Utility-Theoretic Heuristics for Intelligent Adaptive Routing in Large Communication Networks

Armin R. Mikler*, Vasant Honavar† and Johnny S.K. Wong‡
Department of Computer Science
Iowa State University
Ames, Iowa 50011, USA
mikler|honavar|wong@cs.iastate.edu

Abstract

Utility theory offers an elegant and powerful theoretical framework for design and analysis of autonomous adaptive communication networks. Routing of messages in such networks presents a real-time instance of a multi-criterion quasi-optimization problem in a dynamic and uncertain environment. In this paper, we examine several heuristic decision functions that can be used to guide messages along a near-optimal (e.g., minimum delay) path in a large network. We present an analysis of properties of such heuristics under a set of simplifying assumptions about the network topology and load dynamics. In particular, we identify the conditions under which one such utility-theoretic heuristic function is guaranteed to route messages along an optimal (minimum delay) path in a network with uniform load except for a single hotspot (when it is assumed that the network load stays

*Armin Mikler is currently a post-doctoral fellow at the Scalable Computing Laboratory of the Ames Laboratory (U.S. Department of Energy) at Iowa State University, Ames, IA 50011

†Vasant Honavar is grateful to the National Science Foundation for its support of his research through the grant NSF IRI-9409580.

‡The authors wish to thank Yong Zhao, Karthik Balakrishnan, Rajesh Parekh, and Chun-Hsien Chen (all doctoral students in Computer Science at Iowa State University), and Mary Oman (a post-doctoral fellow at Ames Laboratory) for their helpful comments on earlier drafts of the paper.

constant during the time-frame of interest). We derive bounds on the cost of a sub-optimal path (relative to the cost of an optimal path) as well as the probability that a resulting route between a randomly chosen source-destination pair is suboptimal. We then present a modified heuristic function that is guaranteed to eliminate sub-optimal routes under the same set of assumptions about network topology, load, and load dynamics. We conclude with a brief discussion of the relevance of the theoretical results presented in the paper to the design of intelligent autonomous adaptive communication networks and an outline of some directions of future research.

1 Introduction

With the unprecedented growth in size and and complexity of communication networks, the development of intelligent and adaptive approaches to network management (including such functions as routing, congestion control, etc.) have assumed considerable theoretical as well as practical significance. Knowledge representation and heuristic techniques of artificial intelligence, utility-theoretic methods of decision theory, as well as techniques of adaptive control offer a broad range of powerful tools for the design of intelligent, adaptive, and autonomous communication networks. This paper develops and analyzes some utility-theoretic heuristics for adaptive routing in large communication networks.

Routing [1, 2, 16, 15, 18] in a communication network refers to the task of propagating a message from its source towards its destination. A routing algorithm has to, among other things, select at each node, for each message received by, or originating at that node, a neighboring node to which the message is to be sent (unless the receiving node happens to be the destination). Such a routing algorithm may be required to meet a diverse set of often conflicting performance requirements (e.g., average message delay, network utilization, etc.). This makes routing an instance of a multi-criterion optimization problem.

In general, for a network node to be able to always make an optimal routing decision (as dictated by the relevant performance criteria) it requires not only up-to-date and complete knowledge of the state of the entire network (including such things as network connectivity and load at each node) but also, an accurate prediction of the behavior of the network during propagation of the message through the network (in other words, a precise knowledge of the dynamics of the network). A moment's reflection shows that this is impossible in practice. Thus, a routing algorithm must be capable of adapting to network state changes in almost real time in an effort to optimize the desired performance criteria.

In practice, all routing decisions in large communication networks are based on imprecise, uncertain knowledge of the current network state. This imprecision or uncertainty is a function of the network dynamics, the memory available for storage of network state information at each node, the frequency of, and propagation delay associated with, update of such state information. This makes the routing task, a resource-constrained multi-criterion optimization problem. Resource limitations (e.g., memory for storing network state information at each node) generally will not permit storage and use of a precise description of the state of the entire network. Thus, the routing decisions, in practice, have to be based on knowledge of network state over a local neighborhood supplemented perhaps by a summary of the network state as viewed from a given node. Motivated by these considerations, a class of adaptive heuristic routing algorithms have been developed over the past few years [7, 8, 9, 10, 11, 12]. Experiments demonstrate that such algorithms have a number of interesting properties including: automatic load balancing and message delay minimization. To date, the design of such algorithms has relied primarily on intuition and experimentation, in the absence of an adequate theoretical framework for analysis and design of appropriate heuristics. The work described in this paper is a tentative step toward the development of such a framework.

In what follows, we draw upon concepts of utility theory [3, 4, 17] to design and analyze utility-theoretic heuristics for routing in large communication networks. Various heuristics are designed and their properties are precisely analyzed in section 2. The relevance and limitations of the main results of the paper and some directions for further research are discussed in section 3.

2 Utility-Theoretic Heuristics for Routing

Routing messages in large communication networks so as to optimize some desired set of performance criteria presents an instance of resource-bounded, multi-criteria, real-time, quasi-optimization problem. Our proposed solution to this problem involves the use of *utility-theoretic heuristics.* The heuristic function enables each node n_j in the network to select a *best* neighbor in its neighborhood to route a message M (which it has received or generated) towards its destination. Ideally, this decision is to be performed so as to optimize a set of desired performance criteria. Utility theory offers an elegant theoretical framework for exploring the design and analysis of such heuristic functions. This section briefly reviews the key concepts of utility theory in the context of routing in large communication networks, provides examples of utility-theoretic heuristics, and presents a precise characterization of their properties. For the purpose of this analysis, it is assumed that the network is a regular rectangular grid (with adjacent nodes being unit distance of each other). Additional assumptions concerning load and load dynamics are made as necessary.

2.1 Guiding a Message Using Rewards

Since each node has to route each message that it receives towards its destination, it needs some directional guidance to accomplish this objective. We can model this in terms of a *reward* function that assigns a reward to a message upon reaching its destination. Then it is natural to think of each node as receiving a partial reward corresponding to each message that is routed through it towards an appropriate destination node. Such a reward function rewards each node for making a particular routing decision. Since the primary objective of the reward function is to guide messages toward their destinations, we could use a variety of functions that assign rewards to a node n_i according to some inverse function of the distance to the destination n_d of the message M being routed through n_i. The values of partial rewards that will be received by a node n_i for routing a message destined for n_d can be represented using a *reward landscape* R^d.

Let $D_{i,d}$ denote the Manhattan (or city block) distance between a node n_i and n_d (This makes sense in a grid network but other network topologies might call for the use of other distance measures). We define the partial reward for node n_i as follows:

$$R_i^d = f_R(D_{i,d}) \tag{1}$$

where f_R is a *reward function* chosen such that $\forall i \forall j \quad D_{i,d} \leq D_{j,d} \iff f_R(D_{j,d}) \leq f_R(D_{i,d})$.

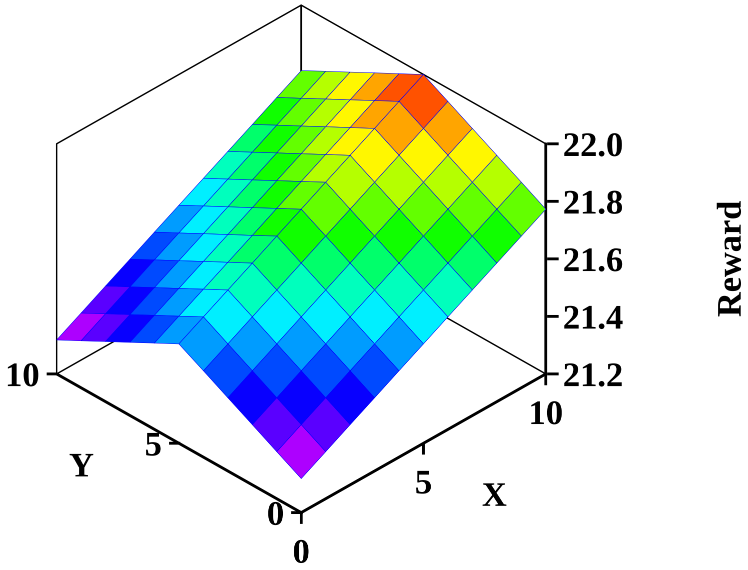

Figure 1: Reward Function $f_R(D_{i,d}) = (m+n) - \frac{D_{i,d}}{m+n}$ for destination $n_d = (10,5)$

There are many possible choices for the reward function $f_R(.)$ A particular example of $f_R(.)$ is given by

$$f_R(D_{i,d}) = (m+n) - \frac{D_{i,d}}{m+n} \tag{2}$$

where n and m are the dimensions of the grid network. The corresponding reward landscape for an 11×11 grid network with destination node $n_d = (10,5)$ is shown in Figure 1. Note that the results that follow are independent of any particular choice of $f_R(.)$ so long as the reward is an increasing function of distance between the node and the destination.

With a slight abuse of notation, assuming additive rewards, we can define a *cumulative reward* R^P that is obtained by a message M traveling along a path P (from its source n_s to its destination n_d) as follows:

$$R^P = \sum_{n_i \in P} R_i^d \tag{3}$$

2.2 Routing Messages Along Optimal Paths in a Uniformly Loaded Network Using Rewards, Penalties, and Payoffs

At each node along its path, the message incurs a penalty. Such penalty incurred by a message M at each intermediate node n_i along a path P can be modeled by a cost C_i. It is assumed that C_i is non-negative and bounded from above by some constant ξ for all messages and for all nodes in the network. That is, $\forall i \; 0 \leq C_i \leq \xi$. It is further assumed that the penalty C_i remains constant during the time it takes to make a routing decision for message M at node n_i. If cumulative delay is a performance criterion that is to be minimized by the routing algorithm a natural interpretation of C_i is the delay suffered by a message (on account of load) at n_i. (However, since delays can become unbounded when there is queueing, it may be necessary to discard some messages in order to keep the delay bounded at the expense of message loss. Alternatively, feedback could be used to discourage messages being routed through n_i. We will not address these issues in this paper.) If cumulative load is the criterion to be minimized, C_i is bounded by the maximum utilization $\rho \leq 1.0$.

The total penalty or cost incurred along a path P is given by:

$$C^P = \sum_{n_i \in P} C_i \tag{4}$$

We can now define the net *partial payoff* Z_i^d received by a message M when it reaches the node n_i on its way to its destination n_d as follows:

$$Z_i^d = R_i^d - C_i \tag{5}$$

Correspondingly, the total payoff along a path P is given by:

$$Z^P = R^P - C^P \tag{6}$$

Let Π be a minimum cost path from a source n_s to a destination n_d. The cost C^Π along this path is given by:

$$C^\Pi = \min_{\forall P}\{C^P\} \tag{7}$$

In the discussion that follows, in order to simplify our analysis, we proceed under the assumption that the network is uniformly loaded. (The implications of this assumption are discussed in section 3). This assumption is captured by the following definition:

Definition 1 *If $\forall i, C_i = \kappa$ ($0 \leq \kappa \leq \xi$), we refer to the network as a uniform cost network.*

Lemma 1 *In a uniform cost network a routing algorithm which propagates a message M such that Z_i^d is maximized at every intermediate step will yield an optimal path Π with cost C^{Π}.*

Proof of Lemma 1
Since in a uniform cost network, $\forall i, C_i = \kappa$, the partial reward Z_i^d from equation 5 can be written as $Z_i^d = R_i^d - \kappa$. Thus, Z_i^d can be maximized at each intermediate node along a path P simply by maximizing R_i^d. Let Λ_P be the number of nodes on a path P. As message M is propagated along a P such that R_i^d is maximized at every intermediate step, in a regular grid network, the property of the reward function (i.e., $\forall i \forall j \quad D_{i,d} \leq D_{j,d} \iff f_R(D_{j,d}) \leq f_R(D_{i,d})$) guarantees that M is propagated along *a shortest path* (as measured by the number of hops) P from the source n_i of the message M to its destination n_d. Thus, from equation 4 we have: $C^P = \kappa\Lambda_P$. Since P is a minimum hop (shortest) path, it follows that $C^P = C^{\Pi}$. $\Box$

It needs to be emphasized that the assumptions of this analysis are difficult to meet exactly in any realistic scenario. If nothing else, even if the network started off with uniform load, the act of routing a message from one node to another will perturb this uniformity. It is easy to see that if the uniform load assumption is not met, a routing algorithm that simply selects a neighboring node that has the largest partial reward cannot guarantee that the resulting paths are minimum cost paths. However, if deviations from the uniform cost assumption are "small" enough to not deflect a message from an optimal path (considering additional delay that is encountered by a message as a result of being routed along a minimum length path - as dictated by the reward function as opposed to a longer path that might have a lower cumulative delay), the result continues to hold. It is also worth pointing out that in large networks with hard to predict network dynamics, routing decisions have to be based on imprecise knowledge of network state and dynamics and there is no way to get around sub-optimal routes. A primary purpose of the preceding analysis as well as the analysis that follows is to develop a class of utility-theoretic heuristics that would be useful in routing messages along near-optimal routes in such networks. Extensions of Lemma 1 to networks that are only *approximately uniform cost* networks, *locally uniform cost networks*, or *probabilistically uniform cost networks* are therefore of interest. However, we will not explore such extensions in this paper. In section 2.3 that follows, we relax the uniform cost assumption by allowing a single hotspot (or a node with a high load) in an otherwise uniformly loaded network.

2.3 Routing with Utilities

Utility is a measure that quantifies a decision maker's preference for one action over another (relative to some criteria to be maximized) [4, 3, 17]. The utility function may be composed of several parts that model different types of payoffs, and weight the various parts as necessary to reflect the preferences of the decision maker. When the result of an action is uncertain, it is convenient to use the *expected* utility of each action to pick actions which maximize the expected utility.

Definition 2 *The utility U_i^d of node n_i (with respect to a destination n_d) is computed by one of its neighboring nodes, n_j, as n_j attempts to route a message M that it has received (or generated), along a desired (e.g., minimum cost) path, to M's destination, n_d. It is typically a real-valued function of the (usually limited) network state information available at node n_i, the message M, and perhaps other parameters.*

Thus, each node can use utilities of its neighbors to pick one that offers the highest utility value. Thus a node n_j preference-orders its neighbors n_i according to their respective utilities. We say that the router at n_j is *indifferent to* the choice between two neighbors n_k and say n_l if $U_k^d = U_l^d$ (where n_d is the destination of the message M being routed by n_j). We denote the indifference between two nodes as $n_k \sim n_l$. We say that a neighboring node n_k is preferred (by the router at n_j) over another neighbor n_l if $U_k^d > U_l^d$. We denote this preference by $n_k \succ n_l$ [3, 4, 17].

In our treatment of uniform cost networks in section 2.2, the partial payoff Z_i^d implicitly served as the utility for the purpose of routing messages along minimum cost paths. That is, we set $U_i^d = Z_i^d$. Let us call this utility function U^0.

Definition 3 *U^0 is a utility function which is given by $U_i^d = Z_i^d$.*

Note that the use of the naive utility function U^0 guarantees minimum cost paths only if a uniform cost network is assumed. The uniform cost assumption renders the cost component in the payoff function irrelevant to the routing decision. This is no longer true when the network is not a uniform cost network. In what follows, we relax the uniform cost assumption by allowing a single *hotspot* (a node with a high load relative to its neighbors) in an otherwise uniform cost network.

Definition 4 *A hotspot, n_h, in an otherwise uniform cost network is a single network node which has a higher load than its neighbors so that a message M traveling through it incurs a cost $C_h > \kappa$ (whereas the same message if it were to travel through a node n_i where $i \neq h$, it would incur a cost $C_i = \kappa$).*

Definition 5 *The neighborhood H_i of a node n_i is the set of nodes n_j, such that there exists a direct communication link $l_{i,j}$ from n_i to n_j.*

Thus, in a regular rectangular grid, each node (with the exception of the boundary nodes) has exactly 4 neighbors.

Note that since the costs C_i are bounded by ξ, it follows that $C_h \leq \xi$. Further note that the above definition of a hotspot does not say anything about the relative difference in costs C_h and C_i. A more realistic definition of a hotspot might insist that the cost of routing a message through a hotspot is *significantly large* relative to routing the same message through a node in the neighborhood of the hotspot. Also, when a network deviates substantially from the uniform cost assumption, it is more useful to focus on the load distribution in the vicinity of a node rather than hotspots per se. However, to make the analysis mathematically tractable, the discussion that follows focuses on routing in an otherwise uniform cost networks with a single hotspot.

2.4 Routing in an Otherwise Uniform Cost Network with a Single Hotspot

As the uniform cost assumption is relaxed by allowing a hotspot n_h with cost $C_h > C_j \;\; \forall j \neq h$ in the network, it is easy to show that relying on partial payoffs alone as utilities for routing messages can result in sub-optimal routes. Consider a grid network with node coordinates increasing as a message M

travels east and south. From the uniform cost assumption, we have $C_i = C_j = \kappa \ \forall i, j \neq h$. Let x_s, y_s, x_d, and y_d be the x and y coordinates of M's source and destination, respectively. Let x_h and y_h be the x and y coordinates of a hotspot in one of the following configurations:

1. $x_s \leq x_h \leq x_d \wedge y_s \leq y_h \leq y_d$ or

2. $x_s \geq x_h \geq x_d \wedge y_s \geq y_h \geq y_d$

In this case, clearly, the probability that a shortest path from n_s to n_d passes through the hotspot n_h is non-zero. Hence, there exists a node n_i with $n_h \in H_i$ that must decide how to route M so as to minimize the total cost incurred by M. As we show below, if this decision is based on a preference ordering induced by the naive utility function U^0 given by $U_i^d = Z_i^d$ (as in section 2.2 above), messages can be routed through the hotspot thereby incurring a higher cost than they would have otherwise.

Assumption 1 *For the discussion below, we assume that the reward functions chosen guarantee that* $\forall n_k \, \forall n_i$ *in the network such that* $| R_i^d - R_k^d | > \xi$ *whenever* $D_{i,d} \neq D_{j,d}$.

This ensures that the cost C_i of a node n_i, (and n_h in particular) does not offset the *guidance* provided through R_i^d unless two nodes with equal rewards are being compared.

In the following we distinguish 4 canonical cases. We focus in our analysis on configuration 1 above. Similar arguments hold for configuration 2.

Case 0

Case 0 combines 4 scenarios of placing nodes n_s, n_d, and n_h in the grid network, each of which presents a trivial routing problem. In these scenarios, at least two of the nodes n_s, n_d, and n_h are identical, that is:

1. $n_s = n_d = n_h$,

2. $n_s = n_d$,

3. $n_s = n_h \neq n_d$, and

4. $n_s \neq n_h = n_d$.

Clearly, in scenarios 1 and 2, no routing decisions are needed as the message source coincides with the destination. Whenever the message source coincides with the hotspot as in scenario 3, the routing algorithm will select a neighbor $n_k \in H_i$ with the highest utility. Hence, the routing algorithm performs as in the case of a uniform cost network (without hotspots). For scenario 4, Assumption 1 assures that n_d yields the highest partial reward R_i^d, $\forall i$, despite the fact that the cost incurred by hotspot conditions reduces its partial payoff. Hence, routing decisions can be made without taking cost C_i into consideration, as in the case of a network without hotspots.

Case 1

Let $P\triangle_{i,j}$ denote the number of minimum hop paths from a node n_i to node n_j. $P\triangle_{i,j}$ is given by:

$$P\triangle_{i,j} = \binom{\mid x_j - x_i \mid + \mid y_j - y_i \mid}{\mid x_j - x_i \mid} \tag{8}$$

Case 1 encompasses all placements of nodes n_s, n_h, and n_d, such that

1. $P\triangle_{s,h} > 1 \ \wedge P\triangle_{h,d} > 1$ or
2. $P\triangle_{s,h} = 1 \ \wedge P\triangle_{h,d} \geq 1$ where $P\triangle_{s,d} > 1$ (see Case 3)

An example of Case 1 scenarios is shown in Figure 2. Here, the hotspot n_h does not share either the x or y coordinates of n_s or n_d. That is, $(x_s < x_h < x_d) \ \wedge \ (y_s < y_h < y_d)$. In this configuration, the number of minimum hop paths from n_s to n_h is $P\triangle_{s,h}$, and any such partial path would lead to a minimum cost path if all nodes n_i that neighbor the hotspot (i.e., $n_h \in H_i$) take action to route M so as to circumvent n_h. Thus, the utility function U^0 given by $U_i^d = Z_i^d$ is guaranteed to route M on a minimum cost path to its destination n_d. This follows from the existence of since minimum hop paths that avoid going through the hotspot.

Lemma 2 *In a uniform cost network with a single hotspot n_h located such that $(x_s < x_h < x_d) \ \wedge \ (y_s < y_h < y_d)$, a routing algorithm which propagates a message M such that U^0 is maximized at every intermediate step will yield an optimal path Π with cost C^{Π}.*

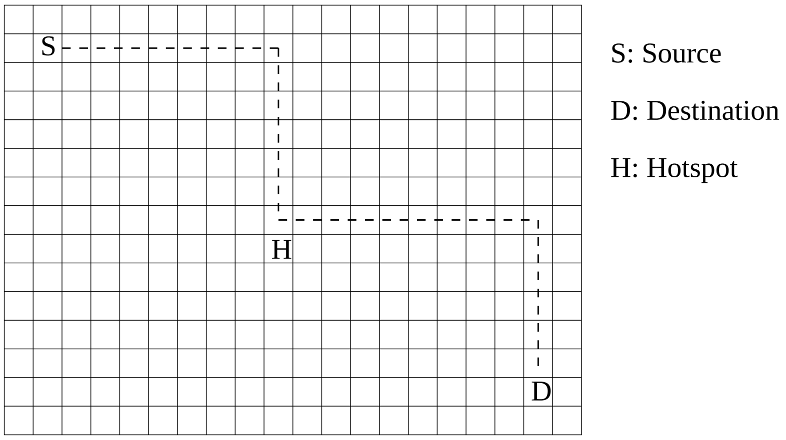

Figure 2: $(x_s < x_h < x_d) \wedge (y_s < y_h < y_d)$

Proof of Lemma 2
Clearly, the only nodes at which a decision has to be made to circumvent n_h are n_i or n_j $(x_h - 1, y_h)$ and $(x_h, y_h - 1)$, respectively. Since $x_h < x_d \wedge y_h < y_d$, there exist nodes n_k and n_l with coordinates $(x_h - 1, y_h + 1)$ and $(x_h + 1, y_h - 1)$, respectively, that lie on a minimum hop path from n_s to n_d. Since $C_k = C_l = \kappa < C_h$ it follows that $Z_k^d = Z_l^d > Z_h^d$. Hence, a routing decision in n_i or n_j that maximizes the partial payoff will choose n_k or n_l to propagate M towards n_d. Since $C_i = \kappa \;\; \forall i \neq h$, and M is propagated along a minimum hop path, Lemma 1 guarantees that M is routed along an optimal path Π. $\Box$

Case 2
Here, n_s, n_d, and n_h are placed such that $(x_s < x_h < x_d) \wedge (y_s < y_h = y_d)$ or $(x_s < x_h = x_d) \wedge (y_s < y_h < y_d)$ (see Figure 3), i.e.; $(P\triangle_{s,h} > 1) \wedge (P\triangle_{h,d} = 1)$.

Assuming the former, there exists a node n_i with coordinates (x_i, y_i) with $(x_s < x_i < x_h) \wedge (y_i = y_h = y_d)$ from which the number of minimum hop routes $P\triangle_{i,d} = 1$. Since in a uniform cost network $n_k \sim n_l$, $\forall k, l \neq h$ the naive utility function U^0 can guide a message M through n_i, thereby committing to a path P with cost $C^P > C^\Pi$. Assuming that M is only routed using utilities to choose among minimum hop routes, the additional cost $(C^P - C^\Pi)$ is inflicted on M by n_h. If M is permitted to deflect from a minimum hop route, the additional cost $(C^P - C^\Pi)$ is inflicted by n_h itself or due to the

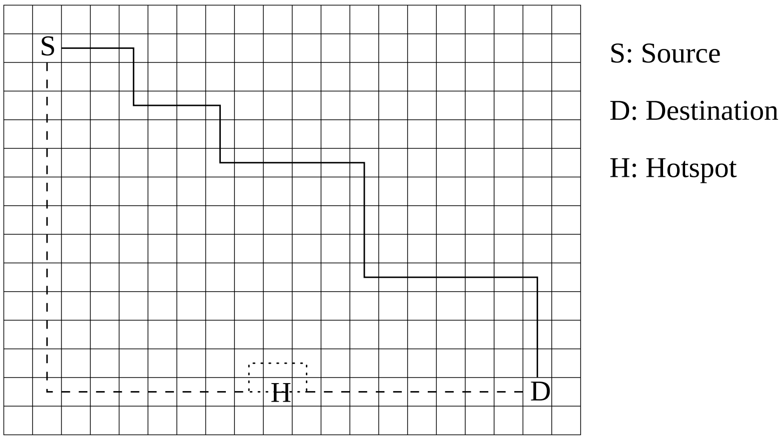

Figure 3: $(x_s < x_h < x_d) \wedge (y_s < y_h = y_d)$

extended length of P in circumventing n_h.

Case 3
This scenario consists of all placements of n_s, n_d, and n_h such that $(x_s = x_h = x_d) \wedge (y_s \leq y_h \leq y_d)$ or $(x_s \leq x_h \leq x_d) \wedge (y_s = y_h = y_d)$ (see Figure 4). Since there is only a single optimal path Π from n_s to n_d, i.e., $P\triangle_{s,d} = 1$, message M must either visit n_h or deflect from the minimum hop path in order to circumvent n_h. U^0, however is not sufficiently informative to guarantee an optimal routing decision. Hence, M may be routed along a path P for which $C^P > C^\Pi$.

Assumption 2 *In the following we assume that a node n_j upon receiving a message M from a neighbor node $n_i \in H_j$ will refrain from propagating M back to n_i.*

This is a natural assumption that is meant to avoid the so-called *bouncing of messages* back to a node from which it was routed.

Lemma 3 *In a uniform cost network with a single hotspot n_h, a routing algorithm based on U^0 will deflect a message M at most once in order to circumvent n_h provided bouncing is avoided (via Assumption 2).*

Proof of Lemma 3
Consider a node n_i with coordinates (x_i, y_i) such that $x_s < x_i = x_h - 1 <$

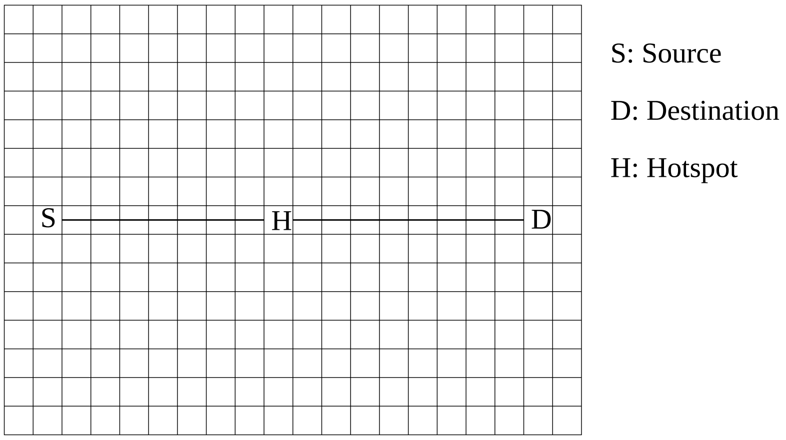

Figure 4: $(x_s < x_h < x_d) \wedge (y_s = y_h = y_d)$

$x_d \wedge y_s < y_i = y_h = y_d$. (Similar analysis holds for the case where $x_s = x_i = x_h < x_d \wedge y_s < y_i = y_h - 1 < y_d$). Node n_i can deflect M to a node n_j with coordinates (x_j, y_j), such that $x_s < x_j = x_h - 1 < x_h < x_d \wedge y_s < y_j = y_h \pm 1$. Clearly, $P\triangle_{j,h} = 2$. Since $x_h < x_d$, $P\triangle_{j,d} > 2$. Hence there must exist a node n_k with $n_k \in H_j$ which lies on a minimum hop path P from n_j to n_d such that $n_h \notin P$. The choice of the reward function 1 guarantees that $R_i^d = R_k^d > R_j^d$. In a grid, $\nexists n_l \; l \neq i, k$ such that $n_l \in H_j \;\; \wedge R_l^d \geq R_k^d$ (since the reward function ensures that the rewards vary inversely with the Manhattan distance). Since $C_j = C_k = C_i = \kappa$, $Z_i^d = Z_k^d > Z_j^d$. This limits the routing choices for message M at n_j to n_i and n_k, of which, by assumption 2, n_k has to be chosen (since otherwise M will be bounced back to n_i, which had routed the message to n_j to begin with thereby violating assumption 2). This ensures that from n_j, M is sent along a minimum hop path P to the destination n_d. Since $n_h \notin P$, Lemma 1 guarantees that M is propagated along P without further deflection. $\Box$

The preceding analysis sets the stage for the following theorem which bounds the cost of a suboptimal path in an otherwise uniform cost network with a single hotspot.

Theorem 1 *In a uniform cost network with a single hotspot n_h with $C_h > \kappa$ (where $\forall i \neq h, C_i = \kappa$), a routing algorithm which propagates a message M such that U^0 is maximized at every intermediate step is guaranteed to yield a path P with cost C^P such that $C^P - C^\Pi \leq \max((C_h - \kappa), 2\kappa)$.*

Proof of Theorem 1

In Case 1, Lemma 2 guarantees that a routing algorithm based on U^0 will find a minimum cost path if n_s, n_h, and n_d are placed such that $(x_s < x_h < x_d) \wedge (y_s < y_h < y_d)$. Hence, $C^P = C^\Pi$ and thus $C^P - C^\Pi = 0 < \max((C_h - \kappa), 2\kappa)$.

Case 2 involves a node n_i with coordinates (x_i, y_i) such that $x_s < x_i = x_h - 1 < x_d \wedge y_s < y_i = y_h = y_d$ or $x_s = x_i = x_h < x_d \wedge y_s < y_i = y_h - 1 < y_d$. Now n_i must decide whether to route message M through n_h or to deflect M from a minimum hop path. Routing through n_h will result in a path cost C^P which is sub-optimal by an amount $C_h - \kappa$. That is, $C^P - C^\Pi = C_h - \kappa \leq \max((C_h - \kappa), 2\kappa)$. If n_i chooses to deflect M so as to circumvent n_h, M is propagated along a path P'. Let Λ_P be the length (in number of hops) of the minimum hop path P from n_i to n_d via n_h and $\Lambda_{P'}$ be the length of path P'. Deflecting from path P in a grid topology yields a path P' with $\Lambda_{P'} = \Lambda_P + 2$. Lemma 3 guarantees that M is deflected at most once, $C^{P'} = C^\Pi + 2\kappa$. Hence $C^{P'} - C^\Pi = 2\kappa \leq \max((C_h - \kappa), 2\kappa)$.

In Case 3, the cost C^Π for a minimum cost path Π between n_s and n_d is given by $C^\Pi = \Lambda_P \kappa + \min(C_h - \kappa, 2\kappa)$. Hence, $C^{P'} - C^\Pi \leq \max((C_h - \kappa), 2\kappa)$.

If n_h coincides with either n_s or n_d, the hotspot cannot be circumvented and $C^P = C^\Pi$. (That is, the minimum cost path has to necessarily pass through the hotspot in this case). Clearly, $C^P = C^\Pi$ and $0 < \max((C_h - \kappa), 2\kappa)$. Therefore, $C^P - C^\Pi \leq \max((C_h - \kappa), 2\kappa)\ \forall P$. $\Box$

2.4.1 The Probability of Sub-Optimal Routing Using U^0

Let the probabilities of a sub-optimal path due to Cases 1, 2, and 3 be ϕ_1, ϕ_2, and ϕ_3, respectively. Only Cases 2 and 3 can result in suboptimal paths when routing decisions are based on U^0. Thus, $\phi_1 = 0$. The probability of sub-optimal routes $P_{\neg\Pi}$ in an N-node $m \times n$ grid network is given by:

$$pr(P_{\neg\Pi}) = \phi_1 + \phi_2 + \phi_3 \tag{9}$$

The probability ϕ_2 is computed as:

$$\phi_2 = \sum_W \sum_{H_x} pr(P_{\neg\Pi,x}) + \sum_W \sum_{H_y} pr(P_{\neg\Pi,y}) \tag{10}$$

where $pr(P_{\neg\Pi,x})$ and $pr(P_{\neg\Pi,y})$ are the probabilities of sub-optimal paths when $x_h = x_d$ and $y_h = y_d$, respectively. In equation 10, W is the set of all possible placements of n_s and n_d such that $\mid x_s - x_d \mid > 0 \wedge \mid y_s - y_d \mid > 0$. H_x, the set of hotspot placements n_h such that x_h lies strictly between x_s and x_d and $y_h = y_d$. H_y, the set of hotspot placements n_h such that y_h lies strictly between y_s and y_d and $x_h = x_d$. Let n_i be a node with coordinates (x_i, y_i) such that $x_i = x_h - 1 \wedge y_i = y_h$ in the case where $x_s < x_h < x_d$ ($x_i = x_h + 1 \wedge y_i = y_h$ if $x_s > x_h > x_d$), with $n_h \in H_x$ (or $n_h \in H_y$ correspondingly). Let $P\triangle_{s,i}$ and $P\triangle_{s,d}$ (respectively) be the number of minimum hop paths from n_s to n_i and from n_s to n_d. The probabilities $pr(P_{\neg\Pi,x})$ and $pr(P_{\neg\Pi,y})$ are given by:

$$\frac{P\triangle_{s,i}}{P\triangle_{s,d} \cdot N^3} \tag{11}$$

where $1/N^3$ is the probability of a particular configuration of n_s, n_d, and n_h in the $m \times n$ grid (assuming each such configuration to be equally likely).

It is clear that a message in transit from a source to a destination in a source-hotspot-destination configuration corresponding to the scenario in Case 3 will be routed through a hotspot if routing decisions at each node are based on U^0. Depending on the relative difference between C_h and C_i, this may or may not yield an optimal path. Thus, we can establish an upper bound on the number of possible sub-optimal routing decisions in the scenario described in Case 3 in terms of the number of such source-hotspot-destination configurations that are possible in an $n \times m$ network. The resulting probability of sub-optimal path arising out of a node placement corresponding to to Case 3, namely ϕ_3, is given by:

$$\phi_3 \leq \frac{2n\binom{m}{3} + 2m\binom{n}{3}}{N^3} \tag{12}$$

$2n\binom{m}{3}$ and $2m\binom{n}{3}$ count the number of possible node placements in each row and column of the grid network that constitute a Case 3 scenario. Equation 12 represents an upper bound on ϕ_3 as it counts some of the trivial scenarios listed under Case 0.

The preceding analysis can be summarized in the form of the following corollary:

Corollary 1 *The expected penalty for choosing sub-optimal routes in a $m \times n$ otherwise uniform cost network with a single hotspot due to routing decisions based on U^0 is bounded by $pr(P_{\neg\Pi}) \max((C_h - \kappa), 2\kappa)$.*

2.5 Eliminating Suboptimal Routing Using A Modified Utility Function

So far, routing decisions in node n_k are based on the simple utility function U^0 given by $U_i^d = Z_i^d$ which preference-orders nodes n_i in H_k according to the payoffs Z_i^d. Note that Z_i^d, is determined solely from local information available when a message M arrives at n_i, such as the reward R_i^d, and the cost C_i.

Sub-optimal routing scenarios discussed above arise primarily as a result of a lack of knowledge at n_i at the time it is routing a message M to a neighbor n_j, regarding the likely cost of completing the path from n_j to the destination of M, namely, n_d. As shown in section 2.4, source-hotspot-destination configurations corresponding to scenarios described in Case 2 and Case 3 can result in sub-optimal routes (i.e., $C^P > C^\Pi$) when routing decisions are based on U^0.

In what follows, we explore modifications of U^0 to obtain a utility function which is guaranteed to eliminate suboptimal routing decisions that arise in source-hotspot destination placements corresponding to the scenarios in Case 2 and Case 3. We proceed in two steps: First, we define a utility function U^1 that eliminates suboptimal routing decisions that arise in scenarios corresponding to Case 3. We then modify U^1 by introducing a cost estimator function to obtain a utility function U^2 designed to eliminate suboptimal routing decisions that arise in Case 2 scenarios as well.

2.5.1 Eliminating Sub-Optimal Routing Decisions in Case 3 Using a Utility Function U^1

Definition 6 *Let U^1 be a utility function given by:*

$$U_j^d = \begin{cases} R_j^d & \text{if } \kappa < C_j < 3\kappa \wedge \nexists k \, (R_j^d = R_k^d) \, \wedge (n_j \neq n_d) \\ Z_j^d & \text{otherwise} \end{cases} \tag{13}$$

U^1 exploits the fact that messages are to be routed in a uniform cost network with a single hotspot. If routing decisions are based on the preference ordering induced by U^1 in an otherwise uniform cost network with a single hotspot, every message originating in a source n_s and a destination n_d that correspond to a source-hotspot destination placement described in Case 3 is guaranteed to be propagated along an optimal path Π between n_s to n_d.

Using U^1, n_i can decide whether or not to propagate M through a hotspot n_h in its neighborhood or to circumvent the hotspot by routing M through a different neighbor $n_k \neq n_h$. In other words, the preference ordering induced by U^1 ensures that at a node neighboring a hotspot in a Case 3 scenario we have:

- $(C_h - C_k) = (C_h - \kappa) > 2\kappa \Longleftrightarrow n_k \succ n_h$ and
- $(C_h - C_k) = (C_h - \kappa) < 2\kappa \Longleftrightarrow n_h \succ n_k$.

Thus all routing decisions based on U^1 in Case 3 scenarios result in optimal (minimum cost) routes. In other words, in equation 12, ϕ_3 is reduced to 0.

It is easy to see that U^1 does not eliminate the possibility of a sub-optimal route in a source-hotspot-destination configurations corresponding to the scenario in Case 2, leaving ϕ_2 in equation 10 unchanged.

The preceding analysis is summarized by Corollary 2:

Corollary 2 *The expected cost penalty associated with a sub-optimal route in an $m \times n$ otherwise uniform cost network with a single hotspot due to routing decisions based on U^1 is bounded by $\phi_2 \min((C_h - \kappa), 2\kappa)$.*

2.5.2 Eliminating Sub-Optimal Routing Decisions in Case 2 Using a Utility Function U^2

As shown by the preceding analysis, U^1 can result in a sub-optimal routing decision in a source-hotspot-destination configuration corresponding to the scenario in Case 2. In particular, any routing decision in a configuration corresponding to Case 2 will result in a sub-optimal path P if it results in the propagation of a message M to a node $n_k \in P$ such that $x_k < x_h < xd \wedge y_k = y_h = y_d$ or $x_k = x_h = xd \wedge y_k < y_h < y_d$. Routing decisions based on a preference ordering induced by U^1 can lead to such a situation since in a neighborhood H_i of n_i such that $n_h \notin H_i$, $\forall n_j \; n_k \in H_i$, $n_k \sim n_j$ provided $R_k^d = R_j^d$. Note that Case 2 scenarios include all placements of n_s, n_h, and n_d, such that $\forall \, \{n_i \mid x_i \neq x_d \wedge y_i \neq y_d\} \; \exists \, k, l$, such that $(n_k \in \Pi) \vee (n_l \in \Pi)$.

These observations suggest the possibility of using an estimate of the cost along paths from n_k to n_d as a component of a modified utility function U^2 so as to induce a preference ordering between nodes (where no such preference ordering is induced by U^1) so as to eliminate suboptimal routing decisions altogether. In other words, U^2 should be able to induce a preference ordering among nodes n_k and n_l in the neighborhood of a node n_i (the node making the routing decision for a message M) such that:

- $(n_k \in \Pi) \wedge (n_l \notin \Pi) \Longrightarrow n_k \succ n_l$ and
- $(n_k \notin \Pi) \wedge (n_l \in \Pi) \Longrightarrow n_l \succ n_k$.

We now proceed to define a cost estimator function E_k^d as follows:

Definition 7 *A cost estimator function $E_k^d(.)$ estimates the cost E_k^d of a minimal cost path to a destination n_d from a node n_k.*

It would be nice if the cost estimator function defined above helps U^2 to induce the desired preference ordering necessary to guarantee routing along an optimal path in the scenario corresponding to Case 2. We capture this property by defining what are called *admissible* cost estimator functions.

Definition 8 *A cost estimator function is said to be* admissible *if forall nodes n_i in the network, for all nodes n_k, n_l in the neighborhood H_i of n_i, the following conditions hold:*

- $(n_k \in \Pi) \wedge (n_l \notin \Pi) \Longrightarrow E_k^d < E_l^d$ *and*
- $(n_k \notin \Pi) \wedge (n_l \in \Pi) \Longrightarrow E_l^d < E_k^d$.

Definition 9 *We define a utility function U^2 as follows:*

$$U^2 = \begin{cases} U^1 & \text{if } x_s = x_d \ \vee \ y_s = y_d \\ U_j^d = R_j^d - C_j - E_j^d & \text{otherwise} \end{cases} \tag{14}$$

In the discussion that follows, we make the following assumptions.

Assumption 3 *The cost estimator function E_k^d used in equation 15 is admissible.*

Clearly, the estimate returned by $E_k^d(.)$ must be based, at the very least, on some knowledge of the current cost distribution in the network. In general, more precise estimates would require knowledge of the network dynamics. If costs associated with each node are allowed to change with time, as would be the case in a more realistic routing task, since E_k^d is computed at the time a message M is being considered for propagation through n_k, to a destination n_d, E_k^d has to change with time as well (so as to reflect the changes in the costs associated with various nodes in the network). However, the representation of network cost distribution cannot be specific to a particular destination node n_d since n_d is specified independently for each message. Thus we need an intermediate representation of the costs associated with each node in the network at a given time which the cost estimator function E_k^d can use. Any such representation, in order to be useful in practice in large networks, must not require the storage and update at (or broadcast to) each node, of cost values for all the nodes in the network regions of the network. Ideally, it must adequately summarize the load values in large regions of the network as viewed from a given node.

These considerations (among others) led us to define a *view*, V_k, which is maintained in every node in the network [7, 8, 9, 10, 11]. In a rectangular grid network, this view consists of four components, one for each of the four directions - north, south, east, and west. Thus we have: $V_k = [V_k^N, V_k^S, V_k^E, V_k^W]$.

Each component $V_k^\delta : (\delta \in \{N, S, E, W\})$ represents a weighted average of costs C_i along the minimum hop path from n_k to the border of the grid network in the direction specified by δ. Consider two nodes, n_i and n_k, located such that $n_k \in H_i$ and n_k is to the east of n_i, i.e., $x_i < x_k \wedge y_i = y_k$. Then V_i^E is given by:

$$V_i^E = \frac{C_k + V_k^E}{2} \tag{15}$$

V_i^N,V_i^S, and V_i^W are computed using analogous formulae.

Note that the views summarize the costs associated with nodes along a general direction. Given the recursive nature of view computation, views take a certain time to stabilize after major load changes in the network. This requires a careful choice of frequency of view update etc. The reader is refered to [7, 8, 9, 10, 11] for a discussion of this and related issues.

Assumption 4 *In the discussion that follows, we assume that sufficient time has elapsed for the view computation to stabilize before the view is used in the computation of cost estimates using $E_k^d(.)$.*

In practice, this assumption need not be satisfied exactly so long as the views are adequately precise to ensure the admissibility of the cost estimator function defined below. Assuming that n_d is located such that $x_s < x_d \wedge y_s <$

y_d. Let $D_i^x = | x_i - x_d |$ and $D_i^y = | y_i - y_d |$ denote the distance from n_i to n_d in x and y direction, respectively. $E_i^d(.)$ is given by:

$$E_i^d(.) = \frac{D_i^x V_i^E + D_i^y V_i^S}{D_i^x + D_i^y} \tag{16}$$

The estimator defined by equation 16 is one of several alternatives that are possible. We will not delve into a consideration of alternative definitions of $E_k^d(.)$ in this paper. It is easy to verify that the estimator given by equation 16 is in fact, admissible. In what follows, we revert back to the primary focus of this section, namely, utility functions for making optimal routing decisions in an otherwise uniform cost network with a single hotspot.

Lemma 4 *For all nodes n_i in the network, for each message M from a source n_s to a destination n_d that reaches a node n_i, the routing decision at n_i based on the preference ordering induced by U^2 will route M along a path P selected only from the set of minimum hop paths from n_i to n_d, unless $P\triangle_{i,d} = 1$ and $(n_h \in P) \wedge (n_h \in H_i)$.*

Proof of Lemma 4:
Consider a routing decision to be made for message M by a node n_i. Since $P\triangle_{i,d} > 1$ and $n_h \notin P$, there must exist at least one node $n_k \in H_i$ such that $n_k \neq n_h$ and $R_i^d < R_k^d$ (i.e., n_k is closer to the destination (n_d) than n_i). For Lemma 4 to hold, we have to show that the router at n_i, based on the perference ordering induced by U^2, will necessarily route M to such a node n_k. That is, U^2 must ensure that n_i will not route M through a node $(n_j \in H_i) \wedge (n_j \neq n_h)$ such that $D_{j,d} > D_{i,d} > D_{k,d}$. In other words, in this scenario we have to show that $n_k \succ n_j$ as per the preference ordering induced by U^2.

Note that by Assumption 1 $(R_i^d - R_j^d) > \xi$ and $(R_k^d - R_i^d) > \xi$, and $R_k^d - R_j^d > 2\xi$ (This follows from the fact that n_i and n_j are one hop from each other, n_k and n_i are one hop from each other, and n_j and n_k are two hops from each other). Since $\forall i,\ C_i \leq \xi$, equation 15 guarantees that $V_i^\delta \leq \xi$. By equation 16 $E_i^d(.) \leq \xi$, and thus $(C_i + E_i^d) \leq 2\xi$. Thus we have $(R_k^d - C_k - E_k^d) - (R_j^d - C_j - E_j^d) > 0$ which implies $U_k^d > U_j^d$. This implies that n_i routes M through n_k. Since $n_k \in P$ and $n_j \notin P$ (where P is a minimum hop path from n_i to n_d), this proves lemma 4. $\square$.

The preceding discussion sets the stage for Theorem 2 that establishes a major property of the utility function U^2, namely, that it eliminates sub-optimal routes in an otherwise unformly loaded grid network with a single hotspot.

Theorem 2 *In a uniform cost network with a single hotspot n_h with an associated cost $C_h > \kappa$ (where $\forall i \neq h, C_i = \kappa$), a routing algorithm which makes routing decisions at each node based on a preference ordering induced by U^2 is guaranteed to propagate each message M along a minimum cost path Π.*

Proof of Theorem 2
Consider the placement of n_s and n_d, such that $(x_s \leq x_d) \ \wedge \ (y_s \leq y_d)$. (Analogous arguments hold for other source-destination configurations). For nodes n_i, n_j, and n_k for which $(x_i, x_j, x_k < x_h)$, $(y_i, y_j, y_k < y_h)$, $(n_j, n_k \in H_i)$ and $R^d_i < R^d_j = R^d_k$, as per preference ordering induced by U^2 for a message to be propagated from n_i, $n_j \sim n_k$. Hence, a message will be propagated through the network along a minimum cost partial path until a routing decision has to be made which involves a node n_k with coordinates $x_k = x_h \ \wedge \ y_k < y_h$ or $x_k < x_h \ \wedge y_k = y_h$. At this point, the utility of n_k is below that of some n_j with coordinates $x_j < x_h \ \wedge y_j < y_h$ on account of the relative values of the cost estimates E^d_j and E^d_k. This causes the message M to be propagated to a node n_l with coordinates $x_l = x_h - 1 \ \wedge y_l = y_h - 1$. We can now show that M will always circumvent n_h and is propagated along Π. We will consider each of the four cases in turn.

Since routing in Case 0 scenarios is equivalent to routing in the absence of hotspots, we have $U^2 = U^1 = U^0$. Hence, a message M will travel along a minimum delay path Π.

As an example for Case 1 scenarios, we have $x_h < x_d \ \wedge y_h < y_d$. Consider the two possible routing decisions n_j and n_k with coordinates $x_j = x_h - 1 \ \wedge$ $y_j = y_h$ and $x_k = x_h \ \wedge y_k = y_h - 1$ respectively. Since both n_j and n_k offer a minimum cost path to n_d, either decision will cause the message M to be routed along an optimal path Π. Since $C_h > C_l = \kappa$ for $n_l \in H_k$ or $n_l \in H_j$ and $x_h < x_d \ \wedge y_h < y_d$, M will circumvent n_h while approaching n_d. Lemma 4 assures us that U^2 will propagate messages only along a minimum hop path and since given the same number of hops, a path that circumvents a hotspot is necessarily of a lower cost than a path that goes through a hotspot, we can say that for all source-hotspot-destination configurations that correspond to the Case 1 scenario, U^2 guarantees that M is propagated along an optimal path Π.

In a Case 2 scenario, the routing algorithm has to choose at a node n_l, a neighbor from among nodes n_j and n_k with coordinates $x_j = x_h - 1 \wedge y_j = y_h = y_d$ and $x_k = x_h \wedge y_k = y_h - 1$. Clearly, a routing decision that would yield n_j will result in a suboptimal path P since $x_j = x_h - 1 < x_d \wedge y_j = y_h = y_d$. We can now prove that a routing decision based on the preference ordering induced by U^2 will necessarily select n_k over n_j thereby circumventing n_h. Clearly, $C_j = C_k = \kappa \; \forall i, k \neq h$. Since all nodes to the east of n_h have cost κ, equation 15 yields $V_h^E = \kappa$. It follows therefore that the east view computed at n_j is $V_j^E = (C_h + \kappa)/2$. Correspondingly, the south view computed in n_k is $V_k^S = (C_h + \kappa)/2$. As n_h does not impact the south view of n_j or the east view of n_k, we have $V_j^S = V_k^E = \kappa$. Since n_j and n_k have the same distance from n_d, we have $R_j^d = R_k^d$. Therefore the preference ordering between n_j and n_k for routing decisions in n_l is determined by the relative values of $E_j^d(.)$ and $E_k^d(.)$. In other words, n_k is prefered over n_j if $E_j^d(.) - E_k^d(.) > 0$. $E_j^d(.)$ and $E_k^d(.)$ are given by:

$$E_k^d(.) = \frac{D_k^x \kappa + D_k^y (\frac{\kappa + C_h}{2})}{D_k^x + D_k^y}$$

$$E_j^d(.) = \frac{D_j^x (\frac{\kappa + C_h}{2}) + D_j^y \kappa}{D_j^x + D_j^y} = \frac{(D_k^x + 1)(\frac{\kappa + C_h}{2}) + (D_k^y - 1)\kappa}{D_j^x + D_j^y}$$

$E_j^d(.) - E_k^d(.)$ is then given by:

$$\frac{(D_k^x + 1)(\frac{\kappa + C_h}{2}) + (D_k^y - 1)\kappa}{D_j^x + D_j^y} - \left(\frac{D_k^x \kappa + D_k^y (\frac{\kappa + C_h}{2})}{D_k^x + D_k^y} \right)$$

Since $(D_j^x + D_j^y) = (D_k^x + D_k^y)$ it suffices to consider the difference

$$(D_k^x + 1)(\frac{\kappa + C_h}{2}) + (D_k^y - 1)\kappa - \left(D_k^x \kappa + D_k^y (\frac{\kappa + C_h}{2}) \right)$$

which simplifies to

$$r = (\frac{C_h - \kappa}{2})(D_k^x - D_k^y + 1)$$

Now, $r > 0 \Rightarrow E_j^d(.) > E_k^d(.)$ and n_k should be prefered over n_j. This is the case when $(D_k^x - D_k^y + 1) > 0$.

Since $x_k < x_d$, $D_k^x \geq 1$. As $y_h = y_d$ and $y_k = y_h - 1$, we must have $D_k^y = 1$. Therefore, $(D_k^x - D_k^y + 1) > 0$ and a routing decision based on U^2 will route the message M to n_k on its way to the destination n_d.

For Case 3 senarios, U^2 uses U^1 which, by Corollary 2, will yield an optimal path Π. This proves Theorem 2 $\Box$.

3 Discussion and Summary

Decision theory and artificial intelligence provide a range of tools that can be useful in the design of intelligent, adaptive, self-managing communication networks. Decision and control tasks that arise in such networks (e.g., routing decisions made at each node, actions taken to balance the load across the entire network, etc.) have to attempt to satisfy as closely as possible, multiple, and often conflicting, performance criteria. Examples of such performance criteria include: network throughput, maximum tolerable delay, maximum tolerable message loss, average delay, degree of load balancing, etc. Conventional routing and control mechanisms rely on relatively up-to-date information about the state of the entire network. Hence, in large communication networks with thousands of nodes distributed over a wide area, they entail tremendous resource overhead in terms of memory needed at individual nodes, computation time for making decisions, and network bandwidth needed to keep the information up-to-date. The overall effect of this phenomenon include: reduced utilization of the network (in terms of network bandwidth used to actually transmit messages as opposed to information needed for network management) and/or deterioration in the quality of routing and control decisions as measured by some performance metric. This requires an understanding of the complex interactions that exist between different measures of network performance and resource requirements and the development of a coherent framework that facilitates a smooth trade-off of some of the performance measures and resource requirements against others on demand.

In this paper, we have formulated some simple utility-theoretic heuristic decision functions for guiding messages along a near-minimum-delay path in a large network. We have analyzed some of the interesting properties of such heuristics under a set of simplifying assumptions regarding network topology and load dynamics. For a regular grid network with uniform load (with the exception of a single hot spot), we have identified the precise conditions under which a simple and computationally efficient utility-theoretic heuristic decision function is guaranteed to route a message along a minimum delay path when it is assumed that the change in network load is negligible during the time it takes to make a routing decision. We have derived an upper bound on the sub-optimality of a path. We have established an upper bound on the probability that a path between a randomly chosen source-destination pair is sub-optimal by considering configurations of uniformly loaded grid networks with single hotspots under the assumption that each source-destination pair is equally likely. We have also designed a modified heuristic function that is guaranteed to yield optimal routes under the same set of assumptions about network topology, load, and load dynamics. The latter is very similar in spirit to utility functions that were developed (mostly based on intuition rather than sound mathematical foundations) and experimentally studied in [7, 8, 9, 10, 11, 12]. Indeed, the study of utility-theoretic heuristics which is described in this paper, was, at least in part, motivated by a desire to formulate the heuristc routing functions and to understand the experimental results in more precise mathematical terms.

Some natural questions to ask at this point include: How realistic or practical are the various assumptions that were made in our development and analysis of utility-theoretic heuristics for routing? How can the results be applied (if at all) to more realistic communication network environments in which assumptions regarding network topology, load, and load dynamics do not hold? How can the analysis be extended to such scenarios? How can computationally efficient utility-theoretic heuristics be designed for different sets of performance criteria for such complex and dynamic networks so that they become essentially autonomous and self-managing? Although this paper does not provide complete and satisfactory answers to all these questions, we believe that it constitutes a useful (albeit perhaps tentative) first step in that direction. In this context, a few comments are in order.

Results of a wide range of experiments using heuristics that are very similar in spirit to U^2 display the property of automatic *load balancing* [8, 9, 10, 11, 12]. This suggests that the simplifying assumption of *uniform network load* (except at a hot spot) is useful at least as a crude first approximation of a more realistic scenario. A hotspot is typically caused in such a network due to extensive influx of traffic to a particular network node (or group of nodes) or a node or link failure (which is generally assumed to be rare in modern communication networks). However, the behavior of the routing functions compensates for this change by redistributing traffic away from the hot spot. Also, given this behavior, it is reasonable to assume that the probability of several hotspots occurring simultaneously within close proximity of each other in such a network is generally quite small. A possible exception to this scenario would be a hotspot region (caused for example, by a failure of an entire subnetwork as could occur in the event of a major natural disaster). When the hotspots are not in close proximity of each other, the single hotspot assumption holds at least locally in a large network. Similarly, the uniform load assumption is also likely to hold (given the load-balancing tendency of the heuristic routing functions), at least locally (except for the discontinuity introduced by a hotspot), in a large network. These observations suggest that our analytical results are likely to be useful (at least in qualitative terms) to guide the design utility-theoretic heuristics for a a more complex network. Of course, this does not mean that it is not worthwhile to extend our analysis to a range of increasingly complex scenarios by removing some of the simplifying assumptions. Some obvious cases to consider include: allowing irregular grids; allowing non-uniform (but relatively smooth) load distribution - except at a hotspot, allowing multiple hotspots or contiguous hotspot regions (of various shapes), etc.

It is perhaps worth emphasizing that the utility function U^2 developed in this paper yields minimum delay paths if certain assumptions regarding network topology, load, and load dynamics hold - by making use of the measured uniform load in the network (and hence the delay per link). Thus, the performance of such utility-theoretic heuristics critically depends on the existence of an adequately precise *estimator* of delay (or some other performance measure) that would result from a particular routing choice. A wide range of such estimators are possible, depending (among other things) on what can be assumed regarding the network topology, load, and network dynamics. It might be useful to analyze a range of such estimators and the resulting heuristics based on different sets of such assumptions - especially since a useful strategy for designing good heuristics for complex problems is based on solution of *simplified* or *relaxed versions* of the original problem [14]. Other interesting research directions include: investigation of methods for adaptation that enable the tuning of heuristics - perhaps parameterized in some manner - using appropriate measurements of network performance as feedback in real-time - perhaps drawing upon the rich literature on adaptive control [13]; and techniques for learning [5, 6] that construct new heuristics or modify existing heuristics) as a function of measured network behavior or as a function of information gathered through directed experiments initiated by the network during otherwise idle periods.

The task of making decisions based on incomplete and uncertain information is by no means limited to communication networks. Load distribution and task scheduling in distributed computing environemnts are other examples of decision mechanisms that are attempting to maximize certain performance criteria without having access to global information upon which there decisions can be based. The tradeoff between the quality of decisions and the resource overhead associated with knowledge acquisition and maintenance is critically important to understand in any complex dynamic environment. Thus, techniques similar to the ones used in this paper might find use in analyzing spatio-temporally distributed dynamic computing and communication environments.

References

[1] Bertsekas, D. and Gallager, R. *Data Networks*. 2nd ed. Prentice-Hall, Englewood Cliffs, NJ. (1992).

[2] Comer, D. *Internetworking with TCP/IP; Principles, Protocols, and Architectures*. Prentice-Hall, Englewood Cliffs, NJ. (1988).

[3] Fishburn P.C. *Utility Theory for Decision Making* John Wiley & Sons, Inc., New York. (1970).

[4] French, S. *Decision Theory:an introduction to the mathematics of rationality.* Halsted Press, New York. (1986).

[5] Honavar, V. Toward Learning Systems That Integrate Multiple Strategies and Representations. In: Honavar, V. and Uhr, L. (ed.) *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration*, Academic Press, Boston (1994), pp. 615-644.

[6] Langley, P. *Elements of Machine Learning*, Morgan Kaufmann, Palo Alto, CA. (1995).

[7] Mikler, A.R., Honavar, V.G., and Wong, J.S.K. Simulating a Traveller: A Heuristic Approach to Routing in Large Communication Networks. *Proceedings of the European Simulation Symposium*, Dresden, Germany (1992), pp. 297-301.

[8] Mikler, A.R., Honavar, V.G., and Wong, J.S.K. Quo Vadis - A Framework for Adaptive Routing in Very Large Communication Networks. In: Alspector, J., Goodman, R., and Brown, T.X. (ed.) *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications*, Lawrence Erlbaum, Hillsdale, New Jersey (1993), pp. 196-202.

[9] Mikler, A.R., Wong, J.S.K., and Honavar, V.G. Quo Vadis - A Framework for Intelligent Routing in Large Communication Networks. Tech. Rep. TR94-24, Department of Computer Science, Iowa State University (1994). To appear in: *The Journal of Systems and Software.*

[10] Mikler, A.R., Wong, J.S.K., and Honavar, V.G. Quo Vadis - Adaptive Heuristics for Routing in Large Communication Networks. *Proc of the 3rd International Conference on Telecommunication Systems, Modeling and Analysis*, Washington, D.C. (1995), pp. 66-75.

[11] Mikler, A.R., Wong, J.S.K., and Honavar, V.G. Quo Vadis - Adaptive Heuristics for Routing in Large Communication Networks. Tech. Rep. ISU CS-TR 95-10, Department of Computer Science, Iowa State University (1995).

[12] Mikler, A.R., Wong, J.S.K., and Honavar, V.G. An Object-Oriented Approach to Modeling and Simulation of Routing in Large Communication Networks, Tech. Rep. ISU CS-TR 95-9, Department of Computer Science, Iowa State University (1995).

[13] Narendra, K.S., and Annaswamy, A.M. *Stable Adaptive Systems.* Prentice-Hall, Englewood Cliffs, NJ. (1992).

[14] Pearl, J. *Heuristics: Intelligent Search Srategies for Computer Problem Solving.* Addison-Wesley, Reading, MA. 1984.

[15] Perlman, R. *Interconnections, Bridges and Routers.* Addison-Wesley, Reading, MA. 1992.

[16] Tanenbaum, A.S. *Computer Networks.* Prentice-Hall, Englewood Cliffs, NJ. 1988.

[17] Von Neumann, J. and Morgenstern O. *Theory of Games and Economic Behavior.* John Wiley & Sons, Inc., New York. 1964.

[18] Wong, J.S.K. and Mikler, A.R. Routing Algorithms for High-Speed Communications Networks. In: Conard, J.W. (ed.), *Broadband Communications Systems.* Auerbach Publications (1993), pp. 97-105.