Pricing and rate control in communication networks

Frank Kelly
Statistical Laboratory
University of Cambridge

The basic model

J — set of *resources*

R — set of *routes*

A route r is a subset $r \subset J$.

Let

$$\begin{aligned} A_{jr} &= 1 \quad \text{if } j \in r \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

For $j \in J$, let C_j be the *capacity* of resource j.

Each route has a *user*: if x_r is the *rate* on route r, then the *utility* to user r is $U_r(x_r)$.

$U_r(\cdot)$ — increasing, strictly concave, continuously differentiable on $[0, \infty)$ — *elastic* traffic.

Let $C = (C_j, j \in J)$, $x = (x_r, r \in R)$. Then

$$Ax \le C.$$

The system problem

$$\begin{aligned} &\text{maximize} && \sum_{r\in R} U_r(x_r) \\ &\text{subject to} && Ax \le C \\ &\text{over} && x \ge 0. \end{aligned}$$

- Maximize aggregate utility, subject to capacity constraints

The user problem

$$
\begin{array}{ll}
\text{maximize} & U_r\left(\dfrac{w_r}{\lambda_r}\right) - w_r \\
\text{over} & w_r \geq 0.
\end{array}
$$

- User r chooses an amount to pay per unit time, w_r, and receives in return a flow $x_r = w_r/\lambda_r$

The network problem

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{r \in R} w_r \log x_r \\ \text{subject to} & Ax \le C \\ \text{over} & x \ge 0. \end{array}$$

- *As if* the network maximizes a logarithmic utility function, but with constants $(w_r, r \in R)$ chosen by the users

Three optimization problems

SYSTEM(U, A, C)

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{r \in R} U_r(x_r) \\ \text{subject to} & Ax \leq C \\ \text{over} & x \geq 0. \end{array}$$

USER$_r(U_r; \lambda_r)$

$$\begin{array}{ll} \text{maximize} & U_r\left(\dfrac{w_r}{\lambda_r}\right) - w_r \\ \text{over} & w_r \geq 0. \end{array}$$

NETWORK$(A, C; w)$

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{r \in R} w_r \log x_r \\ \text{subject to} & Ax \leq C \\ \text{over} & x \geq 0. \end{array}$$

Theorem (problem decomposition)

There exist vectors $\lambda = (\lambda_r, r \in R)$,
$w = (w_r, r \in R)$ and $x = (x_r, r \in R)$ such that

(i) $w_r = \lambda_r x_r$ for $r \in R$

(ii) w_r solves $\mathrm{USER}_r(U_r; \lambda_r)$

(iii) x solves $\mathrm{NETWORK}(A, C; w)$.

The vector x then also solves $\mathrm{SYSTEM}(U, A, C)$.

- Thus the system problem may be solved by solving simultaneously the network and user problems

Sketch of proof

Lagrangian for system:

$$L_{\text{system}}(x, z; \mu) = \sum_{r \in R} U_r(x_r) + \mu^T(C - Ax - z)$$

$$\frac{\partial}{\partial x_r} L_{\text{system}}(x, z; \mu) = U_r'(x_r) - \sum_{j \in r} \mu_j$$

Lagrangian for network:

$$L_{\text{network}}(x, z; \mu) = \sum_{r \in R} w_r \log x_r + \mu^T(C - Ax - z)$$

$$\frac{\partial}{\partial x_r} L_{\text{network}}(x, z; \mu) = \frac{w_r}{x_r} - \sum_{j \in r} \mu_j$$

User objective: $U_r\left(\frac{w_r}{\lambda_r}\right) - w_r$.

$$\frac{\partial}{\partial w_r}\left[U_r\left(\frac{w_r}{\lambda_r}\right) - w_r\right] = \frac{1}{\lambda_r}U_r'\left(\frac{w_r}{\lambda_r}\right) - 1.$$

If user r chooses w_r so that $U_r'\left(\frac{w_r}{\lambda_r}\right) = \lambda_r$, then earlier derivatives become identical, under the identification

$$x_r = \frac{w_r}{\lambda_r}, \quad \lambda_r = \sum_{j\in r}\mu_j.$$

The dual problem

$$L_{\text{network}}(x, z; \mu) = \sum_{r \in R} w_r \log x_r + \mu^T (C - Ax - z)$$

$$\begin{aligned} \frac{\partial}{\partial x_r} L_{\text{network}}(x, z; \mu) &= \frac{w_r}{x_r} - \sum_{j \in r} \mu_j \\ \frac{\partial}{\partial z_j} L_{\text{network}}(x, z; \mu) &= -\mu_j \,. \end{aligned}$$

Thus the optimal rate is

$$x_r = \frac{w_r}{\sum_{j \in r} \mu_j} \,,$$

where

$$\mu \geq 0, \quad Ax \leq C, \quad \mu^T (C - Ax) = 0.$$

Hence

$$\max_{x,z\geq 0} L_{\text{network}}(x,z;\mu) = \sum_{r\in R} w_r \log \frac{w_r}{\sum_{j\in r}\mu_j} + \mu^T C,$$

and so the dual to NETWORK$(A, C; w)$ becomes:

DUAL$(A, C; w)$

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{r\in R} w_r \log\Big(\sum_{j\in r}\mu_j\Big) - \sum_{j\in J}\mu_j C_j \\ \text{over} & \mu \geq 0. \end{array}$$

Max-min fairness

$x = (x_r, r \in R)$ is *feasible* if $x \geq 0$ and $Ax \leq C$.

$x = (x_r, r \in R)$ is *max-min fair* if it is feasible, and if, for any other feasible vector y,

$$\exists r : y_r > x_r \Rightarrow \exists s : y_s < x_s < x_r.$$

- The compactness and convexity of feasible region imply that such a vector exists and is unique
- As a concept, max-min fairness has been much discussed by political philosophers (major work: J. Rawls, *A Theory of Justice*, 1971)

Proportional fairness

$x = (x_r, r \in R)$ is *proportionally fair* if it is feasible, and if, for any other feasible vector y, the aggregate of proportional changes is zero or negative:

$$\sum_{r \in R} \frac{y_r - x_r}{x_r} \leq 0.$$

Let $w = (w_r, r \in R)$ be a vector of weights, or charges.

$x = (x_r, r \in R)$ is *proportionally fair per unit charge* if it is feasible, and if, for any other feasible vector y,

$$\sum_{r \in R} w_r \frac{y_r - x_r}{x_r} \leq 0.$$

Proposition

A vector x solves NETWORK$(A, C; w)$ if and only if it is proportionally fair per unit charge.

Sketch of proof

Consider a perturbation from x to y, where $y_r = x_r + \delta x_r, r \in R$. This increases the objective function $\sum_{r \in R} w_r \log x_r$ by an amount

$$\sum_{r \in R} w_r \frac{\delta x_r}{x_r} + o(\delta x).$$

Thus result follows, from convexity of the feasible region and strict concavity of the objective function.

Solution of the network problem

- Strategy: design algorithms, based on the primal and dual problems for the network, to implement proportional fairness
- Joint work with Aman Maulloo and David Tan
- Paper available at `http://www.statslab.cam.ac.uk/~frank/rate.html`

A primal algorithm

$$
\begin{aligned}
\frac{d}{dt}\, x_r(t) \quad &= \kappa\Big(w_r - x_r(t) \sum_{j \in r} \mu_j(t)\Big) \\
\mu_j(t) \quad &= p_j\Big(\sum_{s: j \in s} x_s(t)\Big)
\end{aligned}
$$

where $p_j(\cdot)$ is a non-negative, continuous, increasing function.

Interpretation:

- resource j generates feedback signals at rate $\mu_j(t)$
- signals sent to each user r whose route passes through resource j
- multiplicative decrease in flow x_r at rate proportional to stream of feedback signals received
- linear increase in flow x_r at rate proportional to w_r

Theorem (global stability)

$$\mathcal{U}(x) = \sum_{r \in R} w_r \log x_r - \sum_{j \in J} \int_0^{\sum_{s:j \in s} x_s} p_j(y)dy$$

is a Lyapunov function for the primal algorithm. The unique value x maximizing $\mathcal{U}(x)$ is a stable point of the system, to which all trajectories converge.

Proof

$$\frac{\partial}{\partial x_r} \mathcal{U}(x) = \frac{w_r}{x_r} - \sum_{j \in r} p_j \Big(\sum_{s:j \in s} x_s \Big),$$

and so

$$\frac{d}{dt} \mathcal{U}\big(x(t)\big) = \sum_{r \in R} \frac{\partial U}{\partial x_r} \frac{d}{dt} x_r(t)$$

$$= \kappa \sum_{r \in R} \frac{1}{x_r(t)} \left(w_r - x_r(t) \sum_{j \in r} p_j \Big(\sum_{s:j \in s} x_s(t) \Big) \right)^2$$

$$\geq 0.$$

If

$$
\begin{aligned}
p_j(y) \quad &= \infty \qquad y > C_j \\
&= 0 \qquad y \leq C_j
\end{aligned}
$$

then maximization of the Lyapunov function $\mathcal{U}(x)$ becomes the primal problem NETWORK$(A, C; w)$.

Rate of convergence

Let x be unique vector maximizing $\mathcal{U}(x)$, let $\mu_j = p_j(\sum_{s:j\in s} x_s)$ and suppose $p_j(\cdot)$ is differentiable at this point, with derivative p'_j. Let

$$x_r(t) = x_r + x_r^{1/2} y_r(t).$$

Then, linearizing about x,

$$\begin{aligned} \frac{d}{dt} y(t) &= -\kappa(WX^{-1} + X^{1/2}A^T P' A X^{1/2}) y(t) \\ &= -\kappa \Gamma^T \Phi \Gamma y(t) \end{aligned}$$

where

$$\begin{aligned} X &= \mathrm{diag}(x_r, r \in R) \\ W &= \mathrm{diag}(w_r, r \in R) \\ P' &= \mathrm{diag}(p'_j, j \in J) \\ \Phi &= \mathrm{diag}(\phi_r, r \in R) \qquad \text{– the matrix} \end{aligned}$$

of (positive, real) eigenvalues.

Random effects

Congestion indication with joint feedback

$$dx(t) =$$
$$\kappa\left(w_r dt - x_r(t)\sum_{j\in r}\varepsilon_j dN_j\left(\varepsilon_j^{-1}\int_0^t \mu_j(\tau)d\tau\right)\right)$$

where $(N_j(\tau), \tau \geq 0), j \in J$, are independent unit rate Poisson processes.

Interpretation:

- resource j generates feedback signals at rate $\varepsilon_j^{-1}\mu_j(t)$
- a copy is sent to each user r whose route passes through resource j
- user r reacts to such a feedback signal by reducing $x_r(t)$ by an amount $\kappa\varepsilon_j x_r(t)$

Congestion indication with individual feedback

$$dx_r(t) =$$
$$\kappa\left(w_r dt - \sum_{j\in r} \varepsilon_j dN_{jr}\left(\varepsilon_j^{-1}\int_0^t x_r(\tau)\mu_j(\tau)d\tau\right)\right)$$

where $(N_{jr}(\tau), \tau \geq 0), j \in J, r \in R$, are independent unit rate Poisson processes.

Interpretation:

- feedback signals from resource j to user r arise at rate $\varepsilon_j^{-1} x_r(t)\mu_j(t)$
- user r reacts to a feedback signal by reducing $x_r(t)$ by an amount $\kappa\varepsilon_j$

General Brownian model

$$dy(t) = -\kappa\big(\Gamma^T \Phi \Gamma y(t) dt + F dB(t)\big)$$

$$F - |R| \times |I| \quad \text{matrix}$$

$B(t) = \big(B_i(t), i \in I\big)$ – independent standard Brownian motions.

Stationary solution?

$$y(t) = -\kappa \int_{-\infty}^{t} e^{-\kappa(t-\tau)\Gamma^T \Phi \Gamma} F dB(\tau).$$

But this is just a linear transformation of the Gaussian process $(B(\tau), \tau < t)$. Thus

$$y(t) \sim N(0, \Sigma)$$

where

$$\begin{aligned}
\Sigma &= \mathbb{E}[y(t)y(t)^T] \\
&= \kappa^2 \int_{-\infty}^{0} e^{\kappa\tau\Gamma^T\Phi\Gamma} F F^T e^{\kappa\tau\Gamma^T\Phi\Gamma} d\tau \\
&= \kappa\Gamma^T \left[\int_{-\infty}^{0} e^{\tau\Phi} \Gamma F F^T \Gamma^T e^{\tau\Phi} d\tau \right] \Gamma.
\end{aligned}$$

Define $[\Gamma F;\Phi]$ by

$$\begin{aligned}
[\Gamma F;\Phi]_{rs} &= \left[\int_{-\infty}^{0} e^{\tau\Phi} \Gamma F F^T \Gamma^T e^{\tau\Phi} d\tau \right]_{rs} \\
&= \frac{[\Gamma F F^T \Gamma^T]_{rs}}{\phi_r + \phi_s}.
\end{aligned}$$

Then

$$\Sigma = \kappa\Gamma^T [\Gamma F;\Phi]\Gamma.$$

A dual algorithm

$$\frac{d}{dt}\,\mu_j(t) \quad = \kappa\Big(\sum_{r:j\in r} x_r(t) - q_j\big(\mu_j(t)\big)\Big)$$

$$x_r(t) \qquad = \frac{w_r}{\sum_{k\in r}\mu_k(t)}$$

where $q_j(\cdot)$ is continuous, strictly increasing function, with $q_j(0) = 0$.

Interpretation:

- $q_j(\mu_j)$ is the flow through resource j which generates a price μ_j
- rate of change of prices $\mu_j(t)$ proportional to *excess demand* at these prices

Theorem (global stability)

$$\mathcal{V}(\mu) = \sum_{r \in R} w_v \log\Big(\sum_{j \in r} \mu_j\Big) - \sum_{j \in J} \int_0^{\mu_j} q_j(\eta) d\eta$$

is a Lyapunov function for the dual algorithm. The unique value μ maximizing $\mathcal{V}(\mu)$ is a stable point of the system, to which all trajectories converge.

If

$$q_j(\eta) = C_j \quad \eta > 0$$

then maximization of the Lyapunov function $\mathcal{V}(\mu)$ becomes the dual problem $\mathrm{DUAL}(A, C; w)$.

A random network

$$\begin{aligned}\mu_j[t+1] &= \mu_j[t] \\ &\quad +\kappa\Big(\textstyle\sum_{r:j\in r} N_r[t-d(j,r)] - q_j(\mu_j[t])\Big)\end{aligned}$$

$$x_r[t] = \frac{w_r}{\sum_{k\in r}\mu_k[t-d(k,r)]}$$

$$N_r[t] \sim \text{Poisson}\,(x_r[t])$$

$\lvert J\rvert = 100$	– one hundred resources
$\lvert R\rvert = 1000$	– one thousand routes
$A(j,r) = 1$	with probability 0.1
$= 0$	with probability 0.9
$d(j,r)$	uniformly distributed over $\{0,1,\ldots,10\}$

Thus

- average length of a route is 10 resources
- average number of routes through a resource is 100
- largest time delay between a user and a resource is 10