Resource sharing in networks

Frank Kelly
University of Cambridge
www.statslab.cam.ac.uk/~frank

Chinese University of Hong Kong
8 February 2010
Outline

• Fairness in networks

• Rate control in communication networks (relatively well understood)

• Philosophy: optimization vs fairness

• Ramp metering (very preliminary)
Network structure

$J$ - set of resources

$R$ - set of routes

$A_{jr} = 1$ - if resource $j$ is on route $r$

$A_{jr} = 0$ - otherwise
Notation

\( J \) - set of resources

\( R \) - set of users, or routes

\( j \in R \) - resource \( j \) is on route \( r \)

\( x_r \) - flow rate on route \( r \)

\( U_r(x_r) \) - utility to user \( r \)

\( C_j \) - capacity of resource \( j \)

\( Ax \leq C \) - capacity constraints
The system problem

SYSTEM(U,A,C):

Maximize \( \sum_{r \in R} U_r(x_r) \)

subject to \( Ax \leq C \)

over \( x \geq 0 \)

Maximize aggregate utility, subject to capacity constraints
The user problem

\[ \text{USER}_r(U_r; \lambda_r): \quad \text{Maximize } U_r \left( \frac{w_r}{\lambda_r} \right) - w_r \]

over \quad w_r \geq 0

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

\[ \text{NETWORK}(A,C;w): \quad \text{Maximize} \sum_{r \in R} w_r \log x_r \]

subject to \quad Ax \leq C

over \quad x \geq 0

As if the network maximizes a logarithmic utility function, but with constants \( \{w_r\} \) chosen by the users
Problem decomposition

Theorem: the system problem may be solved by solving simultaneously the network problem and the user problems

Max-min fairness

Rates \{x_r\} are *max-min fair* if they are feasible:

\[ x \geq 0, \quad Ax \leq C \]

and if, for any other feasible rates \{y_r\},

\[ \exists r : y_r > x_r \quad \Rightarrow \quad \exists s : y_s < x_s < x_r \]

Rawls 1971,
Bertsekas, Gallager 1987
Proportional fairness

Rates \( \{x_r\} \) are *proportionally fair* if they are feasible:

\[
x \geq 0, \quad Ax \leq C
\]

and if, for any other feasible rates \( \{y_r\} \), the aggregate of proportional changes is negative:

\[
\sum_{r \in R} \frac{y_r - x_r}{x_r} \leq 0
\]
Weighted proportional fairness

A feasible set of rates \( \{x_r\} \) are such that are weighted proportionally fair if, for any other feasible rates \( \{y_r\} \),

\[
\sum_{r \in R} w_r \frac{y_r - x_r}{x_r} \leq 0
\]
Fairness and the network problem

Theorem: A set of rates \( \{x_r\} \) solves the network problem, \( \text{NETWORK}(A,C;w) \), if and only if the rates are weighted proportionally fair.
Bargaining problem (Nash, 1950)

Solution to $\text{NETWORK}(A,C;w)$ with $w = 1$ is unique point satisfying

- Pareto efficiency
- Symmetry
- Independence of Irrelevant Alternatives

(General $w$ corresponds to a model with unequal bargaining power)
Market clearing equilibrium
(Gale, 1960)

Find prices $p$ and an allocation $x$ such that

\begin{align*}
    p &\geq 0, \quad Ax \leq C & \text{(feasibility)} \\
    p^T (C - Ax) &= 0 & \text{(complementary slackness)} \\
    w_r &= x_r \sum_{j \in r} p_j, \quad r \in R & \text{(endowments spent)}
\end{align*}

Solution solves $\text{NETWORK}(A,C;w)$
Optimization formulation of rate control

Various forms of fairness, can be cast in an optimization framework. We’ll illustrate, for the rate control problem.

\[ n_r \quad - \quad \text{number of flows on route } \ r \]
\[ x_r \quad - \quad \text{rate of each flow on route } \ r \]

Given the vector \( n = (n_r, r \in R) \)

how are the rates \( x = (x_r, r \in R) \)

chosen?
Optimization formulation

Suppose \( x = x(n) \) is chosen to

maximize \[ \sum_r w_r n_r \frac{x_r^{1-\alpha}}{1-\alpha} \]

subject to \[ \sum_r A_{jr} n_r x_r \leq C_j \quad j \in J \]
\[ x_r \geq 0 \quad r \in R \]

(weighted \( \alpha \)-fair allocations, Mo and Walrand 2000)

\( 0 < \alpha < \infty \) \quad (\text{replace } \frac{x_r^{1-\alpha}}{1-\alpha} \text{ by } \log(x_r) \quad \text{if } \alpha = 1 ) \)
Solution

\[
x_r = \left( \frac{w_r}{\sum_j A_{jr} p_j(n)} \right)^{1/\alpha} \quad r \in R
\]

where

\[
\sum_r A_{jr} n_r x_r \leq C_j \quad j \in J; \quad x_r \geq 0 \quad r \in R
\]

\[
p_j(n) \geq 0 \quad j \in J
\]

\[
p_j(n) \left( C_j - \sum_r A_{jr} n_r x_r \right) \geq 0 \quad j \in J
\]

\[p_j(n)\] - shadow price (Lagrange multiplier) for the resource \( j \) capacity constraint
Examples of $\alpha$-fair allocations

maximize \[ \sum_r w_r n_r \frac{x_r^{1-\alpha}}{1-\alpha} \]

subject to \[ \sum_r A_{jr} n_r x_r \leq C_j \quad j \in J \]
\[ x_r \geq 0 \quad r \in R \]

\[ x_r = \left( \frac{w_r}{\sum_j A_{jr} p_j(n)} \right)^{1/\alpha} \quad r \in R \]

- $\alpha \to 0$ \quad ($w = 1$) \quad - maximum flow
- $\alpha \to 1$ \quad ($w = 1$) \quad - proportionally fair
- $\alpha = 2$ \quad ($w_r = 1/T_r^2$) \quad - TCP fair
- $\alpha \to \infty$ \quad ($w = 1$) \quad - max-min fair
Example

max-min fairness:
\[ \alpha \to \infty \]

proportional fairness:
\[ \alpha = 1 \]

maximum flow:
\[ \alpha \to 0 \]

\[ n_r = 1, \quad w_r = 1 \quad r \in R, \]
\[ C_j = 1 \quad j \in J \]
Source: CAIDA - Young Hyun, Bradley Huffaker (displayed at MOMA)
Flow level model

Define a Markov process \( n(t) = (n_r(t), r \in R) \) with transition rates

\[
\begin{align*}
    n_r & \rightarrow n_r + 1 \quad \text{at rate} \quad \nu_r \quad r \in R \\
    n_r & \rightarrow n_r - 1 \quad \text{at rate} \quad n_r x_r(n) \mu_r \quad r \in R
\end{align*}
\]

- Poisson arrivals, exponentially distributed file sizes

Roberts and Massoulié 1998
Stability

Let \( \rho_r = \frac{\nu_r}{\mu_r} \quad r \in R \)

If

\[
\sum_r A_{jr} \rho_r < C_j \quad j \in J
\]

then the Markov chain \( n(t) = (n_r(t), r \in R) \) is positive recurrent

De Veciana, Lee & Konstantopoulos 1999;
Bonald & Massoulié 2001
Heavy traffic: balanced fluid model

The following are equivalent:

• $n$ is an invariant state

• there exists a non-negative vector $p$ with

$$n_r = \frac{\nu_r}{\mu_r} \sum_j A_{jr} p_j \quad r \in R$$

Thus the set of invariant states forms a $J$ dimensional subspace, parameterized by $p$.
Example

\( \mu_r = 1, \quad r \in R \)

Each bounding face corresponds to a resource not working at full capacity.

Entrainment: congestion at some resources may prevent other resources from working at their full capacity.
Williams (1987) determined sufficient conditions, in terms of the reflection angles and covariance matrix, for a SRBM in a polyhedral domain to have a product form invariant distribution – a skew symmetry condition.
Product form under proportional fairness

\[ \alpha = 1, \ w_r = 1, \ r \in R \]

Under the stationary distribution for the reflected Brownian motion, the (scaled) components of \( p \) are independent and exponentially distributed. The corresponding approximation for \( n \) is

\[ n_r \approx \rho_r \sum_j A_{jr} p_j \quad r \in R \]

where

\[ p_j \sim \text{Exp}(C_j - \sum_r A_{jr} \rho_r) \quad j \in J \]

Dual random variables are independent and exponential
Suppose a source-destination pair has access to several routes across the network:

\[ S \] - set of source-destination pairs
\[ r \in S \] - route \( r \) serves s-d pair \( s \)
Example of multipath routing

Routes, as well as flow rates, are chosen to optimize

$$\sum_{s} n_s \log(x_s)$$

over source-destination pairs $$s$$
First cut constraint

\[ n_1 x_1 + n_2 x_2 \leq C_1 + C_2 \]

Cut defines a single *pooled resource*
Second cut constraint

\[
\frac{1}{2} n_1 x_1 + n_3 x_3 \leq C_3
\]

Cut defines a second pooled resource
Product form

\[ \alpha = 1, \ w_r = 1, \ r \in R \]

In heavy traffic, and subject to some technical conditions, the (scaled) components of the shadow prices \( p \) for the pooled resources are independent and exponentially distributed. The corresponding approximation for \( n \) is

\[ n_s \approx \rho_s \sum_j p_j A_{js} \quad s \in S \]

where

\[ p_j \sim \text{Exp}(\overline{C}_j - \sum_s \overline{A}_{js} \rho_s) \quad j \in \overline{J} \]

Dual random variables are independent and exponential

Kang, K, Lee and Williams 2009
Outline

- Fairness in networks

- Rate control in communication networks (relatively well understood)

- Philosophy: optimization vs fairness

- Ramp metering (very preliminary)
What we've learned about highway congestion

Data, modelling and inference in road traffic networks
R.J. Gibbens and Y. Saatci
Phil. Trans. R. Soc. A366 (2008), 1907-1919.

Figure 2. The relationship between the speed and flow of vehicles observed on the morning of Wednesday, 14 July 2004 using the M25 midway between junctions 11 and 12 in the clockwise direction. In the free-flow regime, flow rapidly increases with only a modest decline in speeds. Above a critical occupancy of vehicles there is a marked drop in speed with little, if any, improvement in flow which is then followed by a severe collapse into a congested regime where both flow and speed are highly variable and attain very low levels. Finally, the situation recovers with a return to higher flows and an improvement in speeds.
A linear network

\[ m_i(t) = m_i(0) + e_i(t) - \int_0^t \Lambda_i(m(s)) \, ds, \quad t \geq 0 \]
Suppose the metering rates can be chosen to be any vector \( \Lambda = \Lambda(m) \) satisfying

\[
\sum_i A_{ji} \Lambda_i \leq C_j, \quad j \in J
\]

\[
\Lambda_i \geq 0, \quad i \in I
\]

\[
\Lambda_i = 0, \quad m_i = 0
\]

and such that

\[
m_i(t) = m_i(0) + e_i(t) - \int_0^t \Lambda_i(m(s)) \, ds \geq 0, \quad t \geq 0
\]
Optimal policy?

For each of $i = I, I-1, \ldots, 1$ in turn choose

$$\int_0^t \Lambda_i(m(s))\,ds \geq 0$$

to be maximal, subject to the constraints.

This policy minimizes

$$\sum_i m_i(t)$$

for all times $t$
Proportionally fair metering

Suppose \( \Lambda(m) = (\Lambda_i(m), i \in I) \) is chosen to maximize

\[
\sum_i m_i \log \Lambda_i
\]

subject to

\[
\sum_i A_{ji} \Lambda_i \leq C_j, \quad j \in J
\]

\[
\Lambda_i \geq 0, \quad i \in I
\]

\[
\Lambda_i = 0, \quad m_i = 0
\]
Proportionally fair metering

\[ \Lambda_i(m) = \frac{m_i}{\sum_j p_j A_{ji}}, \quad i \in I \]

where

\[ \Lambda_i \geq 0, \quad i \in I \]
\[ \sum_i A_{ji} \Lambda_i \leq C_j, \quad j \in J \]
\[ p_j \geq 0, \quad j \in J \]
\[ p_j \left( C_j - \sum_i A_{ji} \Lambda_i \right) \geq 0, \quad j \in J \]

\( p_j \) - shadow price (Lagrange multiplier) for the resource \( j \) capacity constraint
Brownian network model

Suppose that \((e_i(t), t \geq 0)\) is a Brownian motion, starting from the origin, with drift \(\rho_i\) and variance \(\rho_i \sigma^2\). Let

\[
X_j(t) = \sum_i A_{ji} e_i(t) - C_j t
\]

Then \(X(t) = (X_j(t), j \in J)\) is a \(J\)-dimensional Brownian motion starting from the origin

with drift \(-\theta = A\rho - C\)

and variance \(\Gamma = \sigma^2 A[\rho] A'\)
Brownian network model

Let \( W = A[\rho]A'R_+^J \)

and \( W^j = \{ A[\rho]A': q \in R_+^J, q_j = 0 \} \).

Define \( W(t) \) by the following relationships:

(i) \( W(t) = X(t) + U(t) \) for all \( t \geq 0 \)

(ii) \( W \) has continuous paths, \( W(t) \in W \)

(iii) for each \( j \in J, U_j \) is a one-dimensional process such that

(a) \( U_j \) is continuous and non-decreasing, with \( U_j(0) = 0 \),

(b) \( U_j(t) = \int_0^t I\{W(s) \in W^j\} dU_j(s) \) for all \( t \geq 0 \).
Brownian network model

If $\theta_j > 0$, $j \in J$, then there is a unique stationary distribution $W$ under which the components of

$$Q = (A[\rho]A')^{-1}W$$

are independent, and $Q_j$ is exponentially distributed with parameter

$$\frac{\sigma^2}{2} \theta_j, \quad j \in J$$

and queue sizes are given by

$$M = [\rho]A'Q$$
Delays

Let 

\[ D_i(m) = \frac{m_i}{\Lambda_i(m)} \]

- the time it would take to process the work in queue \( i \) at the current metered rate. Then

\[ D_i(M) = \sum_j Q_j A_{ji} \]
A tree network
A tree network

\[ Q_1 + Q_4 + Q_5 \]
Route choices
Route choices
Route choices

\[ Q_1 + Q_2 + Q_3 + Q_4 \]
Route choices

\[ Q_2 \sim \frac{\sigma^2}{2} \exp(C_1 + C_2 - \rho_1 - \rho_2) \]

\[ Q_1 + Q_2 + Q_3 + Q_4 \]
Final remarks

• Often the networks we design are part of a larger system, where agents are optimizing their own actions

• Sharing resources fairly may, in certain circumstances, lead to near optimal behaviour of the larger system, if it exposes agents to appropriate shadow prices

• The proportional fairness criterion can give the appropriate shadow prices