternative routing optimization. Without loss of generality we will mainly consider the first connection. The other point-to-point connections can be analyzed in the same way.

Using Assumption 3.1 the demand  $\tilde{D}_i(\cdot)$  can be divided into three different components: direct routing demand, alternative routing demand, and blocked demand. That is,

$$\tilde{D}_i(t) = D_i^d(t) + D_i^a(t) + D_i^b(t) \quad \text{for all } i \in \{1, 2, 3\}, \quad t \in [0, \tau], \quad (3.1)$$

where  $D_i^d(t)$  and  $D_i^a(t)$  are the demands for the direct and the alternative routings at time  $t$  and  $D_i^b(t)$  is the demand that is blocked at time  $t$ .

Using equation (3.1) we can solve the traffic on the physical links. For example, the demand that is routed through the physical link between the up and left points in Figure 1 is given by

$$D_1(t) = D_1^d(t) + D_2^a(t) + D_3^a(t). \quad (3.2)$$

Note that  $D_1$  is the traffic on the physical link between the points and  $\tilde{D}_1$  is the demand that requests the service between these points and it has the representation of equation (3.1).

As mentioned earlier, in our network the direct routing in (3.1) is used within its capacity constraint. That is,

$$D_i^d(t) = \tilde{D}_i(t) \mathbf{1}\{\tilde{D}_i(t) \leq C_i\} + C_i \mathbf{1}\{\tilde{D}_i(t) \geq C_i\}, \quad (3.3)$$

where  $\mathbf{1}\{x \geq y\} = \begin{cases} 1, & \text{if } x \geq y \\ 0, & \text{otherwise} \end{cases}$ . The alternative routing of (3.1) is used only if the direct

routing is full. We approximate the alternative routing as follows

$$D_i^a(t) = r_i(t) \prod_{k \in \{1, 2, 3\} - \{i\}} \mathbf{1}\{\tilde{D}_k(t) \leq C_k - r_i(t)\}, \quad (3.4)$$

where the excess demand  $r_i(t) = (\tilde{D}_i(t) - C_i) \mathbf{1}\{\tilde{D}_i(t) \geq C_i\}$ . According to (3.4), the alternative routing demand is equal to the excess demand if there is free capacity for the whole excess demand. Note that the correct alternative demand is

$$\begin{aligned}
D_i^a(t) &= r_i(t) \prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1}\{\tilde{D}_k(t) \leq C_k - r_i(t)\} + \min_{k \in \{1,2,3\} - \{i\}} [C_k - \tilde{D}_k(t)] \cdot \\
&\quad \mathbf{1}\{\tilde{D}_i(t) \geq C_i\} \prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1}\{C_k - r_i(t) < \tilde{D}_k(t) \leq C_k\},
\end{aligned} \tag{3.5}$$

i.e., we assume that the last term is zero which is the alternative routing in the case there is no capacity for the whole excess demand. Because this last term is nonnegative, equation (3.4) gives the lower bound for the alternative routing demand. Note that even though we use (3.4), it will not affect the blocking probabilities as can be seen from Proposition 4.1.

Because we have (3.4) for all  $i \in \{1,2,3\}$ , in our model  $D_i^a(t)D_k^a(t) = 0$  for all  $i \in \{1,2,3\}$  and  $k \in \{1,2,3\} - \{i\}$ . That is, only one alternative routing can be nonzero at a time and therefore, for instance, in equation (3.2)  $D_2^a(t)D_3^a(t) = 0$ . This means that in our network only one demand can be rerouted at a time since if one demand is rerouted then there is no free capacity for the other demands' alternative routings.

The corresponding blocked demand:

$$\begin{aligned}
D_i^b(t) &= \tilde{D}_i(t) - \tilde{D}_i(t) \mathbf{1}\{\tilde{D}_i(t) \leq C_i\} - C_i \mathbf{1}\{\tilde{D}_i(t) \geq C_i\} - \\
&\quad (\tilde{D}_i(t) - C_i) \mathbf{1}\{\tilde{D}_i(t) \geq C_i\} \prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1}\{\tilde{D}_k(t) \leq C_k - r_i(t)\} \\
&= r_i(t) \left[ 1 - \prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1}\{\tilde{D}_k(t) \leq C_k - r_i(t)\} \right].
\end{aligned} \tag{3.6}$$

Equations (3.4) and (3.6) imply that in our network if  $D_i^b(t) > 0$  then  $D_i^a(t) = 0$ . Note that because (3.4) is the lower bound for the alternative routing, equation (3.6) gives the upper bound for the blocked demand. If only a small fraction of  $r_i(t)$  cannot be rerouted then in (3.6) the whole  $r_i(t)$  is blocked.

Due to the alternative routing modeling we explicitly state the following assumption that is the same as equation (3.4) with  $t = T$  and the expected excess demand,  $r_i(t, T)$ , in the indicator. This assumption is needed in Proposition 3.1, where we calculate the expected alternative routing demand. In Section 5 we analyze the error term from this approximation.

*ASSUMPTION 3.2 We model the expected  $i$ 'th alternative routing at time  $T$  based on the information at time  $t$  as follows*

$$E[D_i^a(T)|F_t] = E\left[r_i(T) \prod_{k \in \{1,2,3\}-\{i\}} \mathbf{1}\{\tilde{D}_k(T) \leq C_k - r_i(t, T)\} \middle| F_t\right] \quad (3.7)$$

for all  $i \in \{1, 2, 3\}$ ,  $t \in [0, T]$ ,  $T \in [0, \tau]$ ,

where the expected excess demand  $r_i(t, T) = E[r_i(T)|F_t]$  and the excess demand  $r_i(T) = (\tilde{D}_i(T) - C_i) \mathbf{1}\{\tilde{D}_i(T) \geq C_i\}$ . Equation (3.7) implies that in the modeling of alternative routing demand we assume: (i) the second term in (3.5) is zero, (ii) inside the indicator of (3.7) the expected excess demand,  $r_i(t, T)$ , is used.

In the modeling of alternative routing demand we use both (i) and (ii) in Assumption 3.2, and in the calculating of blocking probabilities we need only (ii). As we will see in Section 5, this gives that our blocking probability model is close to the true value. Note that the blocking probability model is more important in the network routing, because the objective in the routing is to minimize the network blocking.

The term  $r_i(T)$  in (3.7) is the excess demand of the  $i$ 'th direct routing at time  $T$  and the term  $\prod_{k \in \{1,2,3\}-\{i\}} \mathbf{1}\{\tilde{D}_k(T) \leq C_k - r_i(t, T)\}$  is the indicator of the case where there exists enough free capacity on the alternative routing for the expected excess demand. Thus, in (3.7) we assume

that if on the alternative routing there is free capacity for the expected excess demand then the whole excess demand can be routed. The term  $r_i(t, T)$  is an approximation for  $r_i(T)$  and therefore the expected alternative routing from (3.4) is as follows

$$E[D_i^a(T)|F_t] = E\left[\left(\tilde{D}_i(T) - C_i\right)\mathbf{1}\{\tilde{D}_i(T) \geq C_i\} \prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1}\{\tilde{D}_k(T) \leq C_k - r_i(T)\} \middle| F_t\right].$$

The term  $r_i(t, T)$  in Assumption 3.2 can be viewed as a trunk reservation that is used in practice to avoid instability in telecommunications networks (see e.g. Gibbens, Hunt, and Kelly [9]).

Using Keppo [12] and equations (2.1) and (3.3)-(3.7) we get the following proposition.

**PROPOSITION 3.1** *The expected first connection's demand terms in equation (3.1) are given by*

$$\begin{aligned} D_1^d(t, T) &= \tilde{D}_1(t, T)N(d_1(t, T) - \sigma_1\sqrt{T-t}) + C_1N(-d_1(t, T)) \\ D_1^a(t, T) &= \tilde{D}_1(t, T)G(-d_1(t, T) + \sigma_1\sqrt{T-t}, d_2(t, T), d_3(t, T), -\rho_{1,2}, -\rho_{1,3}, \rho_{2,3}) - \\ &\quad C_1G(-d_1(t, T), d_2(t, T), d_3(t, T), -\rho_{1,2}, -\rho_{1,3}, \rho_{2,3})) \\ D_1^b(t, T) &= \tilde{D}_1(t, T)\left[N(-d_1(t, T) + \sigma_1\sqrt{T-t}) - \right. \\ &\quad \left. G(-d_1(t, T) + \sigma_1\sqrt{T-t}, d_2(t, T), d_3(t, T), -\rho_{1,2}, -\rho_{1,3}, \rho_{2,3})\right] - \\ &\quad C_1\left[N(-d_1(t, T)) - G(-d_1(t, T), d_2(t, T), d_3(t, T), -\rho_{1,2}, -\rho_{1,3}, \rho_{2,3})\right] \end{aligned} \tag{3.8}$$

for all  $t \in [0, T]$ ,  $T \in [0, \tau]$ ,

where  $D_1^a(t, T) = E[D_1^a(T)|F_t]$ ;  $D_1^b(t, T) = E[D_1^b(T)|F_t]$ ;  $\tilde{D}_1(t, T) = E[\tilde{D}_1(T)|F_t]$ ;  $N(\cdot)$  is a cumulative standard normal distribution;  $G(-d_1, d_2, d_3, -\rho_{1,2}, -\rho_{1,3}, \rho_{2,3})$  is the area under a standard trivariate normal distribution function covering the region from  $-\infty$  to  $-d_1$ ,  $-\infty$  to  $d_2$ , and  $-\infty$  to  $d_3$ , the three random variables have correlations  $-\rho_{1,2}$ ,  $-\rho_{1,3}$ , and  $\rho_{2,3}$ ;

$$d_1(t, T) = \frac{\ln\left(\frac{c_1}{\tilde{D}_1(t, T)}\right) + \frac{1}{2}\sigma_1^2(T)(T-t)}{\sigma_1(T)\sqrt{T-t}}; \quad d_k(t, T) = \frac{\ln\left(\frac{c_k - r_1(t, T)}{\tilde{D}_k(t, T)}\right) + \frac{1}{2}\sigma_k^2(T)(T-t)}{\sigma_k(T)\sqrt{T-t}} \quad \text{for all } k \in \{2, 3\};$$

$$\text{and } r_1(t, T) = \tilde{D}_1(t, T)N\left(-d_1(t, T) + \hat{\sigma}_1\sqrt{T-t}\right) - C_1N\left(-d_1(t, T)\right).$$

PROOF. See Appendix.

By using equation (3.2) and Proposition 3.1 with all the point-to-point connections we get

$$\begin{aligned} E[D_1(T)|F_t] &= D_1^d(t, T) + D_2^a(t, T) + D_3^a(t, T) \\ E[D_2(T)|F_t] &= D_2^d(t, T) + D_1^a(t, T) + D_3^a(t, T) \\ E[D_3(T)|F_t] &= D_3^d(t, T) + D_1^a(t, T) + D_2^a(t, T). \end{aligned} \tag{3.9}$$

That is, by using Proposition 3.1 and equation (3.9) we can calculate the expected call traffics on the physical links in Figure 1. Then these expected traffics can be used, e.g., in the network design. Note that due to the indicator functions in the direct and alternative routing demands the direct modeling of the  $D$ -processes of (3.2) and (3.9) is difficult. That is, from Proposition 3.1 we can see that the process parameters of the direct and alternative routing demands are not deterministic because they depend (nonlinearly) on the  $\tilde{D}$ -processes. Since these  $\tilde{D}$ -processes are independent of the routing options, they are more natural modeling objects than the  $D$ -processes. Once the routing independent  $\tilde{D}$ -demands are modeled, e.g., based on historical data then the direct and alternative routing demands can be solved from Proposition 3.1.

Usually network data consists of observations from the  $D$ -processes of equations (3.2) and (3.9). However, as mentioned earlier, in practice the direct modeling of these  $D$ -processes is difficult and our model implies that this is due to the nonlinear option type characteristic in

the direct and alternative routing processes. If the direct and alternative routing processes are modeled directly then this is similar situation as the modeling of a stock option's process in financial markets from the option's price history. Usually this financial modeling is done by first modeling the underlying stock price from the market data, because its process parameters can be assumed to be constant, and then the stock option's price is described as a nonlinear function (for instance Black and Scholes model [5]) of the underlying stock. Now this financial framework is translated into our network model by viewing the routing processes as options and the  $\tilde{D}$ -demands as the underlying assets. Further, even though there were no observations on the  $\tilde{D}$ -processes their implied values could be estimated from  $D$ -demand data and equations (3.8) and (3.9).

According to Proposition 3.1 the probability that only the direct routing is used is  $N(d_1(t, T))$  and that the alternative routing is used is  $G(-d_1(t, T), d_2(t, T), d_3(t, T), -\rho_{1,2}, -\rho_{1,3}, \rho_{2,3})$ . Thus, the expected routing processes can be solved analytically in terms of trivariate normal distributions. For the calculation of trivariate normal distributions see, e.g., Genz [7]. Due to the importance of the blocking probabilities, we next study those in greater detail.

#### 4. Blocking Probability

By the conditional probability, the blocking probability of a particular connection can be represented as follows

$$\begin{aligned}
P(\text{blocking}) &= P(\text{blocking of the direct and alternative routings}) \\
&= P(\text{blocking of the direct routing}) \cdot \\
&\quad P(\text{blocking of the alternative routing} \mid \text{blocking of the direct routing}).
\end{aligned} \tag{4.1}$$

Hence, the blocking indicator of the  $i$ 'th connection is as follows

$$\mathbf{1}\{\text{blocking of the } i\text{th connection}\} = \mathbf{1}\{\tilde{D}_i(t) \geq C_i\} \left( 1 - \prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1}\{\tilde{D}_k(t) \leq C_k - r_i(t)\} \right), \tag{4.2}$$

where  $\mathbf{1}\{\tilde{D}_i(t) \geq C_i\}$  is the indicator that the direct routing is full,

$\prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1}\{\tilde{D}_k(t) \leq C_k - r_i(t)\}$  is the indicator that both the alternative routing's point-to-point

connections are free and, therefore,  $1 - \prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1}\{\tilde{D}_k(t) \leq C_k - r_i(t)\}$  is the indicator that there

is no enough free capacity on the alternative routing. Thus,

$\mathbf{1}\{\text{blocking of the } i\text{th connection}\} = 1$  implies an agent placing a call will not get the service since the call is blocked. As we noted earlier, equation (4.2) implies that only (ii) in Assumption 3.2 is used in the calculating of the blocking probability.

From Proposition 3.1 we get the following result that gives the blocking probability.

**PROPOSITION 4.1** *The first connection's blocking probability at time  $T$  based on the information at time  $t$  is given by*

$$N(-d_1(t, T)) - G(-d_1(t, T), d_2(t, T), d_3(t, T), -\rho_{1,2}, -\rho_{1,3}, \rho_{2,3}) \quad \text{for all } t \in [0, T], \quad T \in [0, \tau]. \tag{4.3}$$

**PROOF.** From Proposition 3.1 and equation (4.2) we get (4.3). □

Proposition 4.1 gives the first connection's probability for not getting the service by routing the call either directly or indirectly. Traditionally, the blocking probability of one direct point-to-point connection is calculated by using the Erlang's loss probability formula:

$$E(\nu, C) = \frac{\nu^C}{C!} \left[ \sum_{n=0}^C \frac{\nu^n}{n!} \right]^{-1}, \quad (4.4)$$

where the point-to-point calls are modeled by using a Poisson process with intensity  $\nu$  and  $C$  is the point-to-point capacity. In equation (4.4) a call is blocked and lost if the point-to-point demand is greater than the capacity. Note that in our model we use different demand processes and consider the alternative routing and, therefore, also the correlations between the point-to-point demands. Equation (4.4) is extended, e.g., by Gibbens, Hunt and Kelly [9]. In their model  $K$  nodes are linked to form a graph and then the loss probability is given by

$$B = e(\nu[1 + 2B(1 - B)], C), \quad (4.5)$$

where  $e(\cdot)$  is the Erlang's formula,  $B$  is the blocking probability,  $\nu[1 + 2B(1 - B)]$  is the modified Poisson process rate that replaces  $\nu$  in (4.4), and  $C$  is the capacity between any pair of nodes. Thus, the blocking probability  $B$  is the solution to the fixed-point problem of (4.5). As with the Erlang's function, the differences between Proposition 4.1 and (4.5) are that our model considers alternative routing and different demand processes as well as the demand correlations.

#### 4.1 Connection to DAR

Proposition 4.1 can be used in the alternative routing selection in cases where we have a more complex network structure than was presented in Figure 1. If several alternative routing candidates exist then the optimal one can be chosen according to

$$\min_{x \in \{1, \dots, m\}} \left\{ N(-d_1(t, T)) - G(-d_1(t, T), d_2^x(t, T), d_3^x(t, T), -\rho_{1,2}^x, -\rho_{1,3}^x, \rho_{2,3}^x) \right\}, \quad (4.6)$$

where  $m$  is the number of alternative routings that have free capacity at time  $t$ ,  $N(-d_1(t, T)) - G(-d_1(t, T), d_2^x(t, T), d_3^x(t, T), -\rho_{1,2}^x, -\rho_{1,3}^x, \rho_{2,3}^x)$  is the blocking probability corresponding to the alternative routing  $x \in \{1, \dots, m\}$ , and  $T \in [t, \tau]$  can be set to correspond to the expected call duration. Thus, (4.6) can be seen as a method which uses capacity on the alternative route that has the smallest blocking probability due to the first connection's rerouting and this way minimizes the whole network's blocking probability. Because  $d_1(t, T)$  in (4.6) is independent of the alternative routing, the minimization gives

$$\max_{x \in \{1, \dots, m\}} \left\{ \frac{G(-d_1(t, T), d_2^x(t, T), d_3^x(t, T), -\rho_{1,2}^x, -\rho_{1,3}^x, \rho_{2,3}^x)}{N(-d_1(t, T))} \right\}. \quad (4.7)$$

Note that in (4.7) the alternative routings' point-to-point demand variances and correlations are important. Further, if  $m$  is high then heuristic methods can decrease the computational time significantly. In Section 5 we discuss more about this approach.

Equations (4.6) and (4.7) are similar to the DAR method (see Mees [13] and Kelly [11]). For illustration let us consider the simplified version of the DAR model and let  $p_x(i)$  denote the  $i$ 'th connection's long run proportion of calls that are routed by using alternative routing  $x$ .

This alternative routing is used only when the direct routing is full. Let  $q_x(i)$  be the  $i$ 'th connection's long run proportion of calls that are routed through the intermediate node  $x$  and that are blocked. That is,

$$q_x(i) = P(\text{blocking of alternative routing } x \mid \text{blocking of the } i\text{'th direct routing}). \quad (4.8)$$

In DAR the probabilities of (4.8) are estimated directly from the blocking data. Then the  $i$ 'th connection's optimal routing is selected by using routing probabilities  $(p_1(i), \dots, p_m(i))$  and the Random Sticky Principle (see e.g. Mees [13]) as follows

$$p_a(i)q_a(i) = p_b(i)q_b(i) \quad \text{for all } a \in \{1, \dots, m\}, \quad b \in \{1, \dots, m\} - \{a\} \quad (4.9)$$

where  $m$  is the number of alternative routings. Thus, (4.9) solves the probabilities  $(p_1(i), \dots, p_m(i))$  because  $\sum_{x=1}^m p_x(i) = 1$  and  $q_a(i)$  and  $q_b(i)$  are received from the network data.

Using our framework, (4.8) can be modeled in the following way (with  $i = 1$ )

$$q_x(1) = 1 - \frac{1}{T-t} \int_t^T \frac{G(-d_1(t, y), d_2^x(t, y), d_3^x(t, y), -\rho_{1,2}^x, -\rho_{1,3}^x, \rho_{2,3}^x)}{N(-d_1(t, y))} dy, \quad (4.10)$$

where  $T - t$  is the selected time horizon, e.g., the expected call duration, and

$$\begin{aligned} &P(\text{blocking of alternative routing } x \text{ at time } T \mid \text{blocking of the } i\text{'th direct routing at time } T) \\ &= 1 - \frac{G(-d_1(t, T), d_2^x(t, T), d_3^x(t, T), -\rho_{1,2}^x, -\rho_{1,3}^x, \rho_{2,3}^x)}{N(-d_1(t, T))}. \end{aligned}$$

Note that equations (4.9) and (4.10) are similar to (4.7). Equation (4.10) gives the first connection's expected proportion of blocked calls that are routed through the intermediate node  $x$  during  $t-T$ . Thus, by using equation (4.10) our framework can be applied with the DAR

method. The difference between the usual DAR and equations (4.9) and (4.10) is that our model is forward looking, i.e., it calculates the future blocking probabilities based on the current demand processes while the DAR method is basing its blocking probabilities directly on the historical blocking proportions. As we noted earlier, because of the routing options the direct modeling of the blocking processes is difficult and the  $\tilde{D}$ -demands are more natural modeling objects since they are independent of the routing options. Further, it is worth noting that our model can also be used other way round. That is, given the blocking probabilities, for instance from blocking data, we can calculate the implied  $\tilde{D}$ -processes. Then, by Proposition 4.1, these demand processes can be used in the calculation of future blocking probabilities.

## 5. Empirical Analysis

Our empirical analysis is based on the University of Michigan's (Ann Arbor) dial-up data between October 12, 2003 – January 12, 2004. The data consists of the number of dial-up connections every 15 minutes. This implies that we have 8,832 data points. Only the University of Michigan affiliated people use these dial-up connections. There are hourly, daily, and monthly cycles. Here we focus only on the hourly and daily cycles, i.e., we estimate the demand pattern over a period of one week, from Sunday 12:00 to the next Saturday 23:45. The data points are represented as follows:

*Data point at time  $t_k = 100 \cdot \text{average number of ports in use on } [t_{k-1}, t_k] / \text{total number of ports}$*

for all  $k \in \{1, 2, \dots, 672\}$ , where  $t_0 = \text{Sunday 12:00}$ ,  $t_{672} = \text{Saturday 23:45}$ , 672 is the number of observations during a week, and the total number of dial-up ports is 1,104. For instance, the value 70 means that 70% of the dial-up ports are in use.

We use the dial-up data to model the point-to-point call demand process in Assumption 2.1. Assumption 2.1 implies that the total point-to-point call demand follows a lognormal distribution. In our dial-up data set all the demand values are less than 100, i.e., there is no blocking and, hence, the data is convenient for modeling the total call demands. We assume that the dial-up data has similar characteristics as the call point-to-point demands. For instance, since there are high daily cycles in the dial-up demand we expect similar cycles in the call demands.

Figure 2 illustrates the hourly and daily cycles in the dial-up data. In the figure there are average demand (solid line) and the 95% confidence interval (dotted lines). The average demand for each 15-minute interval is calculated by taking the average demand of that 15-minute time interval on the same weekday between October 12, 2003 and January 12, 2004. The average demand gives directly the expected curve  $\tilde{D}(0, \cdot) : [\text{Sunday 12:00}, \text{Saturday 23:45}] \rightarrow \mathbf{R}_+$  in Assumption 2.1. From Figure 2 we get, for instance,  $\tilde{D}(0, \text{Sunday 16:30}) = 57$  and  $\tilde{D}(0, \text{Monday 21:15}) = 68$ . And  $\tilde{D}(0, \text{Sunday 16:30})$  is calculated by taking the average demand at 16:30 on Sundays. As discussed earlier, there are cycles in the demand and similar characteristics can be observed in other non-storable commodity markets, for instance in electricity markets (for the modeling of electricity demands see e.g. Räsänen, Ruusunen, and Hämäläinen [21, 22]). We also note from Figure 2 that the

demand pattern is different during regular weekdays than during weekend. Thus, the call demand is similar on Saturday and Sunday and it is similar on weekdays (Monday - Friday).

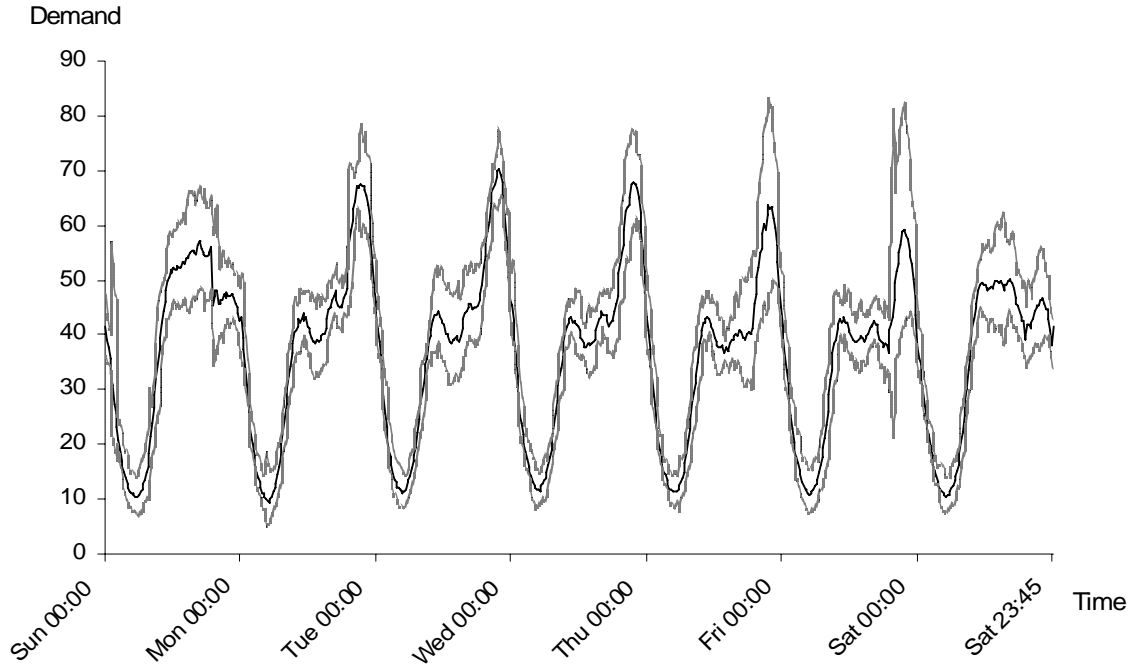
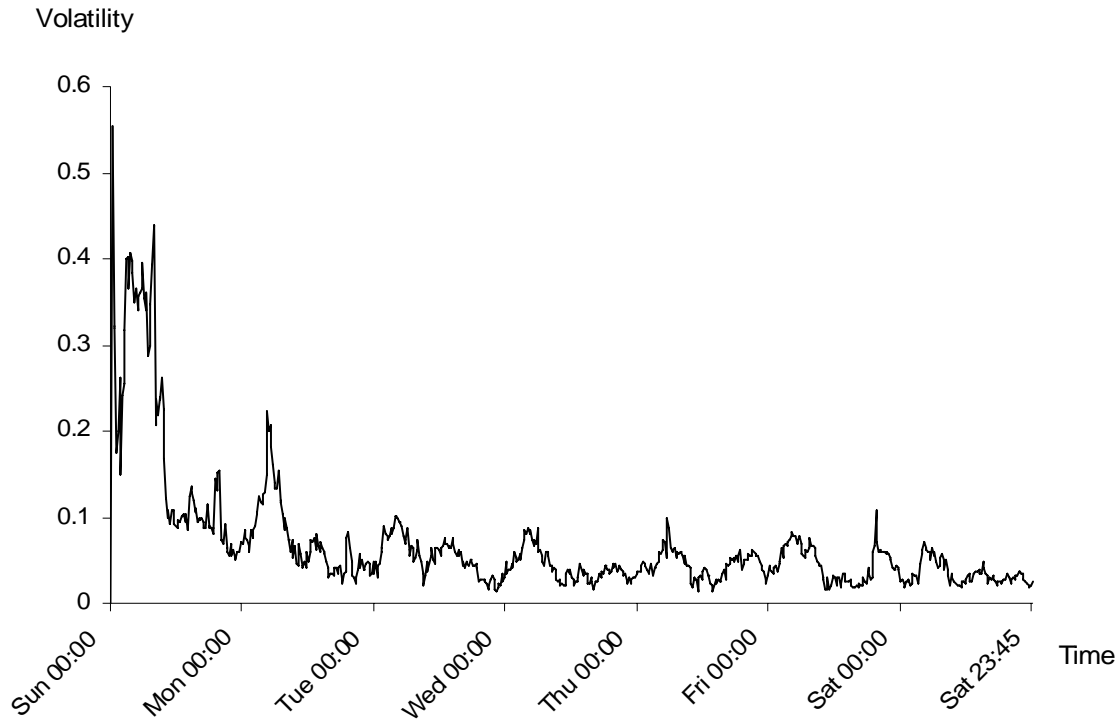


Figure 2. Average dial-up demand (solid line) and the 95% confidence interval (dotted lines).

Figure 3 shows the volatility structure of the demand based on Figure 2 and assuming that the present time is Sunday 12:00. This volatility structure is calculated by using Assumption 2.1, i.e., by using the variance equation:  $\tilde{D}^2(0, T) [\exp(T\hat{\sigma}^2(T)) - 1]$ , where  $\sigma(T)$  is the demand volatility for time  $T$ . As can be seen from Figure 3, there are no as strong cycles in the volatility pattern as there are in the expected demand. This is because there are no strong proportional cycles in the width of the confidence interval in Figure 2. Note that the volatilities are higher than the volatilities, e.g., in financial markets. The average daily volatility is 6.8% which corresponds to 129.9% annual volatility.



*Figure 3. The volatility structure of dial-up demand. In the calculation of the structure time is measured in days.*

As we noted earlier, Assumption 2.1 implies that the demand follows a lognormal distribution. According to Figure 2 all weekdays behave similarly and therefore we map them together and analyze the resulting histogram. Figure 4 illustrates the histogram of the dial-up demand for one 15-minute interval (16:00 - 16:15 on weekdays) and the corresponding lognormal function. According to the figure the demand distribution is close to the lognormal distribution.

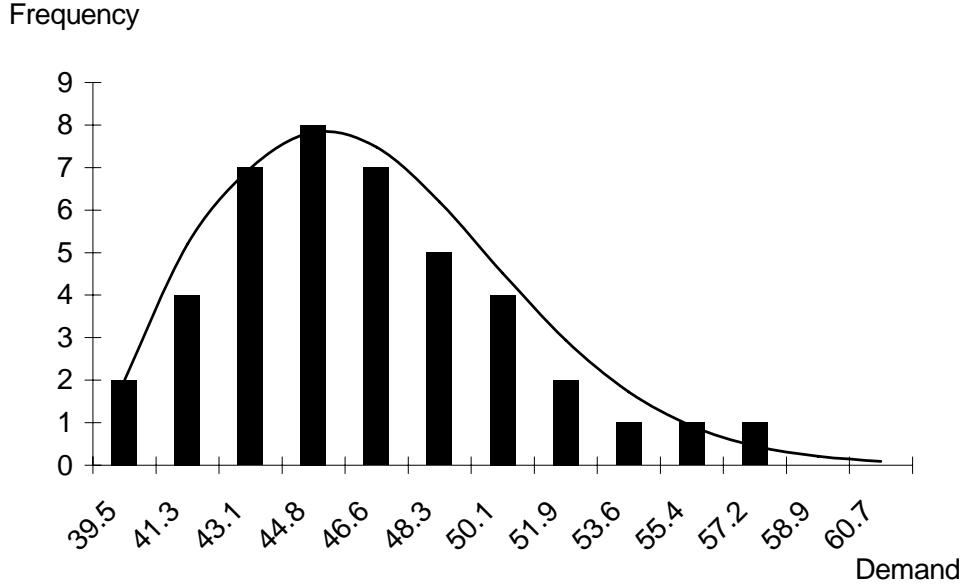


Figure 4. Histogram of the dial-up demand between 16:00 –16:15 on weekdays and the corresponding lognormal function.

In order to analyze further Assumption 2.1 we use the one dimensional D’Agostino-Pearson’s test (see e.g. D’Agostino and Stephens [6]) for the log-demand. According to this test 608 out of the 672 (672 periods of 15 minutes over a week) log-demands are normally distributed. In the test we used 5% significance level. Thus, about 90% of the log-demands passed the D’Agostino-Pearson’s normality test. This implies that Assumption 2.1 is close to the true expected demand process.

## 6. Example

This section illustrates our framework with numerical examples by using the estimated demand processes in Section 5. We assume that all the three point-to-point demand processes

follow the process illustrated in figures 2 and 3. However, in order to have blocking we assume  $C_i = 70$  for all  $i \in \{1,2,3\}$ , i.e., we decrease the capacities from 100 to 70. We assume that the present time is Sunday 12:00. Then we calculate the first connection's blocking probability with different correlation structures by using Proposition 4.1. Note that with these values there is no blocking on Sunday at 12:00 because  $\tilde{D}_1(\text{Sunday 12:00}) < C_1$ . The following figure illustrates the first connection's blocking probability during the week with two different correlation structures.

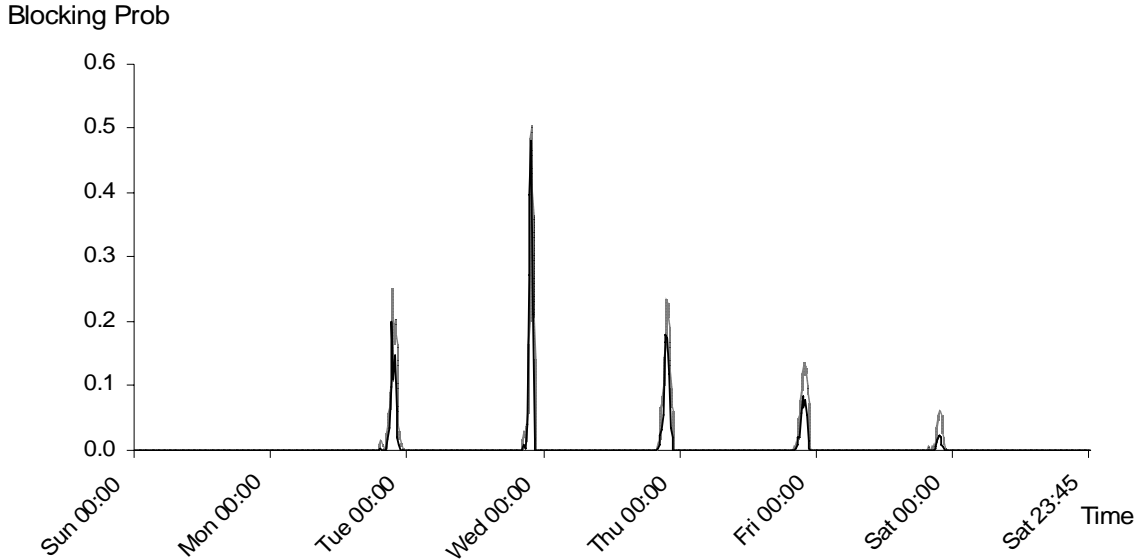


Figure 5. First connection's blocking probabilities with correlations  $(0.5, 0.5, 0.5)$ , dotted line, and  $(0, 0, 0)$ , solid line.

In Figure 5 the dotted line corresponds to the case where the correlations between the demands are 0.5 and the solid line represents the zero correlation case. The average percentage difference between the blocking probability spikes is 22%. Thus, increasing the correlations

from zero to 0.5 increases the blocking probability spikes by 22%. This implies that the correlations have a strong effect on the blocking probability. As can be seen from the figure, blocking probability is higher during the weekdays than during the weekend. This is due to the expected demand in Figure 2.

Next we select two times: Monday 21:30 and Wednesday 21:30, and analyze the expected routing processes and blocking probabilities at those times. We assume that for both the times the expected demand process is given as follows

$$d\tilde{D}_i(t, T) = 0.04 \cdot \tilde{D}_i(t, T)dB_i(t); \tilde{D}_i(\text{Sunday 12:00}, T) = 68 \quad \text{for all } i \in \{1, 2, 3\}, \quad (6.1)$$

where  $T \in \{\text{Monday 21:30}, \text{Wednesday 21:30}\}$  and the process parameters are obtained from figures 2 and 3.

The following table illustrates the first connection's expected routing demands and blocking probabilities with different time horizons and correlation structures.

*Table 1. First connection's expected routing demands and blocking probabilities.*

$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{2,3}$	$T$	$D_1^d$	$D_1^a$	$D_1^b$	Blocking Probability
0	0	0	Monday 21:30	67.277	0.298	0.424	0.170
0	0	0	Wednesday 21:30	66.681	0.422	0.896	0.233
0.5	0.5	0.5	Monday 21:30	67.277	0.331	0.392	0.226
0.5	0.5	0.5	Wednesday 21:30	66.681	0.475	0.843	0.283

In Table 1 the two first lines correspond to the case where the correlations between the demands are zero and the next two lines the case with correlations equal to 0.5. Table 1

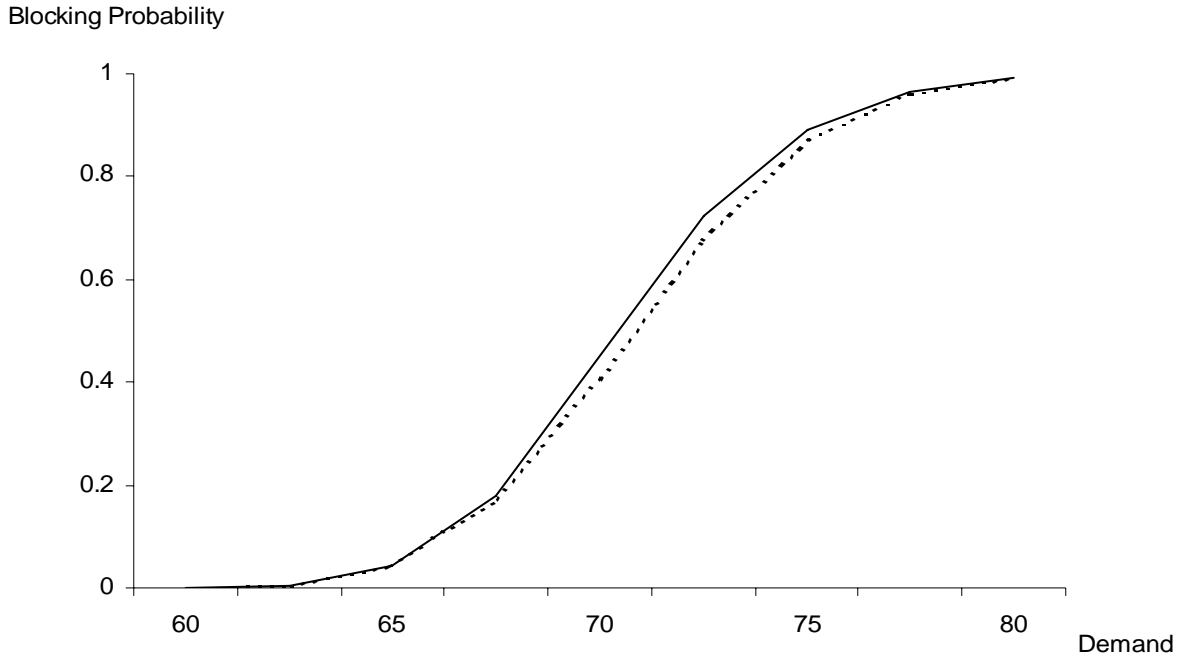
indicates that due to the initial demand situation  $\tilde{D}_1(\text{Sunday 12:00}) < C_1$  the blocking probability is an increasing function of time  $T$ . Further, the higher the time  $T$  the lower the expected direct routing demand and the higher the alternative routing and blocking demands. This is because on Sunday at 12:00 the excess demand is zero and with high  $T$  the demands have a high variance and, therefore, there is a high probability for very high and low demand values. From Table 1 we also see that the greater the correlations between the demands the higher the blocking probability. Comparing the blocking probabilities at  $T = \text{Monday 21:30}$  and Wednesday 21:30 we see that if correlations are increased from 0 to 0.5 then the blocking probability increases almost 21.5%. Note that if the correlations are high then the blocking of the direct routing indicates high blocking probability for the alternative routing. Due to (i) in Assumption 3.2 the expected blocked demand  $D_1^b$  decreases as a function of the correlations. This is an approximation error and since it is not used in the blocking probability calculation, the probability is an increasing function with respect to the correlations.

In order to analyze the error from (ii) in Assumption 3.2 we compare our analytical blocking probability, Proposition 4.1, with Monte Carlo Simulation with 40 000 simulations. The simulation is based on

$$E \left[ \mathbf{1} \left\{ \tilde{D}_i(T) \geq C_i \right\} \left( 1 - \prod_{k \in \{1,2,3\} - \{i\}} \mathbf{1} \left\{ \tilde{D}_k(T) \leq C_k - r_i(T, T) \right\} \right) \middle| F_t \right]$$

i.e. we analyze the error from  $r_i(t, T) [\neq r_i(T, T)]$  in Proposition 4.1. We study the error term as a function of the initial demands, i.e., we change the initial demand in (6.1). The correlation between the demands is 0.5 and  $T = \text{Monday 21:30}$ . Figure 5 illustrates the difference between

the simulated and analytical results. The analytical and simulated blocking probabilities are close to each other. The small difference is due to (ii) in Assumption 3.2 but also partly from the numerical integrations with respect to the trivariate normal distribution.



*Figure 5. The difference between the simulated and analytical blocking probabilities. Dotted line is the result from the simulation (40 000 runs) and the solid line is the analytical result.*

Let us assume that there are two alternative routing candidates and we use Table 1. The first two rows of the table correspond to the first alternative routing and the last rows to the second alternative routing. From equation (4.6) we get that the first candidate is better alternative routing since it has smaller blocking probabilities. Further, if we use equations (4.9)

and (4.10) then the first alternative routing will be given a higher routing probability, i.e.,  $p_1 > p_2$ .

## 7. Conclusions

This paper considered optimal call routing in the case where demand processes are distributed according to lognormal distributions. This demand assumption was theoretically justified by a limit argument and empirically by using a dial-up data. Throughout our model, we analyzed a simple network of three nodes because this is the basis in British Telecom's Dynamic Alternative Routing (DAR). We illustrated how our model can be extended to more complex networks.

We solved the expected routing processes and the corresponding blocking probabilities analytically in terms of trivariate normal distributions. This way we were able to show that the demand uncertainty levels and correlations have a significant impact. Further, the analytical formulas are easily implemented to the everyday industry practice

## Acknowledgement

Research supported partly by Socio-Technical Infrastructure for Electronic Transactions at the University of Michigan. The authors are grateful to Jeffrey K. MacKie-Mason, Gaurav Shah, Xu Meng, and Sophie Shive for useful discussions. The authors also thank conference participants at INFORMS Annual Meeting 2003, the 28<sup>th</sup> Conference on Stochastic Processes

and their Applications, seminar participants at Georgia Institute of Technology, Purdue University, and University of Michigan for helpful comments. The authors also thank Michelle Marcouiller for helping with the data.

## References

- [1] Addie, R.G., Zukerman, M., & Neame, T. Fractal traffic: Measurements, modeling and performance evaluation. Proceedings of INFOCOM 95, 1995, Boston (MA), 985-992.
- [2] Ash, G. R. *Dynamic Routing in Telecommunications Networks*, McGraw-Hill, New York, 1998.
- [3] Ash, G. R., & Oberer, E. Dynamic routing in the AT&T network-improved service quality at lower cost. Proceedings in IEEE Global Telecommunications Conference, 1989, Dallas, TX.
- [4] Ash, G. R., Chen, J.S., Frey, A.E., & Huang, B.D. Real-time network routing in a dynamic class-of-service network. Proceedings of 13th Tele-traffic Congress, 1991, Copenhagen, Denmark.
- [5] Black, F., & Scholes, M. The pricing of options and corporate liabilities. *J. Political Econom.* 1973, 81, 659-683.
- [6] D'Agostino, R. B., Stephens, M. A. *Goodness-of-Fit Techniques*, Marcel Dekker, New York, 1986.

- [7] Genz, A. Numerical Computation of Bivariate and Trivariate Normal Probabilities. Preprint, 2001, Washington State University.
- [8] Gibbens, R.J. Some aspects of Dynamic Routing in Circuit-Switched Telecommunications Networks. Statistics Laboratory, 1986, University of Cambridge.
- [9] Gibbens, R.J., Hunt, P.J., & Kelly, F.P. Bistability in Communication Networks. In *Disorder in physical systems*, G. Grimmett and D. Welsh, Eds. Oxford University Press, 1990, 113-128.
- [10] Kelly, F.P. Network routing. *Phil. Trans. Roy. Soc. Ser.* 1991, A337, 343-367.
- [11] Kelly, F.P. Modelling Communication Networks, Present and Future. Clifford Paterson lecture, in *Philosophical Transactions of the Royal Society of London*, 1996, Series A354, 437-463.
- [12] Keppo, J. Pricing of point-to-point bandwidth contracts, *Mathematical Methods of Operations Research* 2005, 61, 191-218.
- [13] Mees, A. Simple is the best for dynamic routing of telecommunications. *Nature*, 1986, 323, 108.
- [14] Mitra, D., & Seery, J. B. Comparative Evaluations of Randomized and Dynamic Routing Strategies for Circuit-Switched Networks, *IEEE Trans. Communications*, 1990, 39, 102-116.

- [15] Mitra, D., Gibbens, R. J., & Huang, B. D. Analysis and Optimal Design of Aggregated-Least-Busy-Alternative Routing on Symmetric Loss Networks With Trunk Reservations, Proc. 13th International Teletraffic Congress, 1991, Copenhagen, 495-500.
- [16] Mitra, D., Morrison, J. A., & Ramakrishnan, K.G. ATM Network Design and Optimization: A Multirate Loss Network Framework, IEEE/ACM Transactions on Networking, 1996, 4, 531-543.
- [17] Mitra, D., Morrison, J. A., & Ramakrishnan, K.G. Optimization and Design of Network Routing using Refined Asymptotic Approximations, Performance Evaluation, 1999, 36, 267-288.
- [18] Norros, I. A storage model with self-similar input. Queuing Systems, 1994, 16, 387-396.
- [19] Paxson, V., & Floyd, S. Wide-Area Traffic: The Failure of Poisson Modeling. IEEE/ACM Transactions on Networking, 1995, 3, 226-244.
- [20] Ryan, S.M. Capacity Expansion for Random Exponential Demand Growth with Lead Times. Preprint, 2002, Industrial and Manufacturing Systems Engineering, Iowa State University.
- [21] Räsänen, M., Ruusunen, J., & Hämäläinen, R. Customer Level Analysis of Dynamic Pricing Experiments using Consumption Pattern Models, Energy, 1995, 20, 897-906
- [22] Räsänen, M., Ruusunen, J., & Hämäläinen, R. LoadLab - Object-Oriented Software for Electric Load Analysis and Simulations, Simulation, 1997, 13, 365-403.

- [23] Schwartz, E. The stochastic behavior of commodity prices, *Journal of Finance*, 1997, 52, 923-973.
- [24] Stacey, R. R. & Songhurst, D. J. Dynamic Alternative Routing in the British Telecom Trunk Network. *Proceedings of the International Switching Symposium*, 1987, Phoenix, Arizona.
- [25] Zhao, Y. & Kockelman, K.M. The Propagation of Uncertainty through Travel Demand Models. *Annals of Regional Science*, 2002, 36, 145-163.

### Appendix: Proof of Proposition 3.1

From Assumption 2.1 we get

$$\tilde{D}_i(T) = \tilde{D}_i(t, T) \exp\left(-\frac{1}{2}\sigma_i^2(T)(T-t) + \sigma_i(T)\sqrt{T-t}Z_i\right), \quad (\text{A.1})$$

where  $Z_i$  is the standard normal variable corresponding to the  $i$ 'th demand. Equation (3.3) gives

$$E[D_i^d(T)|F_t] = E\left[\tilde{D}_i(t)\mathbf{1}\{\tilde{D}_i(t) \leq C_i\} + C_i\mathbf{1}\{\tilde{D}_i(t) \geq C_i\}|F_t\right] \quad (\text{A.2})$$

where  $\mathbf{1}\{\tilde{D}_i(t) \geq C_i\}$  and (A.1) indicate  $Z_i \geq \frac{\log\left(\frac{C_i}{\tilde{D}_i(t, T)}\right) + \frac{1}{2}\sigma_i^2(T)(T-t)}{\sigma_i(T)\sqrt{T-t}}$ , and  $\mathbf{1}\{\tilde{D}_i(t) \leq C_i\}$

gives  $Z_i \leq \frac{\log\left(\frac{C_i}{\tilde{D}_i(t, T)}\right) + \frac{1}{2}\sigma_i^2(T)(T-t)}{\sigma_i(T)\sqrt{T-t}}$ . Combining (3.7), (A.1), and (A.2) gives

$$\begin{aligned}
E[D_i^d(T)|F_t] &= \int_{-\infty}^{d_1(t,T)} \tilde{D}_1(t,T) \exp\left(-\frac{1}{2}\sigma_1^2(T)(T-t) + \sigma_1(T)\sqrt{T-t}u\right) f(u) du + C_1 \int_{d_1(t,T)}^{\infty} f(u) du \\
E[D_i^a(T)|F_t] &= - \int_{-\infty}^{d_3(t,T)} \int_{-\infty}^{d_2(t,T)} \int_{d_1(t,T)}^{\infty} C_1 g(u_1, u_2, u_3, \rho_{1,2}, \rho_{1,3}, \rho_{2,3}) du_1 du_2 du_3 + \\
&\quad \int_{-\infty}^{d_3(t,T)} \int_{-\infty}^{d_2(t,T)} \int_{d_1(t,T)}^{\infty} \tilde{D}_1(t,T) \exp\left(-\frac{1}{2}\sigma_1^2(T)(T-t) + \sigma_1(T)\sqrt{T-t}u_1\right) g(u_1, u_2, u_3, \rho_{1,2}, \rho_{1,3}, \rho_{2,3}) du_1 du_2 du_3,
\end{aligned} \tag{A.3}$$

where  $f(u)$  is the density function of standard normal distribution,  $g(u_1, u_2, u_3, \rho_{1,2}, \rho_{1,3}, \rho_{2,3})$  is a trivariate normal density function, and the terms  $d_2(t, T)$  and  $d_3(t, T)$  are calculated according to Proposition 3.1. The term  $r_1(t, T)$  in  $d_2(t, T)$  and  $d_3(t, T)$  is solved by using Assumption 3.2. This gives

$$\begin{aligned}
r_i(t, T) &= \int_{d_1(t,T)}^{\infty} \tilde{D}_1(t, T) \exp\left(-\frac{1}{2}\sigma_1^2(T)(T-t) + \sigma_1(T)\sqrt{T-t}u\right) f(u) du - C_1 \int_{d_1(t,T)}^{\infty} f(u) du \\
&= \tilde{D}_1(t, T) N\left(-d_1(t, T) + \sigma_1(T)\sqrt{T-t}\right) - C_1 N\left(-d_1(t, T)\right).
\end{aligned} \tag{A.4}$$

From (A.3) we get

$$\begin{aligned}
E[D_i^d(T)|F_t] &= \tilde{D}_1(t, T) N\left(d_1(t, T) - \hat{\sigma}_1\sqrt{T-t}\right) + C_1 N\left(-d_1(t, T)\right) \\
E[D_i^a(T)|F_t] &= \tilde{D}_1(t, T) \cdot \\
&\quad \left[ M\left(d_2(t, T), d_3(t, T), \rho_{2,3}\right) - G\left(d_1(t, T) - \hat{\sigma}_1\sqrt{T-t}, d_2(t, T), d_3(t, T), \rho_{1,2}, \rho_{1,3}, \rho_{2,3}\right) \right] - \\
&\quad C_1 \left[ M\left(d_2(t, T), d_3(t, T), \rho_{2,3}\right) - G\left(d_1(t, T), d_2(t, T), d_3(t, T), \rho_{1,2}, \rho_{1,3}, \rho_{2,3}\right) \right],
\end{aligned} \tag{A.5}$$

where  $M(d_2, d_3, \rho_{2,3})$  is the area under a standard bivariate normal distribution function covering the region from  $-\infty$  to  $d_2$  and  $-\infty$  to  $d_3$ , and the two random variables have correlation  $\rho_{2,3}$ . Equation (A.5) gives directly the first equation of (3.8) and  $E[D_i^a(T)|F_t]$  in (A.5) can be

written in the form of the second equation of (3.8). Using the first two equations of (3.8) and equation (3.6) we get the third equation of (3.8). □