Stochastic Formulation of Scalability and Quality-of-Information Satisfiability in Wireless Networks

***Abstract*—This is where the abstract will be....essentially we're extending the ideas in the DCOSS submission to include probabilistic requirements.**

I. QoI Model

Let us define the following terms:

- W = Channel Rate (bits/second)
- T = Timeliness Requirement (seconds)
- k_{req} = Number of required images
- I_S = Size of each image (bits)
- CF = Channel Factor
- TF = Traffic Factor
- P_S = Packet size
- DF = Delay Factor
- PL = Path Length

In the DCOSS submission we derived the following equation for scalability that uses each of these defined terms:

$$W \cdot T - k_{req} \cdot I_S \cdot CF \cdot TF - P_S \cdot DF \cdot (PL - 1) \geq 0 \quad (1)$$

If we rearrange this equation, we can view the satisfiability of timeliness in terms of delay components:

$$T \geq \frac{k_{req} \cdot I_S \cdot CF \cdot TF}{W} + \frac{P_S \cdot DF \cdot (PL - 1)}{W} \quad (2)$$

In our previous analysis, we strive to determine the limits of this timeliness satisfiability by utilizing some static values and some average values where appropriate in this relation. The resulting analysis provided approximate values for QoI satisfiability and network scalability, but what if we want to expand satisfiability to a stochastic definition? And/or can we provide more accurate estimations by using more detailed models of the actual values of the parameters in the above list?

To begin answering these questions, we can look at some of these parameters in more detail and use more accurate descriptions of them by classifying them as Random Variables with appropriate probability density distributions. In this case, we start by examining a line network. Its simple structure and routing make it a nice, simple topology to use as an exploratory model. Here, the same traffic model as in the DCOSS submission is also used. In this model, each node is a source of a query that is delivered to a randomly chosen destination.

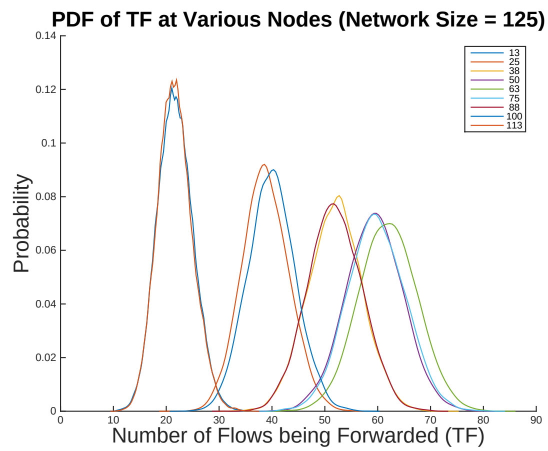


Fig. 1. Plotting frequency of experienced Traffic Factors for different node positions in a line network shows that TF is best modeled by a Normal Random Variable.

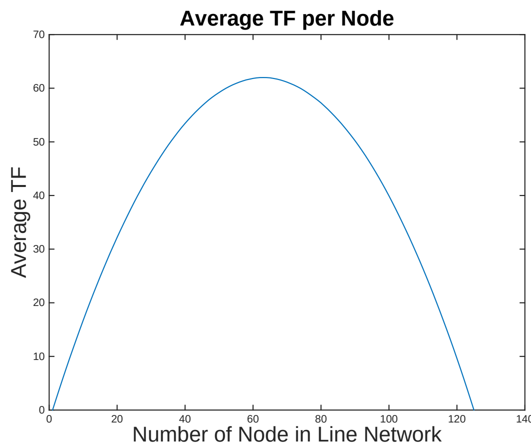
A. Traffic Factor

In order to help with modeling the parameters contributing to delay in satisfying queries, we first examine the results of statistics for Traffic Factor from simple simulations implementing the traffic model. Here, we simply simulated a large number of trials in which destinations were chosen at random for each source and the actual traffic factor and path length statistics were recorded for each node and each trial.

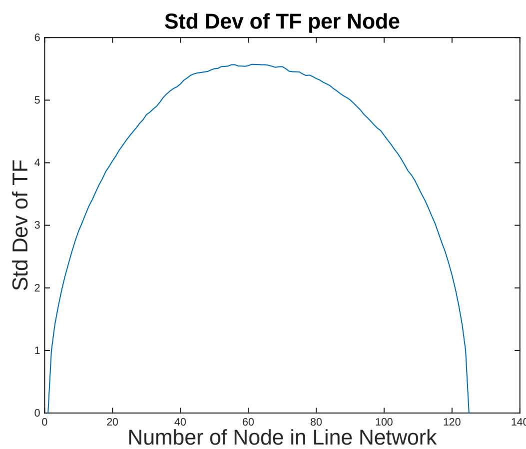
Figure 1 shows the empirical PDF of the TF for a sample of the nodes along the line network. Clearly, the distribution of the TF at each node follows a Normal distribution with varying mean and standard deviation values. This result is expected and should follow as an application of the Central Limit Theorem (need to define RVs appropriately to make this point more rigorous). We plot these mean and standard deviation values, also empirically found, for each node in the line network in Figure 2. Using Matlab's curve-fitting tool, we can give an approximate distribution for the Traffic Factor at node x, i.e. in x_{th} position from one end, in the network:

$$f_{TF_x} = \mathcal{N}(-0.016x^2 + 2x - 2, -0.0011x^2 + 0.138x + 1.42) \quad (3)$$

Let's call these two functions $\mu_{TF}(x)$ and $\sigma_{TF}(x)$. For a flow originating at node i, we want the PDF of the Traffic Factor. Let's call the destination of the flow j. Then, let's use



(a) The average Traffic Factor for each node of a line network.



(b) The standard deviation of the Traffic Factor for each node of a line network.

Fig. 2. We can observe the empirical statistical properties of the Traffic Factor for each node's position in a line network (here with 125 nodes).

P^i_{TF} to represent the PDF of the Traffic Factor for this flow originating at node i. We need to first find the node that has the largest expected TF, so we need to find the node x' that has the maximum expected TF:

$$x' = \underset{x=[\min(i,j),\max(i,j)]}{\arg\max} \mu_{TF}(x) \tag{4}$$

Then, we can say that the distribution of the TF for this flow would be

$$f^i_{TF}(tf) = \mathcal{N}(\mu(x'), \sigma(x')) \tag{5}$$

Let's call the randomly chosen destination of the flow j. If $i < N/2$, since the maximum of μ_{TF} is at $N/2$, the value of x' is given by:

$$x' = \begin{cases} i & j < i \\ j & i < j < \frac{N}{2} \\ \frac{N}{2} & \frac{N}{2} \leq j \leq N \\ 0 & \text{o.w.} \end{cases}$$

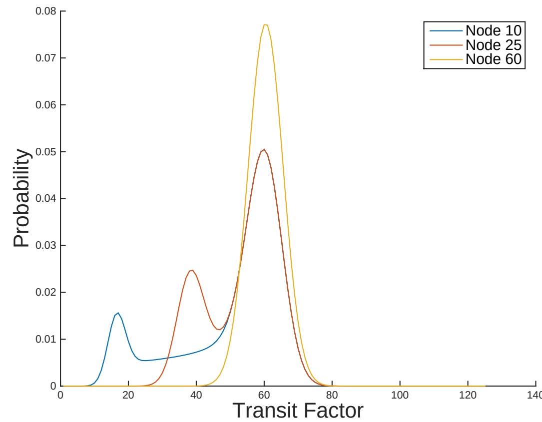


Fig. 3. PDF of Traffic Factor for flows originating in Nodes 10, 25, and 60 in a 125 node line network.

Since j is given by a uniform random variable, the probability distribution of x' for a flow originating in node i would be

$$f^i_{X'}(x') = \begin{cases} \frac{i}{N} & x' = i \\ \frac{1}{2} - \frac{i}{N} & i < x' < \frac{N}{2} \\ \frac{1}{2} & x' = \frac{N}{2} \\ 0 & \text{o.w.} \end{cases} \tag{6}$$

Then, the distribution of the PDF for the Traffic Factor of a flow originating in node i is given by Equation (5), where x' is first sampled from the distribution in Equation (6). This distribution can be fully described with a mixture distribution as follows:

$$\begin{aligned} f^i_{TF}(tf) = \ & \frac{i}{N} \cdot \mathcal{N}(\mu(i), \sigma(i)) \\ & + \sum_{k=i}^{\frac{N}{2}-1} \cdot \frac{\frac{1}{2} - \frac{i}{N}}{\frac{N}{2} - i} \mathcal{N}(\mu(k), \sigma(k)) \\ & + \frac{1}{2} \cdot \mathcal{N}(\mu(\frac{N}{2}), \sigma(\frac{N}{2})) \end{aligned} \tag{7}$$

Figure 3 shows the distribution in Equation 14 for several chosen nodes in a line network. Figure 4 displays the expected value of TF for flows originating in nodes 1 to 62 in a 125 node line network.

B. Path Length

Next, we can capture the distribution of the path length given by flows originating in node i of the network. Since our traffic model is to choose destination nodes with uniform randomness, we can derive the distribution of path lengths for flows with source node i as follows (details can be given if needed). Again, let us just assume that i is less than $N/2$, since we can use symmetry to draw the same conclusion about nodes greater than $N/2$.

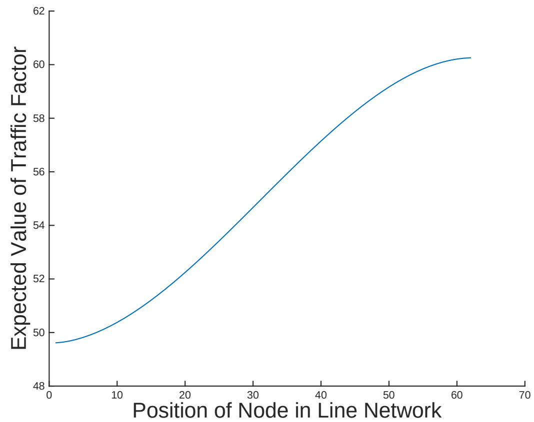


Fig. 4. Expected Value of Traffic Factor for flows originating in Nodes 5-62 in a 125 node line network.

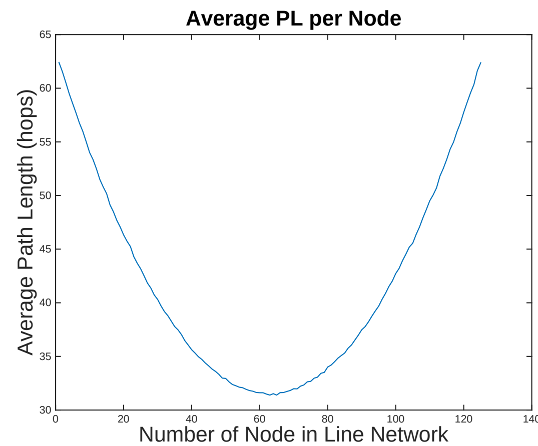


Fig. 6. The average value of path length intuitively peaks at the edges of the line network and is minimum in the middle.

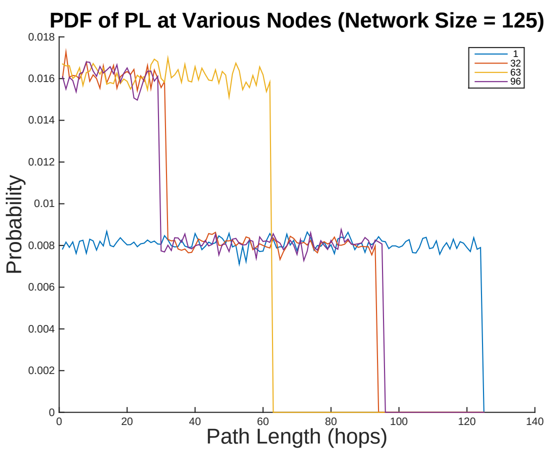


Fig. 5. Plotting the frequency of experienced Path Lengths for different node positions in a line network shows that PL can be modeled as disjoint Uniform Random Variables.

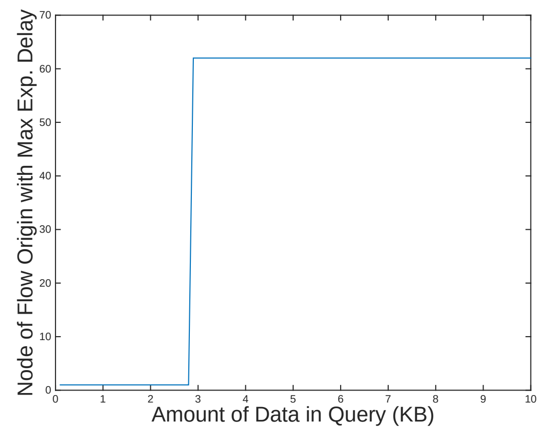


Fig. 7. The value of i (origin of a flow) that causes the maximum expected delay.

$$f^i_{PL}(pl) \quad = \quad \begin{cases} \frac{2}{N} & \text{for } 1 < pl < (i-1) \\ \frac{1}{N} & \text{for } i \leq pl \leq N-i \\ 0 & \text{elsewhere} \end{cases}$$

While the expected value for the path length is not derived here yet, its empirical value for nodes in each position are shown in Figure 6. Not surprisingly, values of mean path length range from $N/2$ at the end of the network to $N/2$ in the middle of the network.

II. Finding Bottleneck Flow

Again, we turn to the satisfiability equation, but use random variables for the Traffic Factor and Path Length, TF_i and PL_i, respectively:

$$T \geq \frac{k_{req} \cdot I_S \cdot CF \cdot TF_i}{W} + \frac{P_S \cdot DF \cdot (PL_i - 1)}{W} \tag{8}$$

where we will call the total delay

$$D_i = \frac{k_{req} \cdot I_S \cdot CF \cdot TF_i}{W} + \frac{P_S \cdot DF \cdot (PL_i - 1)}{W} \tag{9}$$

Now, we want to identify which flow i, identified by its origin node, is most likely to be the flow that cannot be satisfied in the allotted timeliness. To do so, we can simply find the value of i that results in the largest D_i.

Using the expected values for TF_i and PL_i from Figures 4 and 6, we find the value of i that maximizes D_i for different data requirements, $B = k_{req} * I_S$. Figure 7 shows that for low data requirements, the delay of multi-hop paths dominates, causing the "Bottleneck" flow to be those that originate in node 1 and have a larger expected path length. At a point, though, as the amount of data required in the query grows, congestion will be the limiting factor in the network, making the Traffic Factor more important. Thus, the node near the center of the network which will be likely to experience the highest amount of congestion become the source of flows with the highest delay. In this case, we have $i = 62$, since the network has 125 nodes (NOTE: It should probably be 63, but there is an "off-by-one" error here because the definitions/formulations are not carefully derived to be correct on the boundaries).