25th Annual Workshop on
Mathematical Problems in Industry
University of Delaware, June 15–19, 2009

Examination of optimizing information flow in networks

"The Dynamics of a model of a computer protocol" problem presented by
Dr. Fern Y. Hunt
Information Technology Laboratory, MS 8910
National Institute of Standards and Technology
Gaithersburg Maryland

Participants:

Deepak Aralumallige	Michael Levinson	Yi Wang
Richard Braun	Qingxia Li	Romann Weber
Aaron Churchill	Colin Please	Tom Witelski
Garry Halliwell	Ivonne Rivas	

Summary Presentation given by A. Churchill (6/19/09)
Summary Report compiled by T. Witelski

1 Introduction

The central role of the Internet and the World-Wide-Web in global communications has refocused much attention on problems involving optimizing information flow through networks. The most basic formulation of the question is called the "max flow" optimization problem: given a set of channels with prescribed capacities that connect a set of nodes in a network, how should the materials or information be distributed among the various routes to maximize the total flow rate from the source to the destination [7]. This model applies to contexts from computer and telephone networks to shipping and distribution of commercial goods to electrical power and water supply systems. Theory in linear programming has been well developed to solve the classic max flow problem. Modern contexts have demanded the examination of more complicated variations of the max flow problem to take new factors or constraints into consideration; these changes lead to more difficult problems where linear programming is insufficient. In the workshop we examined models for information flow on networks that considered trade-offs between the overall network utility (or flow rate) and path diversity to ensure balanced usage of all parts of the network (and to ensure stability and robustness against local disruptions in parts of the network) [1].

While the linear programming solution of the basic max flow problem cannot handle the current problem, the approaches primal/dual formulation for describing the constrained optimization problem [7] can be applied to the current generation of problems, called *network utility maximization* (NUM) problems. In particular, primal/dual formulations have been used extensively in studies of such networks, see for example [1, 3–5, 8].

In extending the optimization problem, modeling, theoretical consideration and interpretation must be given to the network properties being optimized. In [1], "path diversity" is described as an important consideration to maintain network robustness and to avoid some instabilities due to "route flapping" in path allocations; an "entropy" parameter is used to weight the importance of this factor. Other articles have also introduced concepts of "fairness" in allocating network resources [2, 9, 10]. Game theory and economic interpretations of network optimization also lead to important insights [2, 8]. A key feature of the traffic-routing model we are considering is its formulation as an economic system, governed by principles of supply and demand. Since we are generally dealing with channels of unequal capacity, a consumer-centered intuition might suggest that the higher-capacity channels would command the

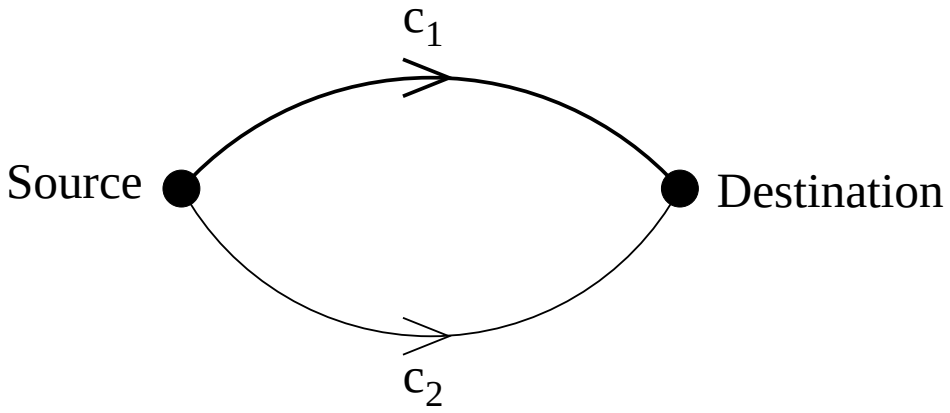

Figure 1: The two-link model network. Each channel or network link is labeled by its capacity, c_i, representing a bandwidth or maximum flowrate.

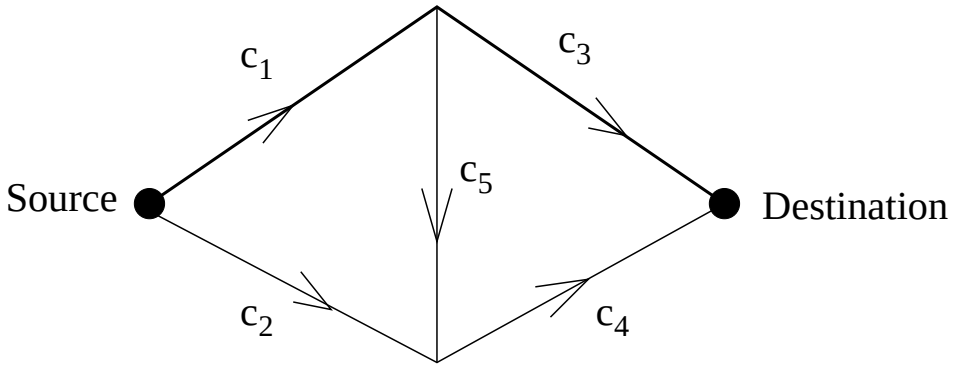

Figure 2: The diamond model network. Here are are three routes from the source to the destination.

higher prices, since they would represent the premium option among the available channels. But the model in [1, 3] describes a control algorithm that uses *price* (taking the form of a dual variable) as a *disincentive* in order to adjust channel traffic toward an optimal distribution [2, 8]. Such an optimal distribution will be made with the channel capacities in mind, so that the portion of the total traffic routed to a particular channel is directly proportional to its capacity. Considering channel capacities as a commodity of limited supply, we might suspect that a system that regulates traffic via a pricing scheme would assign prices to channels in a manner *inversely* proportional to their respective capacities.

Once an appropriate network optimization problem has been formulated, it remains to solve the optimization problem; this will need to be done numerically, but the process can greatly benefit from simplifications and reductions that follow from analysis of the problem. Ideally the form of the numerical solution scheme can give insight on the design of a distributed algorithm for a Transmission Control Protocol (TCP) that can be directly implemented on the network.

At the workshop we considered the optimization problems for two small prototype network topologies: the two-link network and the diamond network. These examples are small enough to be tractable during the workshop, but retain some of the key features relevant to larger networks (competing routes with different capacities from the source to the destination, and routes with overlapping channels, respectively). We have studied a gradient descent method for solving obtaining the optimal solution via the dual problem [1, 3]. The numerical method was implemented in MATLAB and further analysis of the dual problem and properties of the gradient method were carried out. Another thrust of the group's work was in direct simulations of information flow in these small networks via Monte Carlo simulations as a means of directly testing the efficiencies of various allocation strategies.

2 Problem formulation

The basic formulation of the network is a directed graph with a set of nodes (senders/receivers) connected by edges. Each edge will have a given capacity c_i interpreted as a bandwidth or maximum flow rate. The two networks that we focus on are the two-link topology in Figure 1 and the diamond topology in Figure 2. In the two-link model, there are two competing routes from the source to the destination, via channel 1 or via channel 2. In the diamond network, there are five network link $i = 1, 2, \cdots, 5$, but there are only three routes, $j = 1, 2, 3$: (1) via channel 1 to channel 3, or (2) via channel 1 to channel 5 to channel 4, or (3) via channel 2 to channel 4.

The equilibrium optimization problem is to determine probabilities β_j for partitioning a steady flow of information among the different routes. In terms of the diamond network, this would consist of $\beta_1, \beta_2, \beta_3$ that would determine the complete routing through the network of a message, as determined at the source. In practice, once these are determined, they can be translated into probabilities for local routing decisions, namely at the source $\tilde{\beta}_1 = (\beta_1 + \beta_2)$ to use channel c_1 vs. $\tilde{\beta}_2 = \beta_3$ to use channel c_2 and then at the intermediate $c_{1/3/5}$ junction, probability $\tilde{\beta}_3 = \beta_1/(\beta_1 + \beta_2)$ to use channel 3 vs. probability $\tilde{\beta}_5 = \beta_2/(\beta_1 + \beta_2)$ to use channel 5.

There are two sets of fundamental constraints common to all versions (linear and nonlinear programming) of the network optimization problem [7]:

1. Probability constraints: Each β_j must be a valid, non-negative probability,

$$\beta_j \geq 0 \qquad \forall j \tag{2.1}$$

and the probabilities for all network routes must sum to one,

$$\sum_j \beta_j = 1 \tag{2.2}$$

2. Capacity constraints: The maximum flow through each channel is strictly limited by the channel capacity,

$$0 \leq f_i \leq c_i, \qquad \forall i \tag{2.3}$$

If the total flow through the network is x (and inflow equals outflow given the assumption that no data is lost in the network), then the channel flows can be related to sums of the routing fractions $\beta_j x$.

The objective function to be maximized is generally called the network utility, $U(x)$; from theoretical considerations for proving optimality of solutions, this should be a strictly increasing concave function [3]. Violating this slightly, for some sections of this report, we consider the simplest case of maximizing the flow, with $U(x) = x$; this is a special case of the Reno utility function, $U(x) = (x/w)^{\alpha-1}/(\alpha - 1)$ for $\alpha = 2$ and $w = 1$.

To incorporate considerations of path diversity and network robustness, we add an additional constraint:

3. Entropy constraint: A global scalar property (motivated by information theory) of the routing strategy is bounded from below by a prescribed value H_s,

$$-\sum_j \beta_j \log \beta_j \geq H_s. \tag{2.4}$$

If there are J different routes through the network, then the strategy with the largest entropy would have all routes be equally-likely, i.e. $\beta_j = 1/J$ for all j. This sets an upper bound on the values for H_s for which equilibrium solutions are possible,

$$0 \leq H_s \leq H_s^{\max}(J) \equiv \log(J). \tag{2.5}$$

For some dynamic situations, there may be advantages to having an evolving value for H_s, but as described in [1], for the equilibrium problem (2.4) can be reduced to the case of the equality constraint, which we will use for the rest of this report.

Determining β_j to maximize $U(x)$ subject to the constraints (2.1, 2.2, 2.3, 2.4) is a statement of the primal constrained optimization problem for a network. A lot of our further work will focus on making use of the *dual problem* as a means for obtaining the solution. The dual problem is obtained from the primal problem via analysis in terms of Lagrange multipliers (which play the role of the dual variables).

We now focus on specific results for the two networks that we examined.

3 The two-link network and the gradient projection algorithm

The equilibrium optimization problem for the two-link network in Figure 1 for given values of c_1, c_2, H_s is:

$$\underline{\text{objective function}}: \qquad \max_{\beta_1,\beta_2} U(x) \tag{3.1a}$$

$$\underline{\text{probability constraints}}: \qquad \begin{cases} \beta_1 + \beta_2 = 1 \\ \beta_1 \geq 0 \\ \beta_2 \geq 0 \end{cases} \tag{3.1b}$$

$$\underline{\text{capacity constraints}}: \qquad \begin{cases} \beta_1 x \leq c_1 \\ \beta_2 x \leq c_2 \end{cases} \tag{3.1c}$$

$$\underline{\text{entropy constraint}}: \qquad -\beta_1 \log \beta_1 - \beta_2 \log \beta_2 = H_s \tag{3.1d}$$

The solution to this problem can be obtained using a gradient projection algorithm developed by Low and Lapsley [3] and used by Hunt and Marbukh [1]. This is an iterative scheme, with iteration index $k = 1, 2, \cdots$, that will converge to the values of the probabilities β_i for $k \to \infty$ making use of dual variables labeled as prices p_i (which enforce the capacity constraints) and γ (which enforces the entropy constraint). The iteration equations for the prices associated with each of the channels are

$$p_1(k+1) = [p_1(k) - h\{c_1 - \beta_1(k)x(k)\}]^+, \tag{3.2a}$$

$$p_2(k+1) = [p_2(k) - h\{c_2 - \beta_2(k)x(k)\}]^+, \tag{3.2b}$$

where $h > 0$ is a positive numerical stepsize and the ramp function is defined by $[f(x)]^+ = f(x)$ if $f(x) > 0$ and 0 otherwise. The prices will be steady if the capacity constraints hold exactly, $c_i = \beta_i x$, otherwise they will evolve and stay truncated to non-negative values, $p_i \geq 0$. These equations follow from Karush-Kuhn-Tucker (KKT) conditions [6] for the dual variables associated with the capacity constraints.

Given the prices and γ, the probabilities are expressed as

$$\beta_i(k) = \frac{e^{-\gamma(k)p_i(k)}}{Z(k)}, \tag{3.3}$$

where the normalization factor is

$$Z(k) = e^{-\gamma(k)p_1(k)} + e^{-\gamma(k)p_2(k)}. \tag{3.4}$$

Here the exponential form ensures that constraint (2.1) is satisfies for all possible choices of γ, p_i, while the Z normalization ensures that constraint (2.2) is automatically satisfied.

It remains to satisfy the entropy condition (3.1d); if the $p_i(k)$ are assumed, then this yields a nonlinear implicit equation for $\gamma(k)$,

$$\gamma\bar{p} + \log Z = H_s, \tag{3.5}$$

where $\bar{p}$ is defined by

$$\bar{p}(k) = \beta_1(k)p_1(k) + \beta_2(k)p_2(k). \tag{3.6}$$

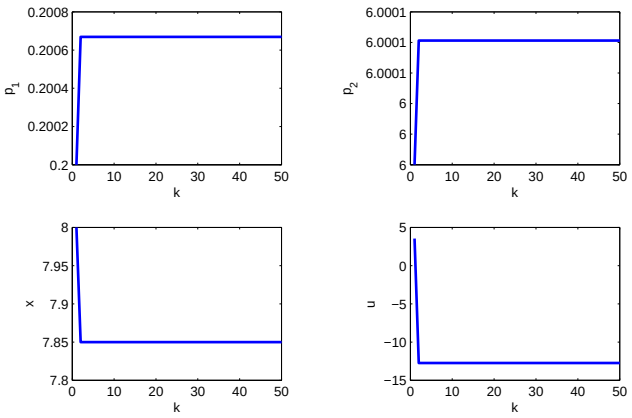

Figure 3: Computed evolution of $p_1(k)$ (upper left), $p_2(k)$ (upper right), $x(k)$ (lower right) and $u(k)$ (lower right) at $H_s = H_{\text{ref}}$. A steady state is rapidly obtained.

In practice, on each k step (3.5) would be solved first for $\gamma(k)$, then the flow can be updated via

$$x(k) = \min\left\{\sqrt{\frac{w}{\bar{p}(k)}}, x_{\max}\right\}, \tag{3.7}$$

and then (3.2ab) can be evaluated to obtain the next iteration of the prices, $p_i(k+1)$. Finally, the updated utility is found for $U(x) = -w/x$

$$U(k) = -\frac{w}{x(k)}. \tag{3.8}$$

Note that $U(x)$ is a monotone increasing function of x. System (3.2ab) is effectively an explicit Forward Euler numerical method for the gradient projection scheme, subject to the algebraic constraint (3.5).

Note that the specific properties of the network infrastructure (i.e. the channel capacities, c_1, c_2) appear only in (3.2ab), not explicitly in any of the other equations in the algorithm.

3.1 Gradient Projection Algorithm Computations for Perturbations to H_{ref}

In this section, we test the gradient projection algorithm on the two-link network for with a value of entropy H_s near the ideal value H_{ref} corresponding to the max-flow solution on the network. Consider a network with prescribed channel capacities $c_1 = 7$, $c_2 = 0.85$. In this case, the optimal solution for the routing probabilities of the reduced problem (3.1a,b,c) is

$$\beta_1 = \frac{c_1}{c_1 + c_2} \approx 0.8917 \qquad \beta_2 = \frac{c_2}{c_1 + c_2} \approx 0.1083. \tag{3.9}$$

These values correspond to

$$H_{\text{ref}} = -\left[\frac{c_1}{c_1 + c_2}\log\left(\frac{c_1}{c_1 + c_2}\right) + \frac{c_2}{c_1 + c_2}\log\left(\frac{c_2}{c_1 + c_2}\right)\right] \approx 0.3429 \tag{3.10}$$

We computed the gradient projection algorithm with parameters $w = 100$, $h = 0.005$, $x_{\max} = c_1 + c_2$, together with initial conditions $p_1(0) = 0.1$, $p_2 = 6$, $x(0) = 8$. $N = 11200$ iterations were used.

We first show results at $H_{\text{ref}} \approx 0.3429$ in Figure 3. It is clear that the steady states for the prices, transmission rate and utility are all obtained rapidly, after a small number of iterations.

We now detune the entropy to $H_s = H_{\text{ref}} + 0.02 \approx 0.3629$. Typical results are shown in Figure 4. A solution is relatively easily found, but as expected, there is no steady state. Here p_1 decreases to nullity in a finite number of steps, and p_2 increases (apparently) linearly after an initial transient. This is due to the choice of $x_{\max} = c_1 + c_2$. Choosing $x_{\max} > \frac{c_1}{\beta_1}$ leads to an increasing p_2 that converges to a limit

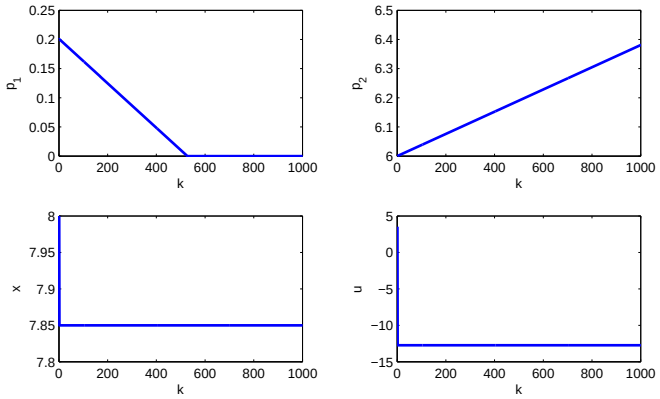

Figure 4: Computed evolution of $p_1(k)$ (upper left), $p_2(k)$ (upper right), $x(k)$ (lower right) and $u(k)$ (lower right) at $H_s = H_{\text{ref}} + 0.02$. No steady state is obtained; p_1 decreases to 0 and p_2 continues to increase.

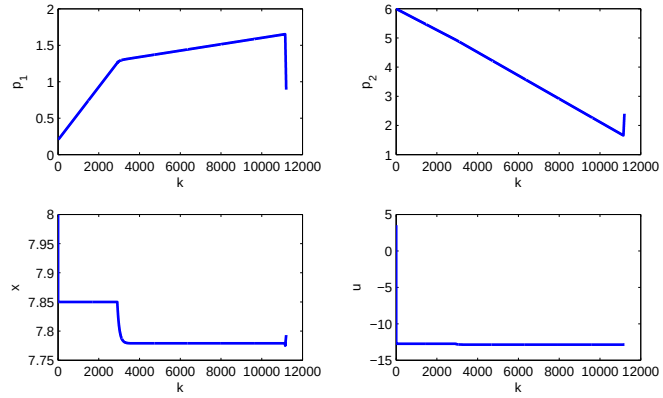

Figure 5: Computed evolution of $p_1(k)$ (upper left), $p_2(k)$ (upper right), $x(k)$ (lower right) and $u(k)$ (lower right) at $H_s = H_{\text{ref}} - 0.02$. No steady state is obtained; p_1 and p_2 merge at the beginning of the sudden change in these quantities. The computation is invalid once the p_i merge.

(as seen in Figure 2 in [1]). When $p_1 = 0$, the transmission rate appears to be unaffected over this range of calculation.

If we compute below H_{ref}, then the behavior changes. Results for $H_s = H_{\text{ref}} - 0.02 \approx 0.3229$ are shown in Figure 5. In this calculation, there is again no steady state. p_1 and p_2 approach each other, in this case with p_2 decreasing, until they become equal at about $k = 11100$. When they become equal, the positive root of $G(\gamma)$ becomes infinite. Thus, the part of the computation where the p_i change suddenly is the end of the believable part of the computation, and it should be ended there.

Thus, for computations at $H_s < H_{\text{ref}}$, there is a finite time for which the model is valid, which is before the merging of the p_i. For $H_s = H_{\text{ref}}$, the steady state can be computed. For $H_s > H_{\text{ref}}$, the solution can be computed but there is no steady state.

3.1.1 Near the "Don't Care" limit

For the two-link network the maximum entropy, $H_s^{\max} = \log(2)$, is attained when both links have the same capacities and hence "we don't care" which link data is routed through. At the peak of the curve for H_s, we did a computation that shows a pitfall of computing with the original algorithm used. Here $c_1 = c_2 = 27$, $p_1(0) = 0.2$ $p_2(0) = 0.8$ and $H_s = \ln(2) - 0.02$. We see in Figure 6 that there are oscillations in p_1 and p_2 that occur from flipping between the roots of $G(\gamma)$. Thus these oscillations are simply an artifact of the computation, not a part of the network algorithm. Clearly some care is needed

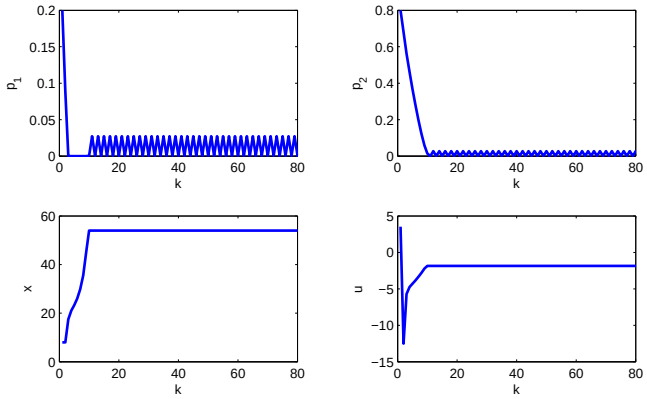

Figure 6: Computed evolution of $p_1(k)$ (upper left), $p_2(k)$ (upper right), $x(k)$ (lower right) and $u(k)$ (lower right) at $H_s = \ln 2 - 0.02$. The oscillations in the p_i is an artifact of the rootfinding, not from the network management algorithm.

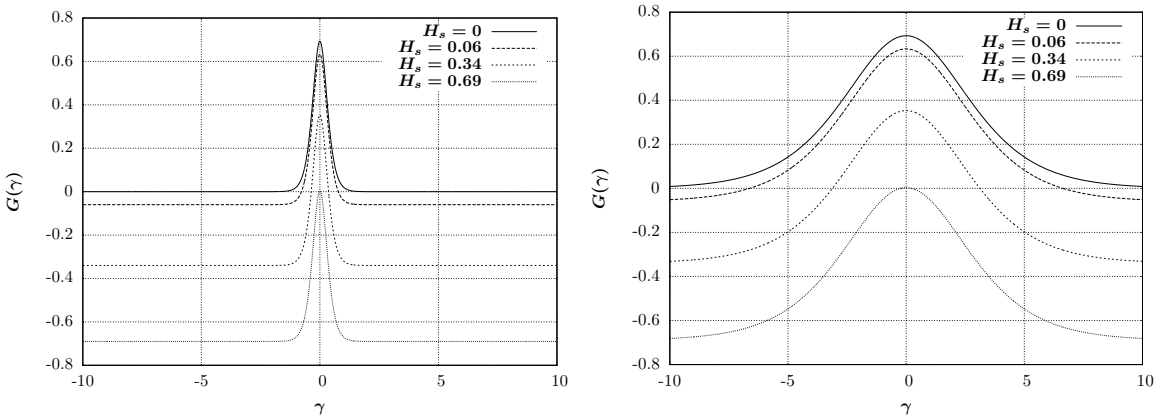

Figure 7: Plots of $G(\gamma)$ with various values of H_s for $p_1 = 0.2$ and $p_2 = 6$ (left) and $p_1 = 0$ and $p_2 = 0.68$ (right).

in the computations.

3.2 Analysis of the implicit equation for γ

The most complicated part of the iteration process is solving the nonlinear equation for γ (3.5). We focus attention on the properties of this problem for given p_1, p_2. Define the function $G(\gamma) \equiv \gamma\bar{p} + \log Z(\gamma) - H_s$, this is a scalar function of γ. The solution of the entropy constraint is then given by the root of $G(\gamma) = 0$. Figure 7 shows the form of $G(\gamma)$. It is not immediately obvious, but $G(\gamma)$ is an even function, $G(-\gamma) = G(\gamma)$, and it has a rapid decay for large γ. Consequently small changes in parameters can yield large changes in the solution γ.

3.3 Reduced gradient scheme without γ

Consider a simplification of the gradient method to eliminate the difficulties connected with the equation (3.5). If we assume $c_1 > c_2$ in the two-link network, then we will obtain a unique solution with $\beta_1 > \beta_2$ from the equations

$$-\beta_1 \log(\beta_1) - \beta_2 \log(\beta_2) = H_s, \qquad \beta_1 + \beta_2 = 1.$$

In particular, writing $\beta_1 \equiv \beta$ and $\beta_2 = 1 - \beta$, we can express this in term of

$$H(\beta) = H_s \qquad H(\beta) \equiv -\beta \log(\beta) - (1-\beta)\log(1-\beta), \tag{3.11}$$

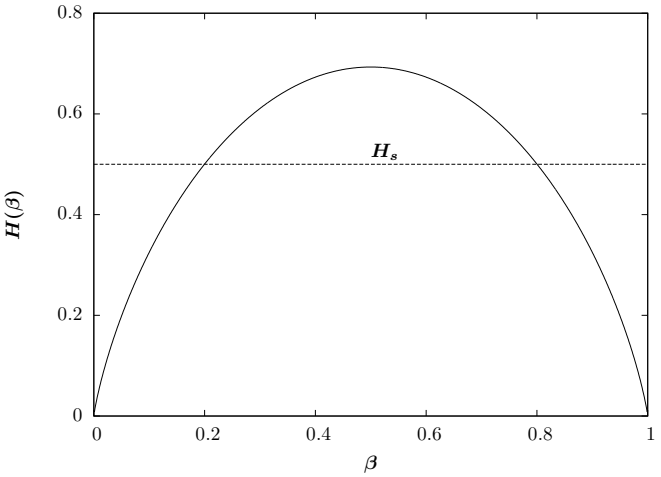

Figure 8: Plot of $H(\beta)$, equation (3.11), for the two-link network.

see Figure 8. It is clear that $H(\beta)$ is even with respect to $\beta = \frac{1}{2}$ where H takes its maximum value $H^{\max} = \log(2) \approx 0.6931$. It immediately follows that equilibrium solutions only exist for $H_s \leq H_s^{\max}$.

Solving $H(\beta) = H_s$ yields the values for the β_i. Using these fixed values, the gradient equations (3.2) can then be iterated to update the prices using (3.7) without reference to γ.

Then a pseudo-code algorithm for computing the network utility over a range of entropies is given by:

Input: $p_1(0), p_2(0), c_1, c_2, H_s, w, x_{\max}, h, N$
Output: U_{av}
foreach H_s **do**
  Compute β_1 and β_2 from equation (3.11)
  foreach k **do**
    Compute x_s, p_1 and p_2 from equation (3.2ab) and x from equation (3.7).
  end
  Compute U from equation(3.8)
end

(3.12)

Taking parameter values $p_1(0) = 0.2$, $p_2(0) = 6$, $c_1 = 7, c_2 = 0.85$ with $N = 1500$ being the total number of iterations, stepsize $h = 0.005$ and $w = 100$, $x_{\max} = 8$, we reproduce the two-link utility plot obtained in [1], see Figure 9 (left). The prices generated are shown in the right panel of that figure; it is interesting that for long times, they will eventually cross, yielding $p_1 > p_2$.

It is interesting to note that this approach effectively solves the primal problem with respect to the β_i directly, then uses the dual problem to determine the flow x and the prices.

3.4 Conversion of the gradient algorithm to an ODE system

Suppose that the prices remain strictly positive, $p_i > 0$, so that the truncation from the ramp functions in (3.2) does not come into play. Then those equation can be re-written as

$$\frac{p_1(t+h) - p_1(t)}{h} = \beta_1(t)x(t) - c_1, \qquad \frac{p_2(t+h) - p_2(t)}{h} = \beta_2(t)x(t) - c_2. \tag{3.13}$$

where we have replaced the discrete iterative index k with a continuous variable t. Let h represent a stepsize, as in the explicit forward Euler finite difference scheme. Consider the limit $h \to 0$ to recover a set of ordinary differential equations,

$$\frac{dp_1}{dt} = \beta_1(t)x(t) - c_1, \qquad \frac{dp_2}{dt} = \beta_2(t)x(t) - c_2. \tag{3.14}$$

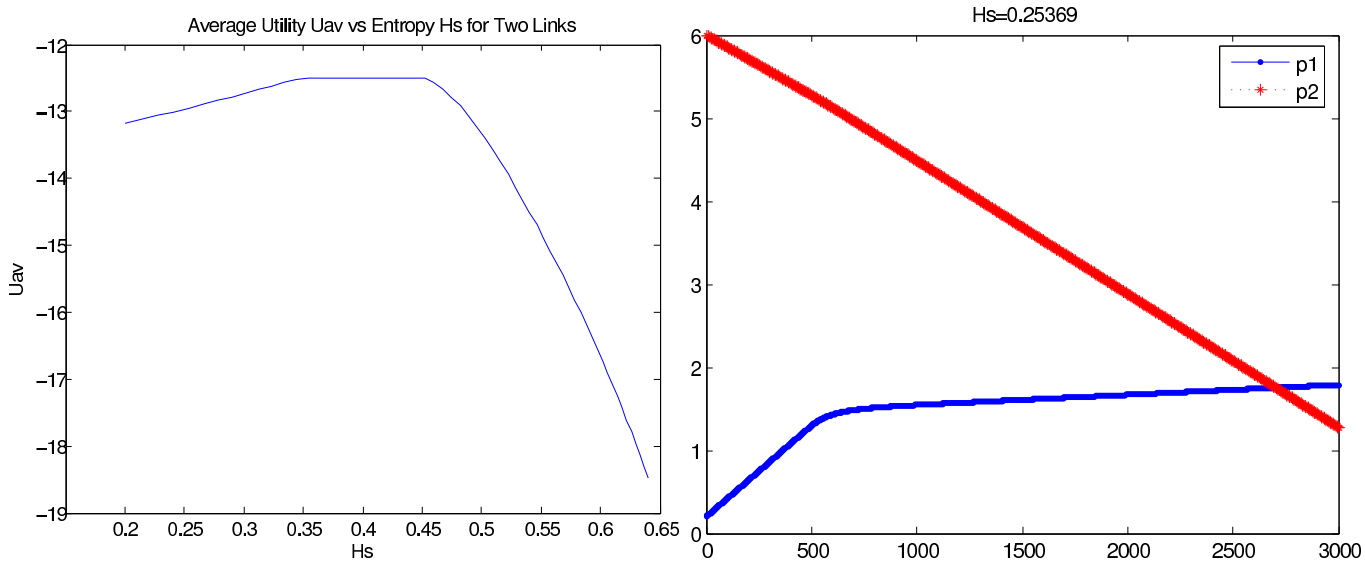

Figure 9: (Left) The utility of the two-link network computed using algorithm (3.12). (Right) The corresponding prices generated.

At this point we need to decide how to treat the $\beta_i(t)$; there are two choices:

1. $\beta_i(t)$ defined by (3.3) in terms of $p_i(t)$ and $\gamma(t)$. Then coupling the ODE system for the $p_i(t)$ to the constraint equation (3.5) for γ yields a difficult differential-algebraic (DAE) system.

2. Follow the route of the previous section in solving the primal problem to obtain constant β_i determined as a function of the entropy parameter H_s.

We use the second approach for further work; hence, given an allowable value of H_s, β_1, β_2 are determined constants.

Consider the steady-state problem

$$0 = \beta_1 x - c_1, \qquad 0 = \beta_2 x - c_2. \tag{3.15}$$

Setting $\beta = \beta_1$ and $\beta_2 = 1 - \beta$, We can solve the system for x,

$$\frac{c_1}{\beta} = x = \frac{c_2}{1-\beta}. \tag{3.16}$$

Consequently, steady states only exist if c_1, c_2 satisfy the relation

$$c_2 = \frac{1-\beta}{\beta} c_1. \tag{3.17}$$

Recalling that $x = \sqrt{w/\bar{p}}$, we now have $w(\beta/c_1)^2 = \bar{p}$ where $\bar{p} = \beta p_1 + (1-\beta)p_2$. Consequently we get

$$\beta p_1 + (1-\beta)p_2 = w\left(\frac{\beta}{c_1}\right)^2. \tag{3.18}$$

Namely for any $0 < \beta < 1$ and w, c_1 there is a solution where the two prices p_1, p_2 are linearly related. Hence we get an entire line of steady states. It is convenient to re-write (3.14) as

$$\frac{dp_1}{dt} = \beta\left[\sqrt{\frac{w}{\beta p_1 + (1-\beta)p_2}} - \frac{c_1}{\beta}\right], \qquad \frac{dp_2}{dt} = (1-\beta)\left[\sqrt{\frac{w}{\beta p_1 + (1-\beta)p_2}} - \frac{c_1}{\beta}\right]. \tag{3.19}$$

Linearizing about a (p_1, p_2) steady state, we compute the Jacobian matrix,

$$\mathbf{J} = \frac{-w^{1/2}}{2\bar{p}^{3/2}} \begin{bmatrix} \beta^2 & \beta(1-\beta) \\ \beta(1-\beta) & (1-\beta)^2 \end{bmatrix} \tag{3.20}$$

Examining the eigenvalues, it is immediate that we have one zero eigenvalue and one positive eigenvalue,

$$\lambda_1 = 0, \qquad \lambda_2 = -\frac{1}{2}\left(\frac{w}{\bar{p}^3}\right)^{1/2}(1 - 2\beta + 2\beta^2) < 0. \tag{3.21}$$

Thus, every steady state is stable – attractive and achievable from some set of initial conditions. The zero eigenvalue is a consequence of there being a continuous set of steady states; they are neutrally stable with respect to perturbations (to the initial conditions) that will shift the final state from one value along the line (3.18) to another.

3.5 Geometric constructions of the primal solution

Some insight into the problem can be found by developing the solution to the simple two-link primal problem using geometric constructions in the (β_1, β_2) parameter plane and considering $U(x) = x$.

- Condition (3.1b) yields a line segment from $(0, 1)$ to $(1, 0)$ in the first quadrant of (β_1, β_2) plane
- For any value for x condition (3.1c) selects a rectangular region as shown in the diagram

  $$0 \le \beta_1 \le \frac{c_1}{x} \qquad 0 \le \beta_2 \le \frac{c_2}{x} \tag{3.22}$$

  With the aspect ratio of the rectangle being maintained for all x and the size of the feasible region decreasing as x increases.
- Without the entropy constraint, the solution is found by increasing x until corner of the rectangle touches the diagonal line segment yielding

  $$\beta_1 = \frac{c_1}{c_1 + c_2} \qquad \beta_2 = \frac{c_2}{c_1 + c_2}$$

  This is the expected result that, with no entropy constraint, the optimal flow is distributed to fill the capacity of the two links.
- To include the entropy condition (3.1d) we indicate the level curves of the function $H(\beta_1, \beta_2) = -\beta_1 \log \beta_1 - \beta_2 \log \beta_2$. For any specific value of H_s the inequality version of (3.1d) describes the interior of one of the closed level curves with $H(\beta_1, \beta_2) = H_s$ on the curve. If $H_s < \log(2)$ then the feasible set consists of a finite segment of $\beta_1 + \beta_2 = 1$ within the level curve; for $H_s = \log(2)$ the level curve is tangent to the line and the only possible solution is $\beta_1 = \beta_2 = 1/2$, see Figure 10. For $H_s > \log(2)$ there are no solutions as the level curves do not intersect the line.

  The possible solutions are now restricted to the diagonal line segment between $(\beta_1(H_s), \beta_2(H_s))$ and $(\beta_2(H_s), \beta_1(H_s))$. As we increase x a solution can be selected in two possible ways. Firstly if H_s is sufficiently low then the solution found without this entropy condition remains the solution. For larger values of H_s the solution occurs when the corner of the rectangle touches the bounding level curve the entropy condition.

We note that Figure 2 in [1] shows some of the geometry of the dual problem.

4 The diamond network

We now consider the optimization problem for the diamond network, see Figure 2. Here there are constraints on the capacity of each link and we assign probabilities to each to each of the possible routes through the network. These probabilities are designated as $(\beta_1, \beta_2, \beta_3)$

$$\underline{\text{objective function}}: \qquad \max_{\beta_1, \beta_2, \beta_3} U(x) \tag{4.1a}$$

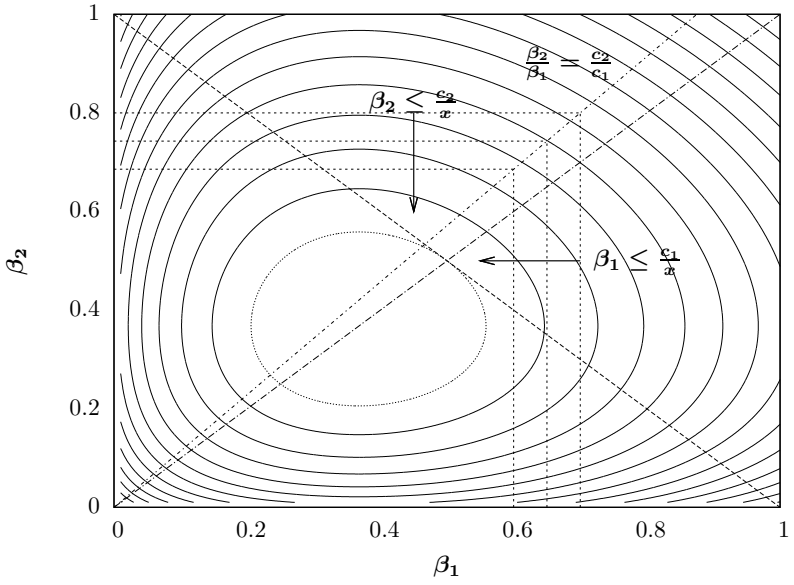

Figure 10: Geometric construction for the two-link primal problem. Curves show level curves of constant entropy.

$$\underline{\text{probability constraints}}: \quad \begin{cases} \beta_1 + \beta_2 + \beta_3 = 1 \\ \beta_1 \geq 0 \\ \beta_2 \geq 0 \\ \beta_3 \geq 0 \end{cases} \tag{4.1b}$$

$$\underline{\text{capacity constraints}}: \quad \begin{cases} \beta_1 x + \beta_3 x \leq c_1 \\ \beta_2 x \leq c_2 \\ \beta_1 x \leq c_3 \\ \beta_2 x + \beta_3 x \leq c_4 \\ \beta_3 x \leq c_5 \end{cases} \tag{4.1c}$$

$$\underline{\text{entropy}}: \quad -\beta_1 \log \beta_1 - \beta_2 \log \beta_2 - \beta_3 \log \beta_3 = H_s \tag{4.1d}$$

Two approaches were pursued for this network.

4.1 The mixed/reduced primal-dual gradient scheme

Analogous to the approach used in section 3.3, we can use equation (4.1c) and (4.1d) to determine the β_i from the primal problem. This will yield a one-parameter family of solutions, we reduce this to a single state with the choice $\beta_1 = \beta_2$. Subsequently the iteration equations for the prices and the overall flow are

$$p_1(k+1) = [p_1(k) - h\{c_1 - (\beta_1 + \beta_2)x(k)\}]^+ \tag{4.2a}$$
$$p_2(k+1) = [p_2(k) - h\{c_2 - \beta_2 x(k)\}]^+ \tag{4.2b}$$
$$p_3(k+1) = [p_3(k) - h\{c_3 - \beta_1 x(k)\}]^+ \tag{4.2c}$$
$$p_4(k+1) = [p_4(k) - h\{c_4 - (\beta_2 + \beta_3)x(k)\}]^+ \tag{4.2d}$$
$$p_5(k+1) = [p_5(k) - h\{c_5 - \beta_3 x(k)\}]^+ \tag{4.2e}$$

$$x(k) = \min\left\{\sqrt{\frac{w}{\beta_1 d_1(k) + \beta_2 d_2(k) + \beta_3 d_3(k)}}, x_{\max}\right\} \tag{4.3}$$

where $d_1(k) = p_1(k) + p_3(k)$, $d_2(k) = p_2(k) + p_4(k)$, $d_3(k) = p_1(k) + p_5(k) + p_4(k)$. Using this system, with the parameters and $c = [0.8, 0.3, 0.3, 0.8, 0.5], M = 30, N = 200, h = 0.005, w = x_{\max} = 40.$ and initial conditions on the prices, $p(0) = [12, 8, 8, 12, 4]$ we compute the utility for the diamond network in Figure 11.

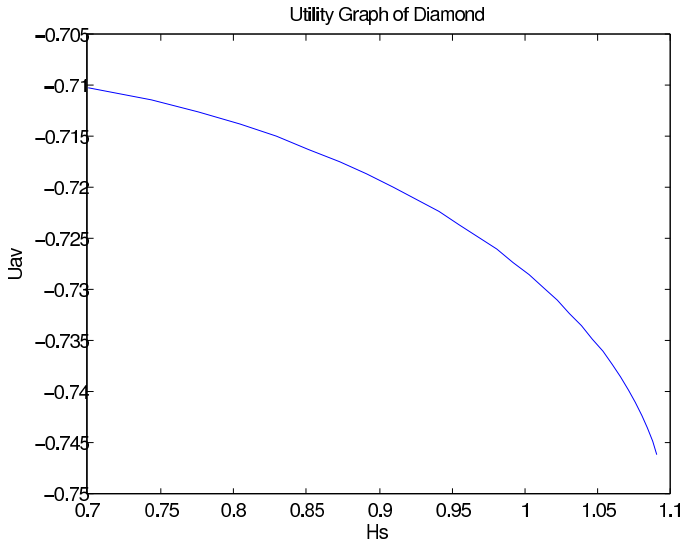

Figure 11: The utility of the diamond network computed the reduced dual gradient scheme.

4.2 Geometric construction of the primal solution

A geometric approach similar to the construction from section 3.5 can be done in three dimensions for the diamond-topology network.

- Condition (4.1b) yields a triangular surface from $(0,1,1)$ to $(1,1,0)$ to $(1,0,1)$ in the first octant of $\beta_1, \beta_2, \beta_3$ space, see Fig. 12

- For any value for x condition (4.1c) selects a rectangular region. The boundary of this region cuts the triangular surface in straight lines parallel to the edges of the triangle. The five constraints can be written in a simple form by using the probability conditions as

$$1 - \frac{c_4}{x} \leq \beta_1 \leq \frac{c_3}{x} \qquad 1 - \frac{c_1}{x} \leq \beta_2 \leq \frac{c_2}{x} \qquad \beta_3 \leq \frac{c_5}{x} \tag{4.4}$$

  We then note that the first two sets of inequalities only have a possible solution if

$$x \leq c_1 + c_2 \qquad \text{and} \qquad x \leq c_3 + c_4 \tag{4.5}$$

  and this corresponds to the obvious upper bound constraints on x dictated by the maximum flow out of the source or into the receiver.

- Without the entropy constraint, the solution is found by increasing x until corner of the rectangle region touches the triangular plane. The solution here depends on which link is the limiting capacity of the system. In particular the solution can either be a point or it can be a line segment (if the feasible region finally collapses due to either of the condition in (4.5)).

- To include (4.1d) we indicate the contours of the function $-\beta_1 \log \beta_1 - \beta_2 \log \beta_2 - \beta_2 \log \beta_2$ on the triangular plane (a projection of this plane is shown in the diagram). For any specific value of H_s (4.1d) is the interior of one of the convex closed contours on this triangular plane. This interior is of non-zero size as long as $H_s \leq \log(3)$.
  Hence the solution is now restricted to the interior of one of these contours on the triangular plane As we increase x the five lines representing the constraints reduce the possible region where the solution can occur. A solution again occurs in a number of possible ways. For example if H_s is sufficiently small then the solution found without this entropy condition remains the solution. For larger values of H_s the solution may occur in different ways either as a line segment or when several straight line constraints touch at the bounding contour of the entropy condition.

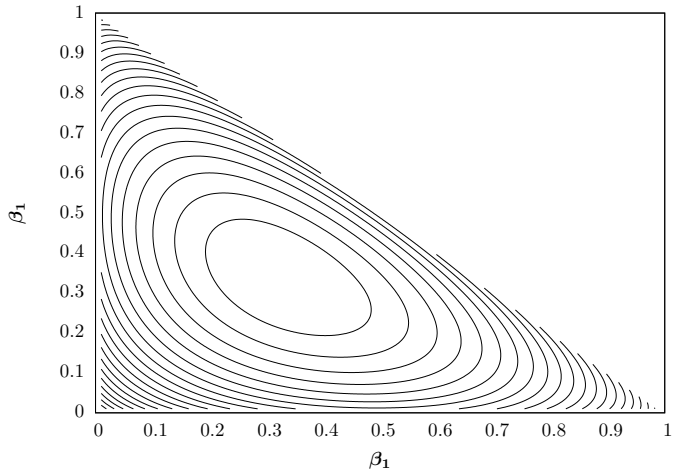

Figure 12: Projection of the contours of the entropy constraint onto the set of feasible probabilities in the (β_1, β_2) with $\beta_3 = 1 - \beta_1 - \beta_2$.

5 Monte Carlo Numerical Simulations of Network Traffic

We return to the simple two-link network, Figure 1, to examine insights that can be gained from observations of direct numerical simulations of network activity. Instead of studying the gradient algorithm for the equilibrium optimization dual problem, or the equilibrium optimization problem directly, here we do numerical simulations in MATLAB of real time-dependent fluctuating stochastic network traffic. If the stochastic processes are stationary, then the results obtained for the long-time behaviors of the simulations should match the equilibrium predictions.

Analytical descriptions of fluctuating network properties are much more challenging to obtain than equilibrium results. Yet, such simulations may hold very valuable information about system performance under wider ranges of conditions. It can be long and computationally-intensive to use such simulations to exhaustively optimize network routing, but they can be used to develop insights from well-selected prototype testing. In particular we examine the performance of several routing strategies that differ in the available network load information that they make use to see if the influence of non-equilibrium effects (i.e. time-gaps between messages, varying message sizes, etc) produce long-term differences in network performance.

We numerically simulate the flow of messages in the two-channel model. We make the following assumptions:

- Channel 1 always has the greater capacity, $c_1 > c_2$.
- New messages entering the network will be randomly generated at the source, one at a time. Messages can be of fixed or variable length and will have randomly distributed inter-arrival times (exponentially distributed inter-arrival times yield a Poisson process).

We developed four simple routing strategies to be tested. They are:

1. **Stacked**: In which we queue the next incoming message on the channel to be freed-up first. This generalizes OSPF (open shortest path first) by calculating when the current load on each channel will be completed to determine when each will be open. This can be further extended by directly considering the predicted arrival time of the message at the destination via either channel.
2. **Unbiased**: Which randomly selects between channels 1 and 2 in an unbiased manner for each message sent.
3. **Biased**: Which randomly selects between channels 1 and 2 for each message based on the equilibrium probabilities $\beta_1 = c_1/(c_1 + c_2)$, $\beta_2 = c_2/(c_1 + c_2)$, but without regard to any waiting times for the channels.

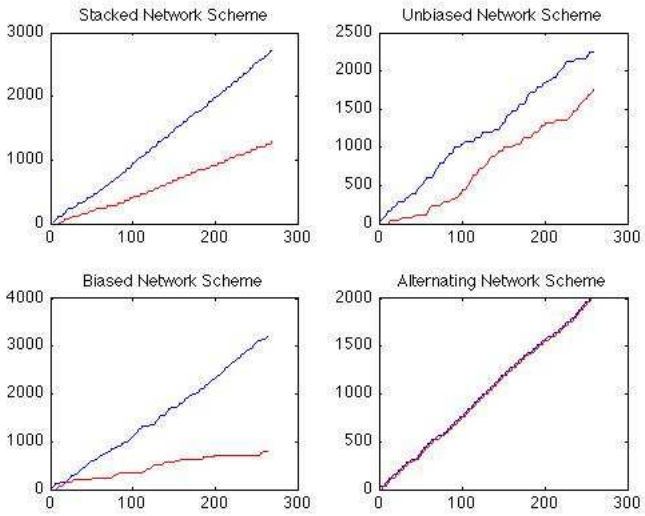

Figure 13: Two-link network traffic simulations via the four routing strategies considered, with $c_1 = 13$ and $c_2 = 6$ with 100 messages each of length 40.

4. **Alternating**: Which deterministically alternates sending messages between channel 1 and channel 2.

5.1 Simulation Results

- The Unbiased and Alternating strategies perform similarly, since both distribute messages evenly on both channels on average.
- For longer message sizes (or lower channel capacities) we observe that the Biased strategy performs better than the others.
- According to our performance measure, the Stacking strategy is the optimal method, because it takes the next free channel into account and more efficiently distributes traffic.

In Fig. 5.1 we simulated network traffic consisting of 100 messages of identical length (40) with exponentially distributed inter-arrival times ($\lambda = 5$). The two-link network considered had capacities $c_1 = 13$ and $c_2 = 6$. The measure of system performance that we considered was the overall network speed:

$$\text{Speed} = \frac{\text{total data sent}}{\text{total time to complete transmissions}}. \tag{5.1}$$

The ideal maximum network speed for this model is $S^* = 40 \times 100/(13 + 6) \approx 210.526$ (i.e. total data divided by total available bandwidth). This ideal speed cannot be achieved in practice since there are random time gaps between messages and messages are not broken up across all available channels. Consequently we can consider the network speeds generated by the different strategies relative to S^* as their performance:

$$P = S^* - S, \tag{5.2}$$

namely, $P \geq 0$ and the smaller the value of P, the closer the performance to optimal speed. The results are summarized in Fig. 14.

5.2 Further simulations

Further simulations considered more realistic models of network message traffic with more degrees of freedom:

Rank	Strategy	Performance
1	Stacked	3.71501
2	Biased	4.14064
3	Unbiased	6.15829
4	Alternating	7.14454

Figure 14: Table comparing performance of four routing strategies for network traffic in the two-link network. Rank 1–4 indicates best to worst overall network speed.

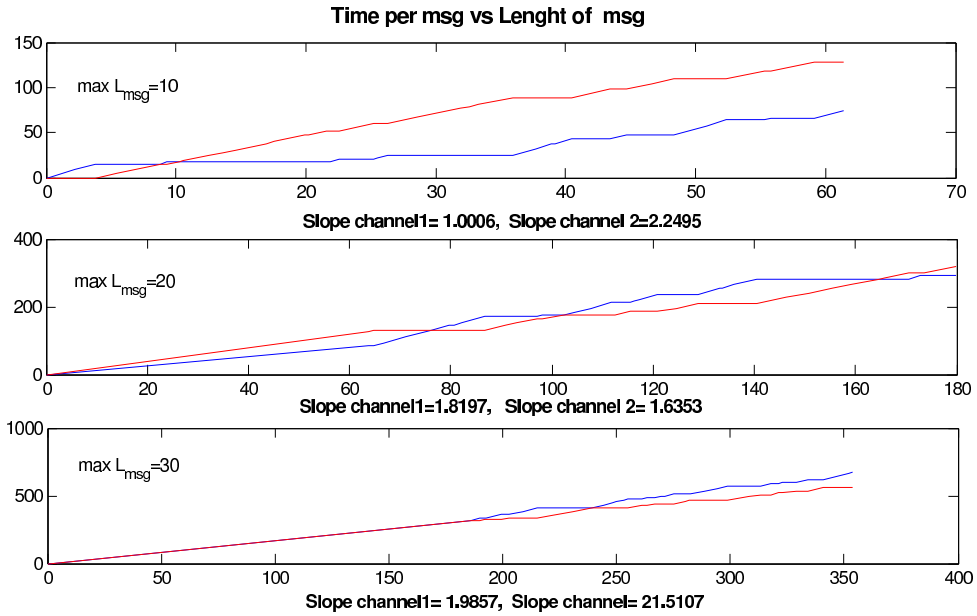

Figure 15: Strategy 1

- The time of arriving messages is random with exponential distribution with parameter λ, this parameter itself varies (uniformly) randomly.
- The length of messages is not constant but is generated randomly using a uniform distribution, in an interval $[0, L]$.

Three simple-to-simulate strategies were considered for this testing:

1. Selecting a channel randomly (Uniformly), i.e. the Biased strategy.
2. Selecting the first channel free using one channel at the time (i.e. only one message can be sent through the network at a time).
3. Selecting the free channel. Both channel can work at the same time, i.e. the Stacked strategy.

The following figures show the speed of the channels when we are considering different length messages and different strategies, also we can see the behavior of the channel. The constant lines means that the channel is busy during this time with the same message.

- **Strategy 1** In this strategy, we see that as expected, the two channels were used to send comparable numbers of messages.
- **Strategy 2** In this strategy, the channel that is free gets the message, making the strategy a little more efficiency than strategy 1, but that the process uses only one channel at a time slows the overall speed.
- **Strategy 3** In this strategy, the channel that is free first gets the message and both channels can be used at the same time, improving speed over the other strategies.

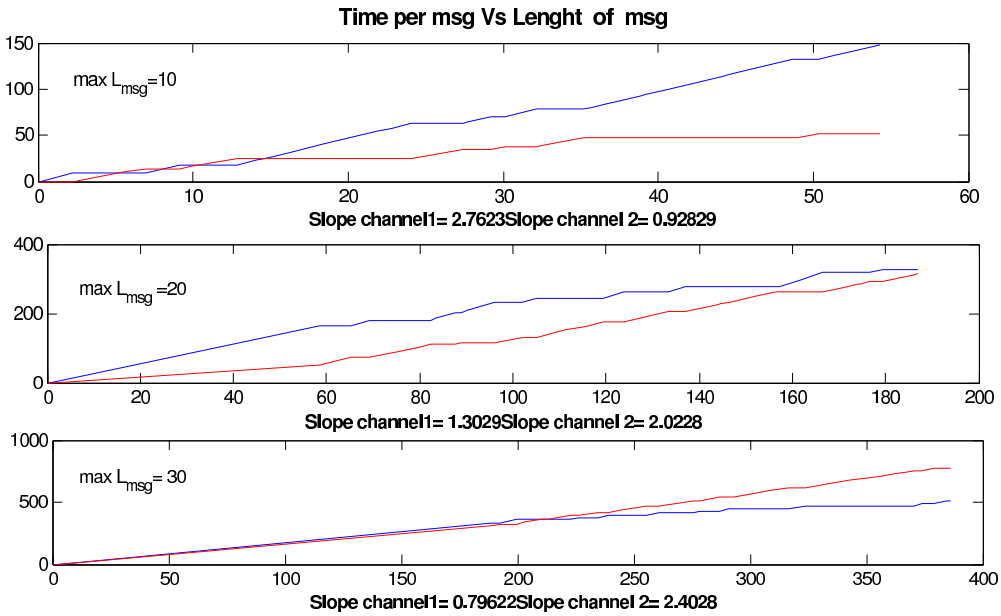

Figure 16: Strategy 2

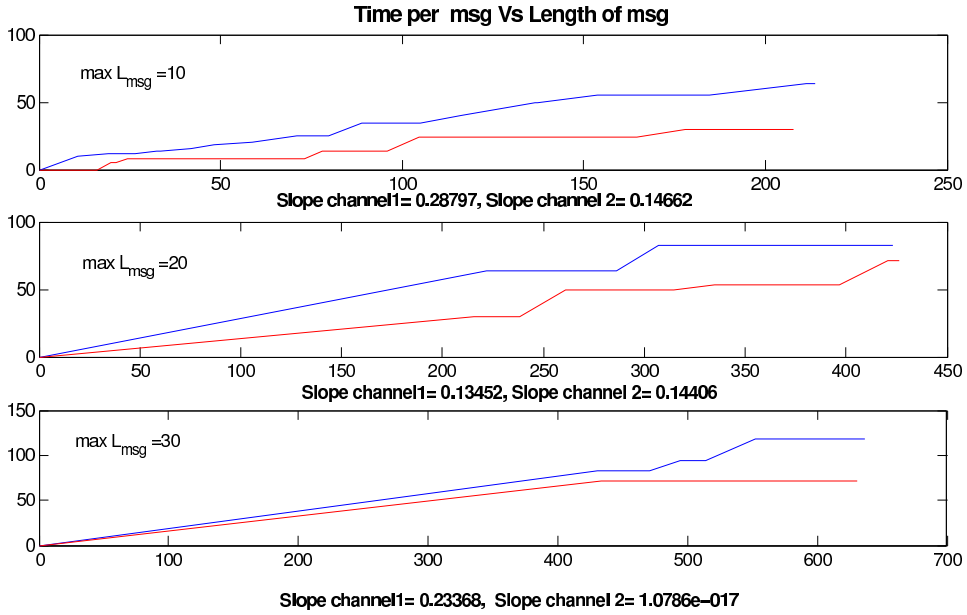

Figure 17: Strategy 3

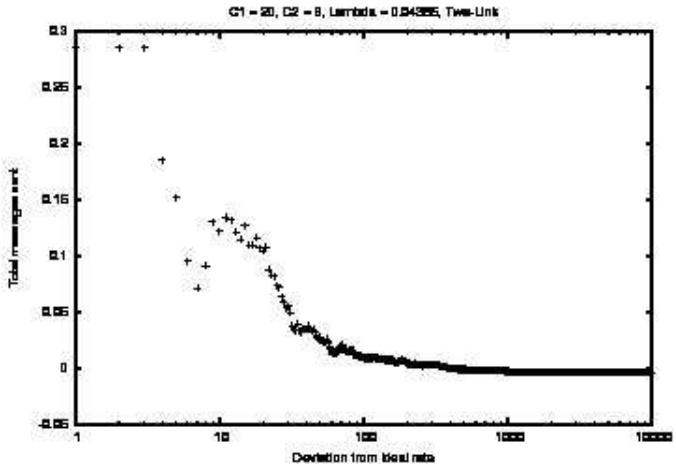

Figure 18: Deviation from expected network performance.

5.3 Consideration of simulation parameters

Finally, we considered the influence of two simulation parameters on the behavior of network traffic:

1. λ: inter-arrival delay – the influence of different values for the average inter-arrival time delay was examined with messages of fixed-size.
   - We tracked the variation in data-delivery rates over a range of λ
   - For some optimal λ, observed and optimal route distributions were very close.
   - As $\lambda \to 0$ (i.e. short time delays, a nearly continuous stream of messages), the channel with the larger capacity is favored.
   - As $\lambda \to \infty$ (i.e. long delays, sparse traffic), the channel choice is then based on whether which is open at the time the message arrives.
2. μ: the average message length with fixed-average-size inter-arrival times.
   - Increasing message size tended to equalize flow (on a per message basis) in the two channels.

The last result points out that depending on how performance of the network is being measured, different conclusions can be reached:

- Measures of bytes transmitted provides elementary view of speed of network transfer (which should directly correspond to channel capacities and β's).
- Measures of messages per time yields different results due to variations in sizes of messages.

References

[1] F. Y. Hunt and V. Marbukh. Measuring the utility/path diversity tradeoff in multipath protocols. *preprint*, 2009.

[2] F.P. Kelly, A. K. Maulloo, and Tan D. K. H. Rate control for communication networks: shadow prices, proportional fairness and stability. *J. Operational Research Soc.*, 49(3):237–252, 1998.

[3] S. H. Low and D. E. Lapsley. Optimization flow control i: Basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, 7(6):861–874, 1999.

[4] A. Nedic and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM J. Optim.*, 19(4):1757–1780, 2009.

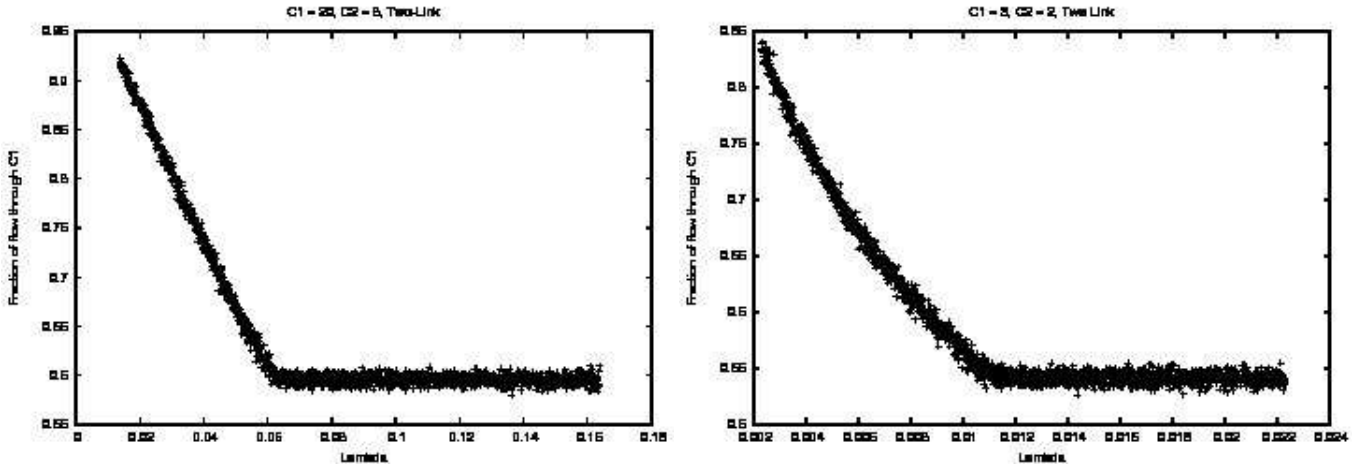

Figure 19: Simulations results with different values of λ.

[5] J. Wang, L. Li, S. H. Low, and J. C. Doyle. Cross-layer optimization in TCP/IP networks. *IEEE/ACM Transactions on Networking*, 13(3):582–595, 2005.

[6] Wikipedia. Karush-Kuhn-Tucker conditions — Wikipedia, the free encyclopedia, 2009. [Online; accessed 20-July-2009].

[7] Wikipedia. Max-flow min-cut theorem — Wikipedia, the free encyclopedia, 2009. [Online; accessed 20-July-2009].

[8] H. Yaiche, R. R. Mazumdar, and C. Rosenberg. A game theoretic framework for bandwidth allocation and pricing in broadband networks. *IEEE/ACM Transactions on Networking*, 8(5):667–678, 2000.

[9] L. Ye, Z. Wang, H. Che, H. B. C. Chan, and C. M Lagoa. Utility function of TCP. *Computer Communications*, 32:800–805, 2009.

[10] M. Zukerman, M. Mammadov, L. Tan, I. Ouveysi, and L. L. H. Andrew. To be fair or efficient or a bit of both. *Computers and Operations Research*, 35:3787–3806, 2008.