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 conference 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
- $p_{i,j}$ = Probability of a flow from i to j

In the conference 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 conference submission is also used. In this model, each node is a source of a query that is delivered to a randomly chosen destination.

A. Traffic Factor

Let $\rho(x)$ be the number of shortest paths of all other nodes that include node x (NOTE: Is this the same as Betweeness Centrality?). Let F_{ij} be represent the existence of a flow existing from node i to node j. We assume that F_{ij} is equal to 1 with probability p_f and 0 otherwise. Then, the traffic factor of a node x, R_x, is given by the sum of F_{ij} for all $\rho(x)$ pairs (i, j) in which x is along the shortest path. Assuming that F_{ij} is i.i.d. for all pairs (i, j), then R_x can be approximated by a Normal RV with mean $\rho(x)p_f$ and variance $\rho(x)p_f(1 - p_f)$:

$$f_{R_x} = \mathcal{N}(\rho(x)p_f, \rho(x)p_f(1 - p_f)) \quad (3)$$

Let's consider a flow originating at node i, and call the destination of the flow j. When characterizing the largest contributor to delay, we need to determine the maximum expected Traffic Factor through which the flow will be forwarded. We will use TF_i to represent that maximum expected Traffic Factor for a flow with origin i. To get a distribution for overall delay, we want to derive a distribution for this value. We will use P_{TF_i} 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 between node i and node j, so we need to find the node x' that has the maximum expected TF. Since p_f is constant, the node with the maximum expected TF is:

$$x' = \underset{x=[\min(i,j),\max(i,j)]}{\arg\max} \rho(x) \quad (4)$$

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

$$f_{TF_i}(tf) = \mathcal{N}(\rho(x')p_f, \rho(x')p_f(1 - p_f)) \quad (5)$$

Given a network graph, values of $\rho(x)$ and the solution of (4) could be analytically determined for that specific case. In networks with regular structure, though, we may be able to develop closed form expressions for

B. Path Length

Next, we can capture the distribution of the path length given by flows originating in node i of the network. Once again, this distribution can be determined for any network given the topology with some simple computation, but we can provide expressions for some regular networks.

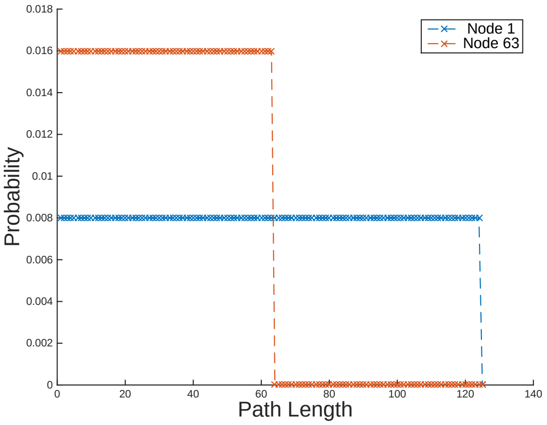

Fig. 1. PDF of Path Lengths for flows originating in edge cases of a line network.

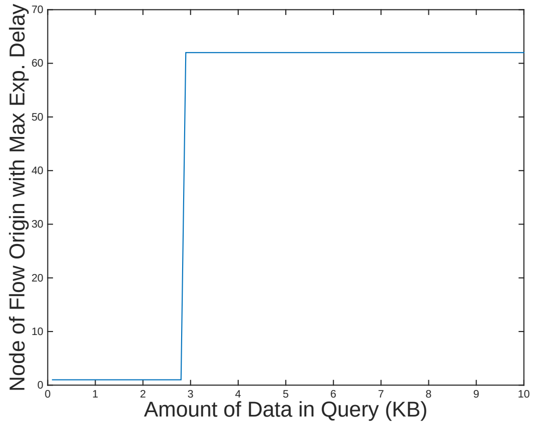

Fig. 2. The value of i (origin of a flow) that causes the maximum expected delay.

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:

$$T \geq \frac{k_{req} \cdot I_S \cdot CF \cdot TF_i}{W} + \frac{P_S \cdot DF \cdot (PL_i - 1)}{W} \tag{6}$$

where we will call the total delay

$$D_i = \frac{k_{req} \cdot I_S \cdot CF \cdot TF_i}{W} + \frac{P_S \cdot DF \cdot (PL_i - 1)}{W} \tag{7}$$

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 8 and 10, we find the value of i that maximizes D_i for different data requirements, $B = k_{req} * I_S$. Figure 2 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).

III. Probability of Satisfying Query

Now, we have an expression for delay that is built up using random variables for the values of TF and PL. The next useful formulation is providing an expression that describes the probability of a bottleneck flow satisfying its timeliness requirement considering the underlying distributions. With this formulation, we can define scalability as a network in which the flow's probability of satisfiability exceeds some threshold P_{thresh}:

$$P(D_i < T) \geq P_{thresh} \tag{8}$$

Once again, we have our total delay equation of a flow originating in node i:

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

Let us define two constants to simplify the expression:

$$C_1 = \frac{k_{req} \cdot I_S \cdot CF}{W}$$
$$C_2 = \frac{P_S \cdot DF}{W}$$

Then, we can express the delay as

$$D_i = C_1 \cdot TF_i + C_2 \cdot PL_i \tag{10}$$

The distribution of D_i is determined by the convolution of TF_i and PL_i:

$$f_{D_i}(d) = \sum_{pl=1}^{N-1} f_{PL_i}(pl) \cdot f_{TF_i}(\frac{d - C_2 \cdot pl}{C_1}) \tag{11}$$

or

$$f_{D_i}(d) = \sum_{tf=1}^{N} f_{PL_i}(\frac{d - C_1 \cdot tf}{C_2}) \cdot f_{TF_i}(tf) \tag{12}$$

The cumulative distribution function of the delay, which is the ultimate goal to provide an expression for Equation (8), then, is

$$F_{D_i}(d) = \sum_{pl=1}^{N-1} f_{PL_i}(pl) \cdot F_{TF_i}(\frac{d - C_2 \cdot pl}{C_1}) \tag{13}$$

Using the definitions in Equations (11) and (13), we can visualize the distribution of delays of flows in the line network. Figure 3 shows the PDF of delays for flows of size $1KB$ that originate in Node 1 and in Node 63 (the center node in the network), and Figure 4 shows the CDF for the same network. As we established in Section II, for this small data size, the

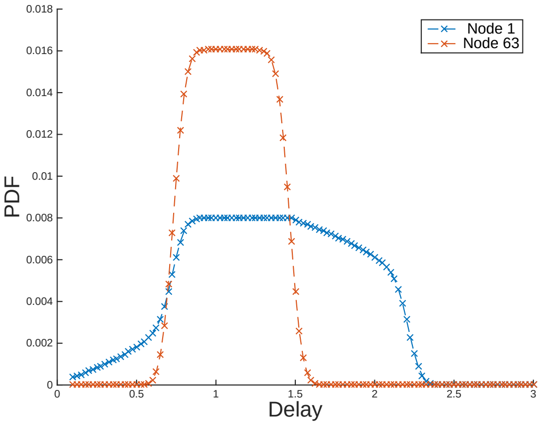

Fig. 3. The PDF of delay for 1 KB flows originating in nodes 1 and 63 in a 125 node line network.

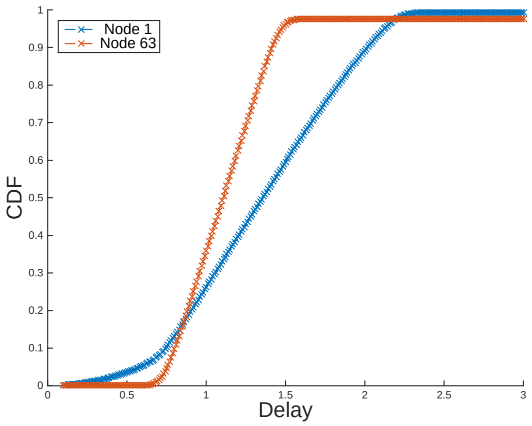

Fig. 4. The CDF of delay for 1 KB flows originating in nodes 1 and 63 in a 125 node line network.

bottleneck flow is likely to occur from the edge of the network. This finding is validated by the distributions for f_{D_1} and F_{D_1} extending to higher delays than those of $f_{D_{63}}$ and $F_{D_{63}}$.

Figures 5 and 6 provide the distributions of flows originating from nodes on the edge and center of the network as well, except for an expected data requirement of $18KB$. Here, we see the ordering of the delay distributions is reversed, also as expected from the findings in Section II.

IV. Example Applications

A. Line Network Example

1) Traffic Factor: First, we can provide an example of developing a distribution for TF_i for a line network.

Let's examine an arbitrary flow from source node i and destination as j. We will construct a distribution for TF_i assuming that $i < \frac{N}{2}$. Since the network is symmetrical, the derivation holds for all nodes. For a node x in the network, the probability that the flow from i to j goes through x is $\frac{2(x-1)(N-x)}{N(N-1)}$. Considering the network has N concurrent flows, the number of expected paths going through x, then, is:

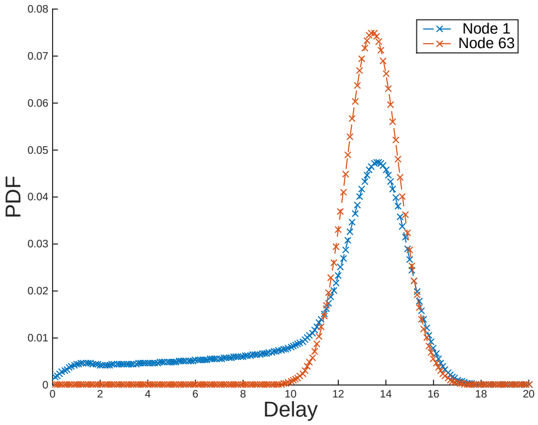

Fig. 5. The PDF of delay for 18 KB flows originating in nodes 1 and 63 in a 125 node line network.

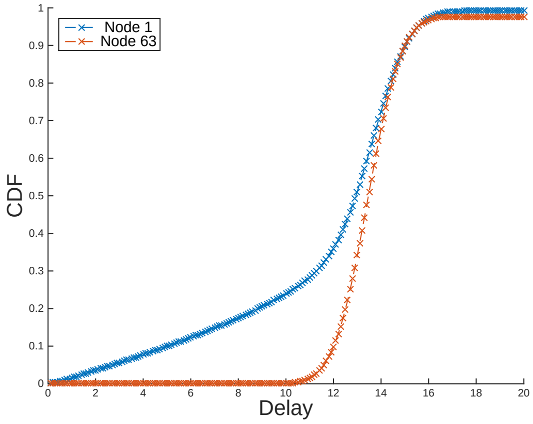

Fig. 6. The CDF of delay for 18 KB flows originating in nodes 1 and 63 in a 125 node line network.

$$\rho(x) = \frac{2(x-1)(N-x)}{N-1}$$

x is a binomial variable with $n = 2(x-1)(N-x)$ and $p = \frac{1}{N-1}$, so it can be approximated with a Gaussian with mean $\frac{2(x-1)(N-x)}{N-1}$ and variance $\frac{2(x-1)(N-x)}{N-1}(1 - \frac{1}{N-1})$ (or maybe $\frac{1}{N}$?)

It is easy to show that $\rho(x)$ is increasing in the domain $[1, N/2)$ with the maximum at $N/2$. Therefore, the value of x' is given by:

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

Assuming that all values of j are equally likely, which holds since p_f is equal for all flows, the probability distribution of x' for a flow originating in node i would be

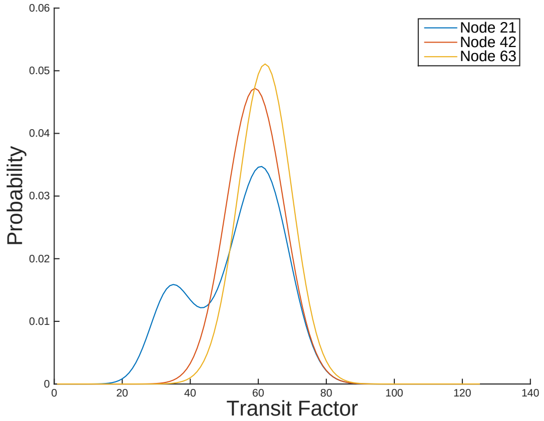

Fig. 7. PDF of Traffic Factor for flows originating in several different nodes in a 125 node line network.

$$f_{X_i'}(x') \quad = \quad \begin{cases} \frac{i}{N} & x' = i \\ \frac{1}{2} - \frac{i}{N} & i < x' < \frac{N}{2} \\ \frac{1}{2} & x' = \frac{N}{2} \\ 0 & \text{o.w.} \end{cases} \tag{14}$$

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 (14). This distribution can be fully described with a mixture distribution as follows:

$$\begin{aligned} f_{TF}^{i}(tf) = \quad & \frac{i}{N} \cdot \mathcal{N}(\rho(i)p_f, \sqrt{\rho(i)p_f(1-p_f)}) \\ + \quad & \sum_{k=i}^{\frac{N}{2}-1} \cdot \frac{\frac{1}{2} - \frac{i}{N}}{\frac{N}{2} - i} \mathcal{N}(\rho(k)p_f, \sqrt{\rho(k)p_f(1-p_f)}) \\ + \quad & \frac{1}{2} \cdot \mathcal{N}(\rho(\frac{N}{2})p_f, \sqrt{\rho(\frac{N}{2})p_f(1-p_f)}) \end{aligned} \tag{15}$$

If the expected traffic is for each node to be the source of 1 flow at a time, on average, then we can substitute a value of $p_f = 1/(N-1)$:

Figure 7 shows the distribution in Equation 15 for several chosen nodes in a line network. Figure 8 displays the expected value of TF for flows originating in nodes 1 to 62 in a 125 node line network.

2) Path Length: Next, we can capture the distribution of the path length given by flows originating in node i of a line 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 $N/2$. The following is the distribution, which is shown in Figure 9 for the edge cases of $i = 1$ and $i = 63$, the nodes on the end and in the middle of the line network, respectively.

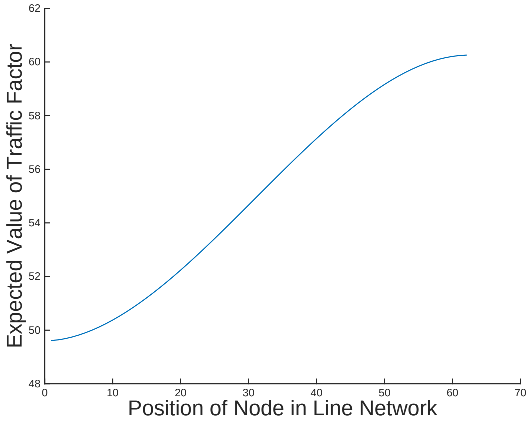

Fig. 8. The expected value of the Traffic Factor is naturally highest in the middle of the line network where more congestion exists.

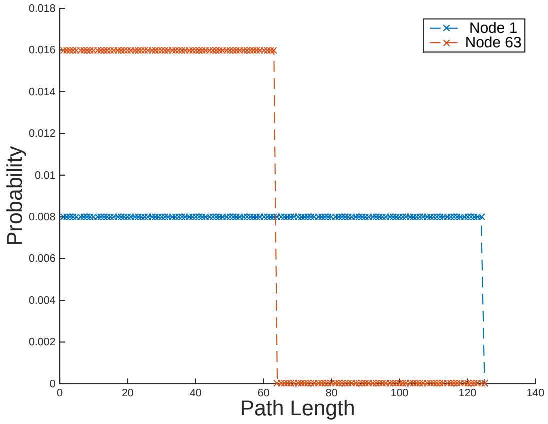

Fig. 9. PDF of Path Lengths for flows originating in edge cases of a line network.

$$f_{PL_i}(pl) \quad = \quad \begin{cases} \frac{2}{N} & \text{for } 1 \leq pl \leq i-1 \\ \frac{1}{N} & \text{for } i \leq pl \leq N-i \\ 0 & \text{elsewhere} \end{cases} \tag{17}$$

From (17), we can derive the following expression for the expected value of a path length for a flow originating in node i.

$$E[PL_i] = \frac{N}{2} + \frac{i^2}{N} - \frac{2 \cdot i}{N} - i \tag{18}$$

These expected values for path length of flows originating in different nodes are shown in Figure 10. Not surprisingly, values of mean path length range from $N/2$ at the end of the network to $N/4$ in the middle of the network.

$$\begin{aligned} f_{TF}^{i}(tf) = \quad & \frac{i}{N} \cdot \mathcal{N}(\frac{2(i-1)(N-i)}{(N-2)(N-1)}, \sqrt{\frac{2(i-1)(N-i)(1-\frac{1}{N-1})}{(N-2)(N-1)}}) \\ + \quad & \sum_{k=i}^{\frac{N}{2}-1} \cdot \frac{\frac{1}{2}-\frac{i}{N}}{\frac{N}{2}-i} \mathcal{N}(\frac{2(k-1)(N-k)}{(N-2)(N-1)}, \sqrt{\frac{2(k-1)(N-k)}{(N-2)(N-1)}(1-\frac{1}{N-1})}) \\ + \quad & \frac{1}{2} \cdot \mathcal{N}(\frac{N(\frac{N}{2}-1)}{(N-2)(N-1)}, \sqrt{\frac{N(\frac{N}{2}-1)}{(N-2)(N-1)}(1-\frac{1}{N-1})}) \end{aligned} \tag{16}$$

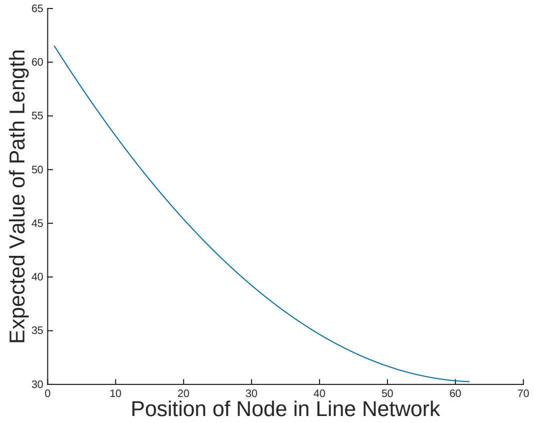

Fig. 10. The expected value of path length intuitively peaks at the edges of the line network and is minimum in the middle.