Optimal Routes and Flows in Congestion Constrained Ad Hoc Networks

Daniel C ONeill, David Julian and Stephen Boyd
Department of Electrical Engineering
Stanford University
dconeill,djulian,boyd@stanford.edu

Abstract— **Future ad hoc and multi-hop networks will simultaneously support many different types of traffic such as streaming video, voice, and data. This is particularly true for emerging 802.11, bluetooth, and other wireless technologies expected to support ubiquitous internet access. Supporting different traffic types requires the network to find the best set of routes from sources to sinks, individual link data rates and individual link transmitter powers, all subject to QoS constraints.**

The paper address this problem. The approach, based on Perron Frobenius matrix theory, yields Pareto optimal values for these system variables and offers a new view of system capacity and cost in terms of associated eigenvalues. Performance metrics can be utility functions or other protocol performance measures. A simple method of solution, the DSM algorithm, is presented. The DSM is iterative and adaptive, responding to changes in the wireless environment by automatically seeking a new set of optimal system variables. Mathematical programming/optimization

Index Terms— **Mathematical programming/optimization, ad hoc networks, multihop networks, power control, routing, Perron Frobenius, utility functions, congestion control.**

I. Introduction

TO fulfill the promise of ubiquitous communication, future ad hoc networks will need to simultaneously support several different types of traffic such as streaming video, voice, and data. Supporting a wide array of traffic types will require the network to find the best set of routes from sources to sinks and the associated data rates along these paths. QoS constraints on individual flows, such as minimum rate and average queueing delay must also be supported. In a variable rate/variable power system the network must translate these characteristics into a set of supporting link rates and transmitter powers while accounting for inter-link interference. Collectively the set of network source rates, paths, link rates, and link transmitter powers selected by the network is termed the operating point of the system.

This paper addresses the problem of finding the optimal operating point for a multi-hop or ad hoc network simultaneously conveying different types of traffic, and subject to limits on network congestion and other QoS constraints . Optimality is measured from a utility function or protocol performance metric point of view. The results, based on Perron Frobenius matrix theory, are Pareto optimal; performance can not be improved for one source with out decreasing it for some other source. The underlying analysis expresses network capacity in terms of the Perron Frobenius eigenvalue with implications for network pricing. The DSM method of solution is adaptive and responds to changes by seeking a new optimum.

The problem of finding optimal transmitter powers for a given set of link rates and link interference gains has been extensively discussed in the literature [1] [2] [3]. More recently, methods to jointly find the best set of link rates for a mix of different types of traffic over a single hop network have been described [4]. Likewise different methods for routing packets in ad-hoc networks have been investigated [5], [6], [7]. This paper combines these objectives by jointly finding the best set of link rates, data source rates, transmitter powers, and routes through the network. It extends our work in [4], which uses Perron Frobenius theory to find

the best set of rates and powers from single hop to multi-hop networks.

The paper divides the problem into three phases. In the first phase the network is abstracted to a feasible rate region, and optimal transfer rates and routes are found for a particular mix of traffic types. Different traffic types can have different performance metrics or protocols. QoS constraints on congestion or minimum link or path rates can be included as can a variety of other convex constraints. In the second phase this abstraction is used to determine the optimal link transmitter rates and powers that meet QoS requirements. In the third phase an iterative method is described that adapts to changes in the ad-hoc network.

The paper is organized into several sections. Sections II and III define an ad-hoc network, notation, the transmission model used and set of feasible routes. Section IV defines the rate region, and shows it is convex. Section V defines performance metrics for the network. These metrics can be different for different flows or protocols. Methods to control congestion are also discussed. Sections VI through VIII analyze the problem and Section IX presents a method of solution. The method is intuitive and adaptive. Section X presents simulation results that illustrate the approach.

II. Notation

The following is a list of network variables used in this paper.

- l : Wireless links numbered $l = 1, \cdots, L$.
- R_l : Physical data rate on link l.
- s : Data sources numbered $s = 1, \cdots, S$. Each source is associated with a single sink.
- k: Path index for source s, $k = 1, \cdots, K_s$.
- $r_{(s,k)}$: Transfer rate from source s along path k.
- r_s : Vector of transfer rates for source s, $r_s = [r_{(s,1)} r_{(s,2)} \cdots r_{(s,K_s)}]^T$.
- r : Vector of transfer rates, $r = [r_1^T r_2^T \cdots r_S^T]^T$.
- $\theta(s,k)$ Set of links used by source s on its kth path from source to destination.
- $\phi(l)$: Set of source-path pairs, (s,k) traversing link l.
- $\Lambda(s)$: Set of paths from source s to its destination.
- A : Routing matrix describing global set of routes.
- ρ_l : Signal to Interference ratio of lth link.
- G_{ij} : Gain from transmitter on link j to receiver in link i.
- p_l : Transmitter power on lth link.
- B : Path aggregation matrix linking source rates to transfer rates.

III. Ad-hoc Wireless Network

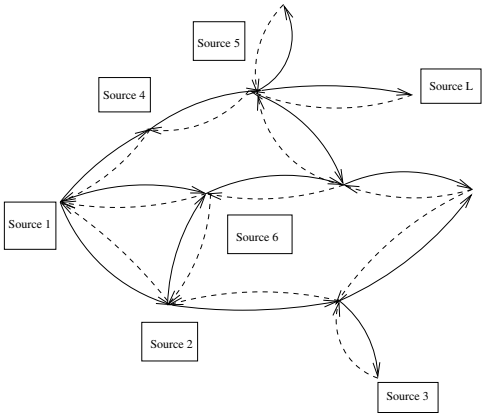

Fig. 1. Schematic of ad-hoc network with L links and S data sources. Solid lines are forward links and dashed lines reverse links.

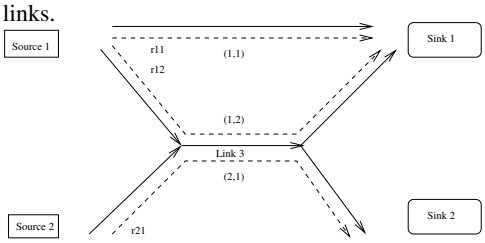

Fig. 2. Path of transfer rates. Source 1 can send packets at rate $r_{(1,1)}$ over path $(1,1)$ and at rate $r_{(1,2)}$ over path $(1,2)$. Source 2 can send only over the single path shown and at rate $r_{(2,1)}$

An ad-hoc wireless network is composed of L unidirectional links and S sources. Figure 1 depicts such a network. Solid lines are forward links and dashed lines reverse links. All links share a common bandwidth and utilize variable rate and variable power CDMA technology. The transmission rate on link l is $R_l > 0$. Each data source $s \in S$ injects packets into the network at aggregate rate $\mathbf{1}^T r_s$ termed the source rate. These packets traverse the network along one or more paths defined by the set $\Lambda(s)$ at rate $r_{(s,k)} \geq 0$, where $\sum_k r_{(s,k)} = \mathbf{1}^T r_s$.

Each path is composed of one or more links or hops. Thus a packet may cross several links along a path when travelling from source to sink . The $r_{(s,k)}$ are termed transfer rates, since this is the rate that source s transfers packets along a particular path $(s,k) \in \Lambda(s)$.

Figure 2 illustrates the flow of packets. Source 1 can send packets at rate $r_{(1,1)}$ over path $(1,1)$ and at rate $r_{(1,2)}$ over path $(1,2)$. Source 2 can only send packets over the single path shown at rate $r_{(2,1)}$. Traffic from both sources are routed across a single link and share the link's capacity. The link's data rate R_3 must be greater than the rate at which the data sources send traffic over the link $r_{(1,2)} + r_{(2,1)} \leq R_3$. Two or more paths associated with a single source can also traverse a single link. In this case the paths diverge at some point between source to sink.

The link rates R_l are assumed to be functions of the Signal to Interference Ratio. The link SIR , ρ_l, is defined as

$$\rho_l = \frac{G_{ll} p_l}{\sum_{j \neq l} G_{lj} p_j}. \tag{1}$$

G_{ll} represents the effective gain between the transmitter and receiver on link l and includes the multiplicative spreading gain, antenna gain, coding gain, and other gain factors. Likewise G_{lj} represents the effective gain from the interfering transmitter on link j to the receiver on link l. The gains $G_{ij} > 0$ are assumed to be positive.

The wireless network is assumed to be interference limited and therefore receiver noise can be neglected to a first approximation. Link SIR, ρ_l, is consequently invariant to power scalings. By choosing $\mathbf{1}^T p = 1$ the p_l can be interpreted as the relative transmitter powers or equivalently as the percent of total transmitted power in the system.

An empirically based link rate model for M-QAM and M-PSK modulation is [8]

$$R_l = \alpha \log(1 + K\rho_l), \tag{2}$$

where $K = (-1.5)/(\ln(5\mathrm{BER}))$ and BER is the target average bit error rate. The constant α is a scaling constant and represents the base of the logarithm used and several other system constants.

This is similar in form to the information theoretic capacity model,

$$C_l = W \log(1 + \rho_l). \tag{3}$$

For reasons that will become clear later, the link rate model used in this paper is a simplified version of equation 2:

$$R_l = \log(\rho_l). \tag{4}$$

The constant K is absorbed into G_{ll}, and α is taken as equal to 1, since this constant has the effect of scaling all rates equally. The assumption $\rho_l \geq 1$ is necessary to prevent negative rates. In most systems $\rho_l \gg 1$ since it represents the effective SIR after spreading gain, antenna gain, and coding gain.

The traffic intensity carried over link l, is necessarily smaller than the link rate

$$\sum_{(s,k) \in \phi(l)} r_{(s,k)} \leq R_l \tag{5}$$

where $\phi(l)$ is the set of source-path pairs (s,k) that traverse a link. This can be rewritten as $a_l^T r \leq R_l$ where a is a binary vector. A one in the (s,k)th place corresponds to source-path pair (s,k) travelling over link l. For example in a network with two sources each with two possible paths, the vector $a_1^T = [1\ 1\ 0\ 1]$ corresponds to source 1 using link 1 on its $(1,1)$ and $(1,2)$ paths from source to sink and source 2 using link 1 only on path $(2,2)$ The remaining source-path $(2,1)$ does not use this link. In matrix form a set of transfer rates are feasible for the system if

$$Ar \leq R \tag{6}$$

where

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_L^T \end{bmatrix} \tag{7}$$

is the routing matrix for the system and inequality is taken element-wise. Transfer rates violating this constraint can't be supported by the network. The columns of A, $\tilde{a}$, likewise describe the set of links traversed by a by source s along its kth path.

Combining equations 4 and 6 yields

$$\begin{aligned} Ar \leq R \quad &\Leftrightarrow \quad p_l \geq \frac{e^{a_l^T r_l}}{G_{ll}} \sum_{j \neq l} G_{lj} p_j \\ &\Leftrightarrow \quad p \geq D(Ar)\tilde{G}p \end{aligned} \tag{8}$$

where $D \triangleq \text{diag}\left(\frac{e^{a_l^T r_l}}{G_{ll}}\right)$ and

$$\tilde{G}_{ij} = \begin{cases} G_{ij}, & i \neq j \\ 0, & i = j. \end{cases} \quad (9)$$

IV. Ad-hoc Network Rate Region

In an Ad-hoc network a vector of transfer rates r is feasible if it is possible for the system to simultaneously support all of these rates. The rate region is the set of feasible transfer rates and can change with changing topologies or routing. Formally the rate region is

$$\mathcal{R} = \{r \in \mathbf{R}_+^n | Ar \leq R(p) \text{ for some } p\}, \quad (10)$$

where $Ar \leq R$ is taken as component-wise, i.e. $a_l^T r \leq R_l$ for all l.

Theorem 1: The rate region $\mathcal{R}$ is convex.

Proof: The set of feasible transfer rates and power pairs, $\mathcal{M}$, is the set of (r, p) such that $a_l^T r \leq \log(\rho_l)$ for all links l. Analytically,

$$\begin{aligned} \mathcal{M} &= \{(r,p) \in \mathbf{R}_+^{2n} | a_l^T r \leq \log(\rho_l), \forall l\} \\ &= \textstyle\bigcap_l \{(r,p) \in \mathbf{R}_+^{2n} | a_l^T r \leq \log(\rho_l)\} \\ &= \textstyle\bigcap_l \mathcal{M}_l. \end{aligned} \quad (11)$$

The $\mathcal{M}_l = \{(r,p) \in \mathbf{R}_+^{2n} | a_l^T r \leq \log(\rho_l)\}$ are convex. This can be seen by the change of variables $x_l = \log p_l$ and rewriting the set qualifier as follows:

$$a_l^T r \leq \log(\rho_l) \quad \Leftrightarrow \quad e^{-a_l^T r} \geq \rho_l^{-1} \quad (12)$$

$$\begin{aligned} &\Leftrightarrow \quad e^{-a_l^T r} \geq \textstyle\sum_{j \neq l} G_{lj} e^{x_j} G_{ll}^{-1} e^{-x_l} \\ &\Leftrightarrow \quad 1 \geq \textstyle\sum_{j \neq l} G_{lj} e^{x_j} G_{ll}^{-1} e^{-x_l} e^{-a_l^T r} \\ &\Leftrightarrow \quad 0 \geq \log(\textstyle\sum_{j \neq l} G_{lj} e^{x_j} G_{ll}^{-1} e^{-x_l} e^{-a_l^T r}). \end{aligned} \quad (13)$$

It is known [9] that the function $\log(\sum \alpha_l e^{y_l})$, for $\alpha_l \in \mathbf{R}_+$ and $y_l \in \mathbf{R}$, is convex in y. Composition with an affine function preserves convexity. Sub-level sets of convex functions always define convex sets, so equation 12 defines a convex set in the variables $\log p_l$ and r_l. Since the intersection of convex sets is convex, $\mathcal{M}$ must also be convex. ■

The rate-region $\mathcal{R}$ is the projection of $\mathcal{M}$ onto the transfer rate space. Linear projection also conserves convexity so the rate region is convex. An example of the rate region is shown in Figure 3.

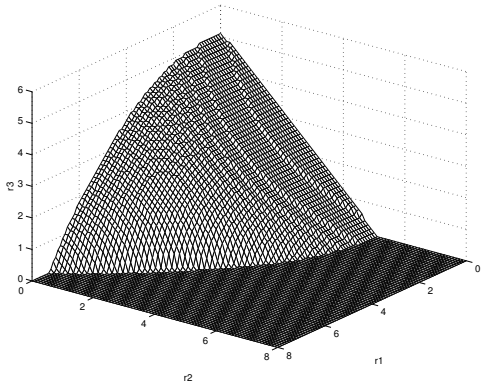

Fig. 3. Ad-hoc Network Rate Region

V. Performance Metrics

System performance is modelled by a performance metric or utility function U. The utility function is assumed to be a function of the total source rates but not the individual transfer rates,

$$U = U(Br) \quad (14)$$

where B is the path aggregation matrix. The matrix B aggregates the set of source-path flows for each source to yield an overall source rate. This formulation measures the ad-hoc network's performance at its input/output ports. This is the performance as seen by users of the system or by end to end network protocols. The routes or paths taken by packets from source s to its sink are treated as internal network decisions that do not directly effect performance as measured at the input/output ports. Similarly, transmitter link powers are treated as network parameters to be adjusted by the system, but which in and of themselves, do not measure system performance. Indirectly, both path selection and power control effect the performance of the system through the feasible transfer rates, and thus effect U through r. By assumption, a higher data rate is valued at least as much as a lower data rate, so U is a non-decreasing function of r. Also, by assumption, there is a diminishing return to additional data rate, so let U be a concave function of r. The performance metric can be expanded to directly include measures of routing and transmitter powers, although this is not done in this paper due to space limitations.

In many situations U can be treated as a separable function,

$$U(Br) = \sum_{s=1}^{S} U_s(\mathbf{1}^T r_s) \tag{15}$$

where each logical link in the network has a performance metric $U_s(\mathbf{1}^T r_s)$. In this view the performance measured by one logical link is not directly affected by the performance on another logical link. Each $U_s(\mathbf{1}^T r_s)$ can represent the utility a user derives from using the system at a particular rate, or can be implied by data protocols, service rate agreements, or other system level metrics [10], [11], [12].

A single voice link might have a protocol that requires a minimum r_s but is indifferent to rates above this rate. An appropriate function is the following:

$$U_s(\mathbf{1}^T r_s) = \begin{cases} -\infty, & \mathbf{1}^T r_s < R_{\min_i} \\ c, & \mathbf{1}^T r_s \geq R_{\min_i}. \end{cases} \tag{16}$$

A wireless Internet user might benefit from an increased r_j and be willing to pay more for this service. A possible link metric is then $U_j(r_j) = \alpha r_j$, where implicitly the user pays more for more bandwidth.

A. QoS

QoS requirements involving r can be addressed in several ways. The simplest is to embed them in U. For example, a minimum rate requirement for user s can be closely approximated by rewriting the performance metric for s as

$$U(r_s)_{\text{new}} = U(r_s) + \gamma \ln(\mathbf{1}^T r_s - r_{th}) \tag{17}$$

where r_{th} is the rate threshold and $\gamma << 1$ is a constant chosen to control the tightness of the threshold. A second example is video, where a user has a required rate to ensure a minimum level of picture quality, but where higher rates are valued according to his personal utility function. Assuming a logarithmic utility function

$$U_s(r_s) = \begin{cases} -\infty, & r_s < R_{\min_i} \\ \alpha \log(r_s) + b, & r_s \geq R_{\min_i}. \end{cases} \tag{18}$$

A second more general approach is to express the QoS requirements as a modified rate region. Complex inter-link requirements can be handled in this way.

B. Congestion control

Congestion is an important parameter in QoS. This is particularly true for delay sensitive data such as video or real time information. Unfortunately analytically modelling the congestion in a network is difficult, and Markovian approximations are often used.

This convention is followed here. Arrivals to the network are assumed to be Poisson(λ) and transmission times across links (service times) are assumed to be distributed exponentially(u). By *Jackson's Theorem* [13] the number of packets waiting or in transmission is

$$E[N] = \frac{\lambda}{u - \lambda} \tag{19}$$

where λ and u are the arrival and service rates and N is the number of packets in the system.

A bound d on congestion can then be expressed as

$$\begin{aligned} E[N] \leq d \quad &\Leftrightarrow \quad \tfrac{\lambda}{u-\lambda} \leq d \\ &\Leftrightarrow \quad \tfrac{1+d}{d}\lambda \leq u \\ &\Leftrightarrow \quad D\lambda \leq u, \end{aligned} \tag{20}$$

where $D = (1 + d)/d$.

The associated queue's arrival rate λ_l is the sum of the transfer rates $a_l^T r$ traversing a link, $\lambda_l = a_l^T r$. Similarly, the service rate u_l is the link rate R_l, $u = R_l$. Thus a QoS bound on average congestion is

$$\begin{aligned} \mathbf{diag}(D_l) A r &\leq R \\ \tilde{A} r &\leq R. \end{aligned} \tag{21}$$

Equation 21 is identical in form to equation 6 and the subsequent analysis remains valid. The interpretations of the rate region $\mathcal{R}$, however, do change. The rate region $\mathcal{R}$ becomes the Congestion Limited Rate Region; the set of feasible rates that meet the congestion bound for the system. Figure 4 shows a Congestion Limited Rate Region corresponding to Figure 3. As intuition suggests the rate region is smaller. This is because transfer rates must be restricted to ensure average delay bounds can be met.

VI. Problem Formulation

The goal is to find the best set of paths from source to sink, transfer rates and transmitter powers such that system performance is maximized. The

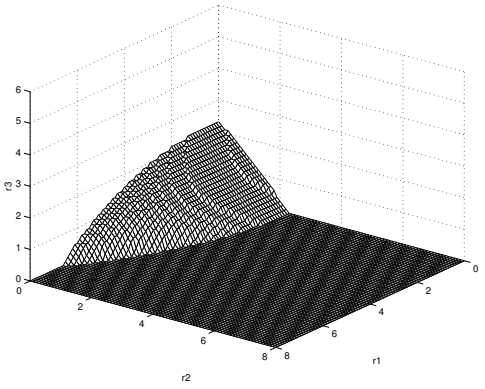

Fig. 4. Ad-hoc Network Congestion Constrained Rate Region. QoS constraints reshape the feasible rate region.

optimal transfer rates are the rates along each path from each source to its destination. A zero transfer rate corresponds to an unused path. A nonzero rate selects the associated path. At optimum a source may utilize one or more paths through the network depending on system typology, interference, and path choices. Formally this can be expressed as

$$\begin{array}{ll} \text{maximize}_r & U(Br) \\ \text{subject to} & r \in \mathcal{R}. \end{array} \tag{22}$$

This is a convex optimization problem in the variable r since the constraint set is convex and the objective is concave. The vector of transmitter powers p is not explicitly used in this formulation. The inter-link interference is captured in the shape of the rate region, as are QoS constraints on transmitter power. As will be subsequently shown, the optimal transmitter powers can be calculated from the optimal transfer rates.

The optimal system operating point will change with changing A, B, or G. Section IX describes a method which adapts to these changes by continuously seeking the best r and p. The method is based on the *Perron Frobenius* theory of positive matrices.

VII. Analysis

A. Perron Frobenius

For a square matrix M the notation $M > 0$ means M is an element-wise positive matrix. The eigenvalue of M with greatest magnitude is the Perron-Frobenius eigenvalue $\lambda_{\text{pf}}(M)$. If the matrix $M > 0$ is regular, meaning that $(M^k)_{ij} > 0$ for all i, j and some positive integer k, then λ_{pf} is strictly positive and unique, and the associated right and left eigenvectors $p > 0$ and $q > 0$ are strictly positive. If $\lambda_{\text{pf}}(M)$ is the Perron-Frobenius eigenvalue for regular $M > 0$, then the inequality $\beta p \geq Mp$ has a feasible $p > 0$ if and only if $\lambda_{\text{pf}}(M) \leq \beta$. Finally, for any positive matrix, the monotone property states if $M_{ij} \leq N_{ij}$ for all i, j with strict inequality for at least one i, j then $\lambda_{\text{pf}}(M) < \lambda_{\text{pf}}(N)$.

Since D is a function of r, in what follows the Perron Frobenius eigenvalue associated with the network is written as $\lambda_{\text{pf}}(D(Ar)\tilde{G})$.

In addition, the following facts are used in what follows.

Theorem 2: The Perron Frobenius eigenvalue of the matrix $M > 0$ is differentiable in m_{ij} and can be found from

$$\frac{\partial \lambda(M)}{\partial m_{ij}} = q^T \frac{\partial M}{\partial m_{ij}} p \tag{23}$$

where q and p are the left and right eigenvectors associated with λ_{pf} and $q^T p = 1$. The proof is an application of *Gerschgorin's* Theorem [14].

By assumption $G > 0$ and consequently $D(Ar)\tilde{G}$ is irreducible and $\lambda_{\text{pf}}(D(Ar)\tilde{G})$ is unique, so the Theorem applies.

Theorem 3: The Perron Frobenius eigenvalue $\lambda_{\text{pf}}(D(Ar)\tilde{G})$ is convex in r. The proof can be found in [4].

B. Pareto Surface

A point $r \in \mathcal{R}$ is *Pareto optimal* for the rate-region $\mathcal{R}$ if there does not exist another point $r' \in \mathcal{R}$ that dominates r. A point $r' \in \mathcal{R}$ *dominates* r if $r'_i \geq r_i$ for all i, and $r'_j > r_j$ for some j. The notation $r' \succeq_d r$ means r' dominates r. The Pareto surface is defined as the set of Pareto optimal points,

$$\mathcal{P} = \{r \in \mathcal{R} | \nexists r' \in \mathcal{R} \text{ s.t. } r' \succeq_d r\}. \tag{24}$$

Theorem 4: An optimal operating point r for the system always lies on $\mathcal{P}$.

Proof: The proof is by construction. Assume there exists an optimal operating point $r^* \notin \mathcal{P}$, then there exists a feasible $r' \in \mathcal{P}$ with $r'_i \geq r^*_i \quad \forall i$. So by the non-decreasing property of U,

$U(r') \geq U(r^*)$. By optimality of r^*, $U(r^*) \geq U(r)$ $\forall$ feasible r, so $U(r') = U(r^*)$ and $r' \in \mathcal{P}$ is also optimal. ■

Theorem 5: At optimality $p = D(Ar)\tilde{G}p$.

Proof: The proof is by contradiction. Assume $p \geq D(Ar)\tilde{G}p$ at optimality. Then $\exists$ an l such that

$$\begin{aligned} a_l^T r &< R_l \\ &= \ln(\tfrac{G_{ll}p_l}{\sum_{j\neq l} G_{lj}p_j}). \end{aligned} \tag{25}$$

Thus the power on the lth link p_l could be decreased with out effecting feasibility. But decreasing p_l would increase ρ_i for some $i \neq l$, since $G > 0$. Thereby finding a new set of transfer rates $\acute{r} > r^*$, contradicting the optimality assumption. ■

Transfer rates r lying in $\mathcal{P}$ can be expressed as the following:

$$\mathcal{P} = \{r|p = D\tilde{G}p\} = \{r|\lambda_{\rm pf}(D(Ar)\tilde{G}) = 1\}, \tag{26}$$

where $\lambda_{\rm pf}(D(Ar)\tilde{G}) = 1$ is the surface of the rate region. Note, $\lambda_{\rm pf}(D(Ar)\tilde{G}) > 1$ corresponds to a point outside of the rate region, i.e. that can not be achieved by the system for any set of powers, and $\lambda_{\rm pf}(D(Ar)\tilde{G}) < 1$ corresponds to a point inside the rate region, i.e. that can be achieved but are no better than those on $\mathcal{P}$.

C. Normal to the Pareto surface

Equation 26 describes the Pareto surface as a level set of the Perron Frobenius eigenvalue, $\lambda_{\rm pf}(D(Ar)\tilde{G}) = 1$. The normal $M(r)$ to the Pareto surface is then

$$\begin{aligned} M(r) &= \nabla_r \lambda_{\rm pf}(D(Ar)\tilde{G}) \\ &= A^T \nabla_x \lambda_{\rm pf}(D(x)\tilde{G})\Big|_{x=Ar}. \end{aligned} \tag{27}$$

The gradient of $\lambda_{\rm pf}$ can be found from Theorem 2. For the (s,k)th component

$$\begin{aligned} \frac{\partial \lambda_{\rm pf}(D(Ar)\tilde{G})}{\partial r_{(s,k)}} &= q^T \frac{\partial D(Ar)\tilde{G}}{\partial r_{(s,k)}} p \\ &= q^T \frac{\partial \text{diag}(e^{a_1^T r}/G_{11}, \ldots)\tilde{G}}{\partial r_{(s,k)}} p \\ &= \sum_{l\in\theta(s,k)} q_l e^{a_l^T r} \frac{I_l}{G_{ll}}. \end{aligned} \tag{28}$$

where $I_l = \tilde{G}p$ is the vector of interference values for the system. At optimality $p = D(Ar)\tilde{G}p$, or equivalently $e^{a_l^T r} = e^{R_l} = \rho_l$ so

$$\frac{\partial \lambda_{\rm pf}(D(Ar)\tilde{G})}{\partial r_{(sk)}} = \sum_{i\in\theta(s,k)} q_i p_i. \tag{29}$$

In matrix form the normal $M(r)$ can be rewritten as

$$\begin{aligned} M(r) &= \nabla_r \lambda_{\rm pf}(D(Ar)\tilde{G}) \\ &= A^T[q_1 p_1, \ldots q_L p_L]^T \\ &= A^T N(r)^T. \end{aligned} \tag{30}$$

The interference term I is missing in equation 30. The effect of interference is instead captured by the left eigenvalue q, $q_i < 1$. The components of the left eigenvector q_l scale the associated link transmitter powers p_l, similar to the way interference scales transmitter power in the SIR term ρ_l. The product $p_l q_l$ can be thought of as a normalized equivalent transmitter power.

The constraint $\lambda_{\rm pf}(D(Ar)\tilde{G}) \leq 1$ can be interpreted as a measure of system capacity utilization. When $\lambda_{\rm pf}(D(Ar)\tilde{G}) = 1$, the system is operating at 100 percent capacity; no transfer rate r_i can be increased without decreasing some other transfer rate r_j, $i \neq j$. When $\lambda_{\rm pf}(D(Ar)\tilde{G})$ is less than one, additional capacity is available in the system. This might happen for example if the system is operated at a set of Pareto suboptimal fixed rates. A multi-hop network dominated by one fixed, high rate user and many slower users is an obvious example. As modelled here, because the utility function U is concave and increasing the system will always be driven to full capacity.

The constraint $\lambda_{\rm pf} \leq 1$ can be thought of as a kind of "capacity constraint" on the network. The network pays for a small increase in transfer rates in $\lambda_{\rm pf}$ dollars. The system is free to spend up to $\lambda_{\rm pf} = 1$, 100% of its capacity, but no more. The term $p_l q_l$ is the marginal cost in system capacity associated with increasing a transfer rate traversing the link. From this point of view equation 30 is intuitively appealing. The marginal cost of increasing source transfer rate $r_{s,k}$ is the sum of the marginal costs associated with all links the flow crosses.

$$c_{s,k} = \sum_{l\in\theta(s,k)} q_l p_l \tag{31}$$

Thus a flow pays a marginal "toll" of $q_l p_l$ when crossing link l. This holds irrespectively of the source or destination of a flow, or the sequence of

links a flow traverses. If r_{sk} is increased, the cost in capacity is

$$\Delta\lambda_{\rm pf}(D(Ar)\tilde{G}) \;=\; c_{s,k}\Delta r_{sk}. \qquad (32)$$

VIII. Optimality Criterion

Because the optimal values of r lie on the Pareto surface $\mathcal{P}$ the network optimization problem can be restated as

$$\begin{array}{ll} \text{maximize}_r & U(r) \\ \text{subject to} & \lambda_{\rm pf}(D(Ar)\tilde{G}) = 1 \\ & r > 0. \end{array} \qquad (33)$$

By *Lagrange's Theorem*[15], at optimality

$$\nabla_r U(Br) = K\nabla_r\lambda_{\rm pf}(D(Ar)\tilde{G}) \qquad (34)$$

where K is a constant of proportionality. In words, the vectors ∇U and $\nabla\lambda_{\rm pf}$ must be parallel at the optimal rates. This is depicted in Figure 5. The gradient ∇U is normal to the level sets of U and $N(r) = \nabla\lambda_{\rm pf}$ is normal to the rate surface. At optimality these normals align and the level set surface for the performance metric is tangent to the rate region.

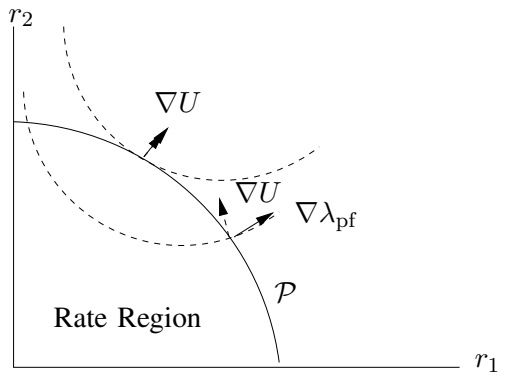

Fig. 5. Optimality condition is $\nabla\lambda_{\rm pf}$ is parallel to $\nabla U(r)$.

The proportionality constant K can be shown to be the ratio of marginal performance to marginal capacity, a measure of system efficiency.

$$\begin{array}{rcl} K & = & \frac{\mathbf{1}^T\nabla U(Br)}{A^T N(r)^T} \\ & = & \frac{\text{Total marginal performance gain}}{\text{Total marginal capacity cost}} \end{array} \qquad (35)$$

IX. Optimal Operating Point

This section describes the Direct Step Method, DSM, for finding the optimal operating point for the system [4]. In so doing the approach calculates the optimal paths from each source to its associated sink, optimal transfer rates, and optimal transmitter powers. Some of the transfer rates are typically zero. The transmitter powers are calculated from $p = D(Ar)\tilde{G}p$. An alternative is the method proposed by G. Foschini and Z. Miljanic [1]. The method is adaptive and continuously seeks the best r for the system.

DSM operates on the Pareto surface $\mathcal{P}$. From an initial point $r_t \in \mathcal{P}$ the method calculates a nearby point r_{t+1} that improves system performance. The process is repeated until an optimal operating point r^*_τ is achieved. The method continues to iterate with $r_t = r_\tau\ t \geq \tau$. Since the feasible operating region is convex and the performance metric concave, a locally optimal r will also be globally optimal. If the system is subsequently disturbed, then the system adapts by again seeking a locally better operating point r and the process continues. Figure 6 depicts the process.

A. Feasible ascent direction

The Direct Step Method is shown in Figure IX Let $M(r)^\perp = \left\{r' | (r-r')^T N(r) = 0\right\}$ be the hyper-plane that is tangent to the Pareto surface at r_c. Since the rate-region $\mathcal{R}$ is convex, $M(r)^\perp$ is a supporting hyper-plane and lies outside of $\mathcal{R}$, except at the point r_c. The supporting hyper-plane $M(r)^\perp$ is a good approximation of $\mathcal{P}$ for small changes in r. For this reason a direction δr is defined to be *feasible* if it lies along $N(r)^\perp$, or equivalently $\delta r^T N(r) = 0$.

A small change δr is defined as an *ascent* direction if $U(r+\alpha\delta r)$ increases for small $\alpha > 0$. Thus, δr is an ascent direction if and only if $\nabla U(r)^T\delta r > 0$. A point that is both feasible and an ascent direction is termed a *feasible ascent* direction.

B. DSM

The DSM is a two phase feasible ascent method. First, in the predictor phase a small feasible change or step δr is calculated. Next, in the corrector phase this point is corrected to lie along $\mathcal{P}$. The method

can be described as a simple two step algorithm. Given the current operating point $r_c(t)$

Algorithm 1:
- Calculate a feasible ascent direction $/deltar$ and predict a new operating point $r_p(t+1) = r_c(t) + \beta\delta r$.
- Correct this estimate by scaling it onto $\mathcal{P}$, $r_c(t+1) = \alpha r_p(t+1)$. Repeat.

End Algorithm

1) DSM: Predictor: The DSM constructs a δr on $M(r)^{\perp}$ from a measure of the sub-optimality of the system. The error estimate is defined as

$$e = \left(\frac{A^T N(r)}{\mathbf{1}^T A^T N(r)} - \frac{\nabla U}{\mathbf{1}^T \nabla U} \right) \tag{36}$$

and compares the normal to the rate region to the normal to the performance metric. At optimality $e = 0$ and the operating point r remains fixed.

Because U is concave and $\lambda_{\rm pf}$ is convex, a rate change δr_i causes the performance metric normal and rate region normal to respond in opposite ways; an increase $\delta r_{{\rm ds},sk} > 0$ causes the (s,k)th component of $\frac{\nabla U}{\mathbf{1}^T \nabla U}$ to decrease and the comparable component of $\nabla \lambda_{\rm pf}$ to increase. Consequently, the decision to increase the (s,k)th component of $\delta r_{\rm ds}$ can be made by comparing the two normals. If $\frac{\nabla U}{\mathbf{1}^T \nabla U}$ is greater than $\nabla \lambda_{\rm pf}$, then the associated rate should be increased. Specifically, for small rate adjustments $\delta r_{\rm ds}$ should have the same sign as $-e$. The direct step method uses this information to find a $\delta r_{\rm ds}$ that is an ascent direction but which is also feasible by construction, that is lies on $N(r)^{\perp}$. Specifically,

$$\begin{aligned} \delta r_{\rm ds} &= -\mathbf{diag}(\tfrac{1}{\sum_{l \in \theta(s,k)} q_l p_l}, \ldots) e \\ &= -\mathbf{diag}(\tfrac{1}{\tilde{a}_{(s,k)}^T N(r)}) e \end{aligned} \tag{37}$$

where $\tilde{a}_{(s,k)}$ are the columns of A.

Substituting e yields

$$\begin{aligned} \delta r_{{\rm ds},({\rm sk})} &= -\left(\tfrac{1}{\tilde{a}_{(s,k)}^T N(r)} \right) \\ &\quad \left(\tfrac{A^T N(r)}{\mathbf{1}^T A^T N(r)} - \tfrac{\nabla U}{\mathbf{1}^T \nabla U} \right)_{(s,k)} . \end{aligned} \tag{38}$$

That $\delta r_{\rm ds}$ lies on the supporting hyper-plane can be seen from

$$\begin{aligned} M(r)^T \delta r_{\rm ds} &= \textstyle\sum \tilde{a}_{(s,k)}^T \left(\tfrac{e_{(s,k)}}{\tilde{a}_{(s,k)}^T} \right) \\ &= \textstyle\sum e_{(s,k)} \\ &= \mathbf{1}^T e \\ &= \mathbf{1}^T \tfrac{M(r)}{\mathbf{1}^T M(r)} - \mathbf{1}^T \tfrac{\nabla U}{\mathbf{1}^T \nabla U} \\ &= 0 \end{aligned} \tag{39}$$

A new rate is calculated as $r_p(t+1) = r_c(t) + \beta\delta r_{\rm ds}$, where $\beta << 1$. This rate lies along the supporting hyper-plane $M(r)^{\perp}$, but, unless this is the optimal operating point, is not on $\mathcal{P}$.

2) DSM: Corrector: The estimated rate $r(t+1)$ is corrected to lie on $\mathcal{P}$ using a scaling method. The scaling method scales each element in the estimated rate vector by a constant α_a. As shown in Figure 7, this represents a movement on a ray from the origin.

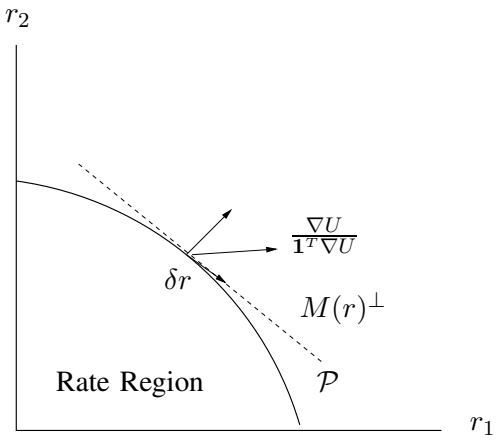

Fig. 6. DSM predictor phase. The optimality error e is used to predict a change rates δr that improves ad-hoc network performance.

The scaling method multiplies each element in the rate vector r_p by a fixed scalar $r_c = \alpha_p r_p$, $\alpha_p > 0$ to find a rate vector $r_c \in \mathcal{P}$. Increasing α increases all rates $r_c = \alpha r_p$, and in turn increases the elements of $D(r_c)\tilde{G}$. By the monotone property for the Perron Frobenius eigenvalue, $\lambda_{\rm pf}(D(\alpha r_p)\tilde{G})$ also increases and is monotonic in α. This leads to a bisection algorithm to find α_p.

The bisection algorithm increases α linearly until $\lambda_{\rm pf}(D(\alpha r_p)\tilde{G}) \geq 1$, so α_p lies between zero and α. Next $\lambda_{\rm pf}(D(\frac{1}{2}\alpha r_p)\tilde{G})$ is computed and compared with one; if it is greater than one then $\alpha_p \in [0, \alpha/2]$ while if it is less than one then $\alpha_p \in [\alpha/2, \alpha]$. If $\alpha_p \in [\alpha/2, \alpha]$, then $\lambda_{\rm pf}(D(\frac{3}{4}\alpha)\tilde{G})$ is computed and compared with one to again reduce the range containing α_p by half. The segment that α_p lies in is reduced through repeated bisections until α_p is known to the desired number of decimal points.

C. Adaptation

The DSM can adapt to changes in system parameters G, the number of sources, and the routing

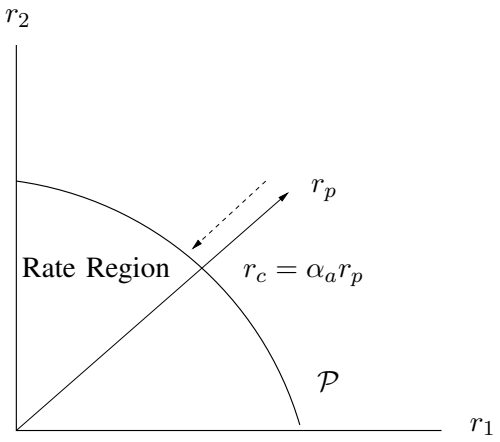

Fig. 7. DSM Corrector Phase: Moving to the rate surface.

matrix A. If a disturbance changes the system and the current operating point is no longer optimal, the error vector e will shift from zero. This in turn will cause the DSM to seek a new updated operating point.

X. Simulation

In this section an equation based simulation is presented. Figure 8 shows the system model used. The network is composed of 8 unidirectional links drawn as solid arrows and 3 sources. Each source sends packets to its sink, which is co-located with the other sinks in the diagram. Each sink is assumed to use a different spreading code. Each source can send packets along several different paths to its source. Source 1 can send packets along three different paths and sources 2 and 3, two paths and a single path respectively. All paths are shown as dashed lines. Time is discrete and can be interpreted as either packet time or as the DSM iteration index.

The data used in the simulation are as follows. The performance metric is given as

$$\begin{aligned} U(r) &= \sum_{s=1}^{S} U_s(\mathbf{1}^T r_s) \\ &= \sum_{s=1}^{3} \xi_s \ln(1 + \mathbf{1}^T r_s) \end{aligned} \tag{40}$$

where ξ are weights associated with the priority given the sources. Equal weighting is used in this example. The gain matrix G was selected at random, and due to space limitations is not presented here. The routing matrix A and the flow matrix B can be inferred from the diagram.

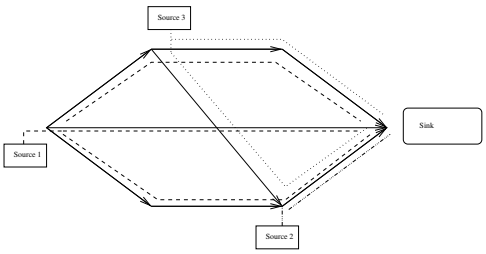

Fig. 8. Simulated ad-hoc network. Each sources has one or more paths to its sink.

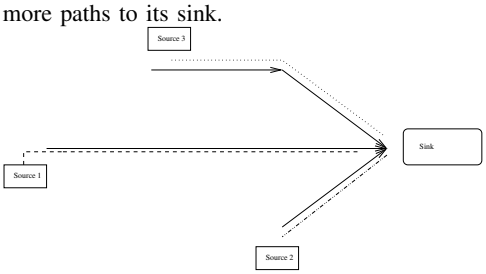

Fig. 9. Links and paths selected by DSM.

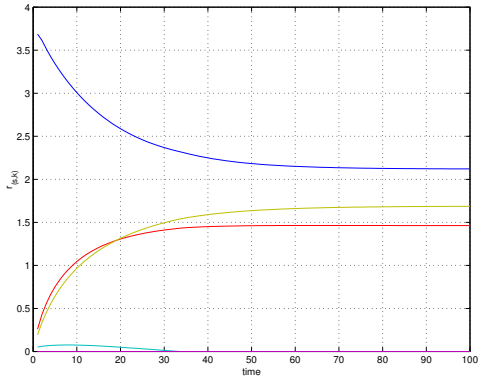

Fig. 10. Ad-hoc network source-path transfer rates.

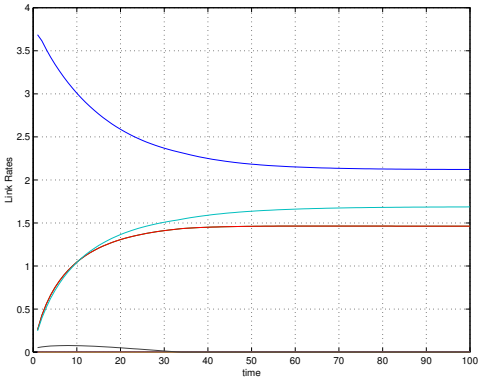

Fig. 11. Network link rates. Shown are the link rates for the links shown in Figure 9 Link rates for the remaining links are zero.

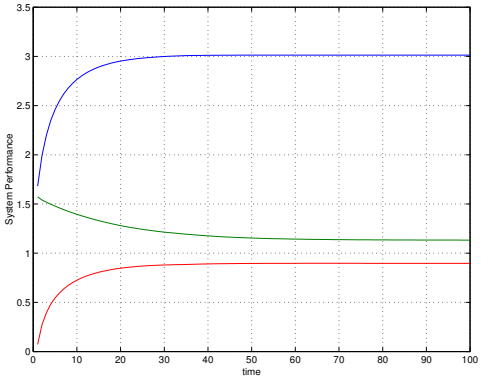

Fig. 12. Performance of Ad-hoc network. The upper curve is total performance and the lower three curves the performance of the three sources

Figure 9 depicts the links and paths selected by DSM. All links other than those shown have a link rate of zero. Each source follows a single path to the co-located sinks. The paths are disjoint in this simple example. System performance is shown in Figure 12. Because the system metric is concave, U not unexpectedly improves monotonically to its globally optimal value. The three lower curves show the performance evolution of each individual source. The trade-off between source 1, the falling curve, and sources 2 and 3, the rising curves, is a result of the shift in transfer rates shown in Figure 10. The transfer rate of source 1 is initially too high and is reduced by DSM allowing more traffic to be sent by the remaining sources. Figure 11 shows the evolution of the link rates. Link rates not used by the selected paths are assigned a zero rate. The close similarity between Figures 10 and 11 is a result of the simplicity of the simulation and will not generally be true.

For this simple model each source is associated with a single path. This will not generally be the case for other typologies or system parameters. If two or more paths associated with source s have equal capacity costs $c_{s,k}$, then all of these paths will be selected. In many data oriented applications the use of multiple paths may be acceptable. In voice applications, for example, the use of multiple paths offers little advantage. This issue can be resolved by applying DSM a second time with all but one of the previously selected optimal paths removed for each source.

XI. Summary

This paper address the problem of finding the best set of routes, link rates and link powers in an ad hoc or multi-hop network supporting different types of traffic and subject to constraints on network congestion and other QoS requirements. The approach is based on Perron Frobenius matrix theory and yields a Pareto optimal operating point. The DSM algorithm is iterative and adaptive.

References

[1] G. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Vehicular Technology*, vol. 42(4), pp. 641–646, Nov. 1993.

[2] N. Bambos, S. Chen, and G. Pottie, "Channel access algorithms with active link protection for wireless communication networks with power control," *IEEE/ACM Trans. on Networking*, vol. 8(5), pp. 583–597, Oct. 2000.

[3] D. Julian, M. Chiang, and D. ONeill, "Qos and fairness constrained convex optimization of resource allocation for wireless cellular and ad-hoc networks," in *Proc. IEEE Infocom*, June 2002, pp. 477–486.

[4] D. ONeill, D. Julian, and S. Boyd, "Seeking foschini's genie: Optimal rates and powers in wireless networks," *submitted to IEEE Trans. Vehicular Technology*, 2003.

[5] H. Li and D. Yu, "Comparison of ad hoc and centralized multihop routing," *Wireless Personal Multimedia Comm, 2002*, pp. 791–795, 2002.

[6] P. Bhagwat and A. Segall, "A routing vector methos for routing in bluetooth scatternets," *IEEE conf on Mobile Multimedia Communications*, pp. 375–379, 1999.

[7] A. Tsirigos and Z. Haas, "Multipath routing in mobile ad hoc networks or how to route in the presence of frequent topology changes," *Milcom 2001*, pp. 878–883, 2001.

[8] A. Goldsmith, *Wireless Communications*. Stanford, CA: Stanford University EE 359 Course Reader, 2001.

[9] S. Boyd and L. Vandenberghe. (2003) Convex optimization. [Online]. Available: http://www.stanford.edu/ boyd/cvxbook.html

[10] S. Low, L. Peterson, and L. Wang, "Understanding vegas: A duality model," *Journal of ACM*, vol. 49(2), pp. 207–235, Mar. 2002.

[11] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communications networks: Shadow prices, proportional fairness and stability," *Journal of Operations Research Society*, vol. 49, pp. 237–252, Nov. 1998.

[12] S. Low. (2002) Duality model of tcp and queue management algorithms. [Online]. Available: http://netlab.caltech.edu

[13] D. Gross and C. Harris, *Fundamentals of Queueing Theory*. New York, NY: John Wiley and Sons, 1985.

[14] G. Stewart and J. Sun., *Matrix perturbation theory*. Boston, Mass.: Academic Press, 1990.

[15] D. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.