Dynamic Path Selection for Source Routing in Time-Varying Lossy Networks

Dacfey Dzung*, Rachid Guerraoui†, David Kozhaya† and Yvonne-Anne Pignolet*
*ABB Corporate Research
†IC, EPFL
Email: firstname.lastname@{*ch.abb.com, †epfl.ch}

***Abstract*—This paper addresses the path selection problem arising in multi-hop sensor networks, e.g., for Smart Grids. In such scenarios, a communication system consisting of multiple multi-hop paths with time-varying hops connects a source and destination. To avoid interference and keep energy consumption low, the source can only send on one path and accrue a reward determined by the state of the traversed hops.**

We provide the first mathematical formulation to the problem of optimal sequential path selection under partially observable Markov decision processes. We unveil an intriguing behavior of the myopic policy, arguably the most appealing known way to tackle this problem. We specifically prove under positively correlated hops, that this policy can get locked, i.e., permanently ignores potentially good paths. We also generalize an empirically proven good approach for the single hop case, the Whittle index, and show its intractability for the problem at hand. We propose a new metric, *Harmonic Discounted Index* (HDI), which (i) circumvents the non-optimal myopic locking and (ii) can be computed efficiently. We evaluate the performance of HDI metric within an index policy in a variety of simulation scenarios and show that routing decisions by the proposed HDI metric outperform those based on alternative index policies.

I. Introduction

A *Smart Grid* is a power distribution system enhanced with intelligent devices, such as sensors and actuators, communicating altogether to deliver new services unattainable over the current power grid. Spread at multiple places along the grid, e.g. at transformers, substations and residential premises, sensors and sensor networks play an outstanding role in areas of remote monitoring and smart metering[1]. Typical sensor networks comprise communication links of which some might be very unreliable [1], [2]. Particularly, parts of the sensor network in Smart grids will employ low power and lossy time-varying communication technologies, such as power line and wireless communication [3]–[5]. Both wireless and power line communication takes place over a shared medium and the link quality can vary a lot, even in a very short time. Due to the transmission ranges and the topologies of these networks, there are typically several multi-hop paths to select from when disseminating information.

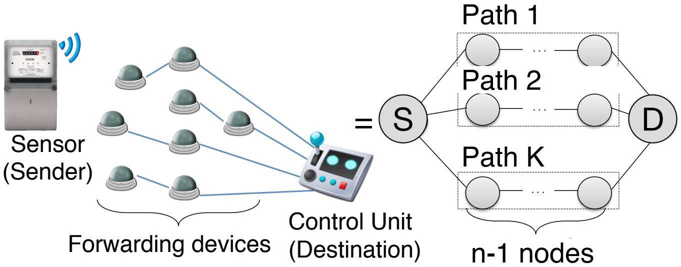

Fig. 1. An example network for automated metering connecting a source S to a destination D, by $\mathcal{K}$ independent n-hop paths.

This paper addresses the path selection problem arising in source routing for multi-hop sensor networks in Smart Grids. Some examples of existing source routing protocols for such networks (which decide on the whole path) are RPL non-storing mode [6] and Dynamic Source Routing (DSR) [7]. To keep the network load low and avoid collisions, the goal is to minimize the number of retransmissions through "smart" routing decisions at the source, under constrained knowledge of the states of the underlying lossy hops. More precisely, we consider a network where a sender has access to multiple independent multi-hop paths (see Fig. 1), but is restricted to transmitting on one of them at any given point in time to avoid interference and keep energy consumption low (energy is not a focus in this work as we consider devices to be main-powered and not battery-powered). We study how a sender can intelligently utilize past observations and the knowledge of the stochastic properties of individual hops to make routing decisions that maximize the number of successfully delivered messages. Existing routing metrics that account for retransmissions, such as the expected transmission count (ETX) [8], may require extra messages in the network by periodically broadcasting "*probe packets*" to measure the delivery ratios of links. In this prospect, optimizing retransmissions under constrained knowledge on the network state reduces the overhead, be it relative to retransmissions or to network state discovery.

We consider individual hops to be lossy time-varying communication links. In most routing studies, the time-varying behavior of hops is not explicitly modeled. In this paper however, each hop is modeled as a 2-state discrete Markov chain (Fig. 2) known as the *Gilbert-Elliot* (GE) [9], [10]. The GE model has been widely used in [11]–[14] and is a simple model of time-varying channel behavior [5], [15], [16]. The reliable state, noted G, for each hop corresponds to a probability of successful transmission $p = 1$. The unreliable state, noted B, corresponds to a transmission success probability of $p = 0$. The transition probabilities between the reliable and the unreliable state can accommodate for the relatively slow processes affecting power line communication quality such

[1] Sensors monitor the functioning of grid devices and temperature, provide outage detection and detect power quality disturbances. Smart Meters allow for real-time determination of energy consumption and for reading the current consumption locally and remotely.

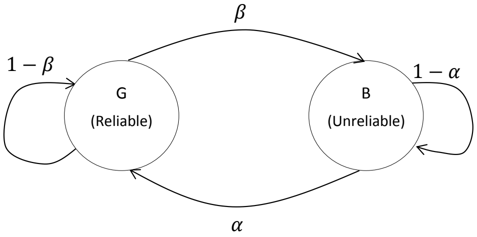

Fig. 2. Gilbert-Elliot model: a communication hop is modeled with a 2-state Markov chain.

as switching of the power grid and activation of electrical equipment, hence level crossings or state transitions typically occur only every few hours [17], [18]. In contrast, typical wireless devices in Smart Grid applications are fixed installations and operate in a steady state. Time varying behavior occurs due to occasional shadowing, but these effects are typically measured in seconds or minutes, i.e., the wireless transition probabilities are different from power line communication links. Transition probabilities can be determined in practice by techniques like [19].

Contributions. Previous path selection work has either focused on single-hop decisions or simpler hop models with constant transmission success probabilities. To the best of our knowledge, our paper is the first to study multiple non-identical hop sequences (paths). This allows for better optimization of source routing decisions in Smart Grid networks where information is most likely to be disseminated over multiple hops. Accordingly, this paper presents the first formulation of the problem at hand in the context of partially observable Markov decision processes, where a successful message delivery is associated with a unit reward. We first prove that the myopic policy, arguably the most appealing known way to tackle the problem, is optimal under *memory-less* hops. However, under positively correlated hops where non-opitmality is expected, we uncover, for the first time, an interesting locking behavior of this policy. That being a behavior where it can permanently stop selecting potentially good paths for transmission. We also generalize an empirically proven good performing index in single-hop cases [20]–[22], the Whittle index, and show its intractability for multi-hop paths.

We present a new metric, *Harmonic Discounted Index* (HDI), which (i) circumvents the non-optimal myopic locking and (ii) can be computed efficiently. HDI measures the attractiveness of transmitting over a path, using Whittle indicies of individual hops[2]. We develop an index-based protocol using our HDI metric, to establish a selection policy that governs the routing decisions taken at the source. We empirically evaluate the performance of our HDI-based policy in a variety of simulation scenarios (i) illustrating HDI's circumvention of the non-optimality of the myopic performance in locking (a performance gap between HDI and myopic of $\sim 20\%$) and (2) showing that the routing decisions relative to the proposed HDI metric outperform all alternative index policies.

Road map. The rest of the paper is organized as follows: Section II presents the related work. Section III defines the model, the assumptions and the behavior of hops in more precise terms. Section IV presents a detailed formulation of the path selection problem. Section V presents and analyses two index policies for n-hop paths: myopic and Whittle. Section VI proposes the new path selection metric, HDI, and the HDI-based index protocol. Section VII explains the experimental setup, describes alternative index policies and performs performance evaluations in various network scenarios. Finally Section VIII concludes the paper and presents potential future extensions of the path selection study.

II. Related work

Partially observable Markov decision processes (POMDPs) are widely used in control theory [25]; however, they are in general notoriously intractable [26], [27]. For single hop paths, the path selection problem can be also considered as a special case of the restless bandit problem first introduced by *Whittle* in [28]. This has been proved by [29] to be PSPACE hard, even in the special case where transition rules are deterministic. Various POMDP and restless bandit formulations have been broadly applied to several domains, of which we mention some. Multichannel opportunistic access is one of these domains. This problem has been studied in [22], [24], [30]–[32] under different assumptions. In general, the multichannel opportunistic access problem considers a sender who has to sense and transmit on one of multiple accessible channels, where each evolves independently, regardless of being sensed or not. In comparison with the path selection problem considered in this paper, the sender in our case has access to n-hop paths, where each hop along a path is an independent Markov process that evolves at all times whether it was used for transmission or not. The work in [30], [31] studied the mutlichannel access problem with channels that are independent and identically distributed (i.i.d.) Markov processes. In fact [30], [31] showed that (a) the myopic policy under these assumptions admits a simple universal structure, and (b) guarantees optimality when channels are assumed to be positively correlated, i.e. $1-\beta > \alpha$ (Section III). The authors in [24], [32] studied the same problem however without requiring channels to be i.i.d. They formulated the problem as a special case of the restless bandit problem and referred to it as *FEEDBACK MAB*. The *FEEDBACK MAB* problem has been studied under the average expected reward and an approximation algorithm is proposed with a performance guarantee of 2. The work in [22] studied a similar formulation of this problem for both positively and negatively correlated channels which are not necessarily identically distributed. In particular, [22] established the indexability of this class of restless bandits and obtained Whittle index in closed form for both discounted and average reward criteria. It was also shown [22] that the Whittle index is optimal under certain conditions when the channels are i.i.d.

III. System Model

We consider a network of $\mathcal{K}$ independent n-hop paths connecting the source (sender) and the destination (Fig. 1). We are interested in a simple non-trivial network where interesting analytic results can be extracted. Specifically, for analytical tractability reasons, paths are assumed independent and the

[2] Despite its wide use in the single hop case, the known theoretical guarantees [21], [23] for the Whittle index (even in such simpler cases) are very weak and theoretical analysis remains elusive and challenging, mainly because of the highly-coupled and complex dynamics it possesses [24].

underlying hops are modeled as independent Markov chains with only two states. At any point in time, the source is restricted to choose exactly one path for transmission. Once a decision is made and a path is selected, the message is transmitted along the selected path going through each of the underlying hops consecutively in 1 time unit. All hops along all paths are assumed to evolve at every time unit, whether they were used for transmission or not. In other words, If a hop is currently in the reliable state (G), it will remain at the next time unit in this reliable state with probability $(1-\beta)$ or will shift with probability β to the unreliable state (B) (analogously the next state is determined by α if it currently is in the unreliable state). In other words, all hops evolve at discrete time units, i.e., at $t = \{0, 1, 2, 3, ..., \infty\}$. Decisions at the source are based on the knowledge of the stochastic properties of individual hops and on past observations. When deciding about what path to use for transmission, the source is assumed to know the α's and β's of all hops[3] (no other node is required to know these parameters) but not the current state of any of the hops. Techniques such as those in [19] can be used to determine these transition probabilities and make them known to the source. Once a choice is made and a message m is sent over the selected path, the following actions are executed.

1. m goes sequentially through the underlying hops of the selected path as long as they are reliable.

2. If m traverses an unreliable hop, it is entirely dropped. All consecutive hops will not be traversed.

3. In case a message is dropped, a *packet-drop detection mechanism* informs the source (before the source decides on a new message) about the hop which led to the message loss[4]. This assumption ensures that in case of message loss the source can rightfully guess state information about the hops from the source up to and including the lossy hop. Alternatively, the source knows that m successfully reached the destination, if nothing is heard from this detection mechanism after n time units of sending m (which conveys state information about all n hops).

4. The source decides on a message to be sent (be it a new message or a retransmission) every n time units, i.e., after the previous message has had the chance to traverse all hops. This essentially leads the decision times at the source, $T = \{T_0, T_1, T_2, ...T_\infty\}$, to be deterministic, i.e. occurring at $t = \{0, n, 2n, 3n, ..., \infty\}$. This assumption is solely considered for "neater" mathematical illustration of results and to avoid unneeded notation complexity. Our theoretical results (sections V-A and V-B) hold under a relaxed version which assumes that the source can transmit a new message either after n time units (if the previous transmission is successful) or in the time unit directly following a notification from the packet-drop detection mechanism (if the previous transmission failed).

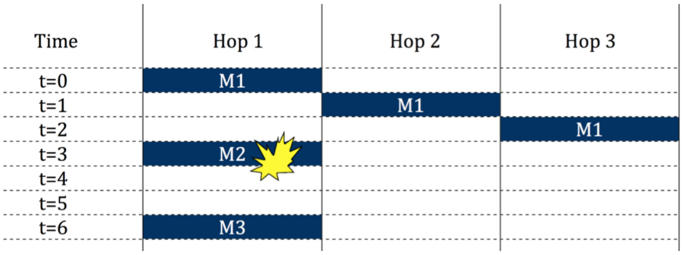

Fig. 3. Illustration of message propagation in a single 3-hop path.

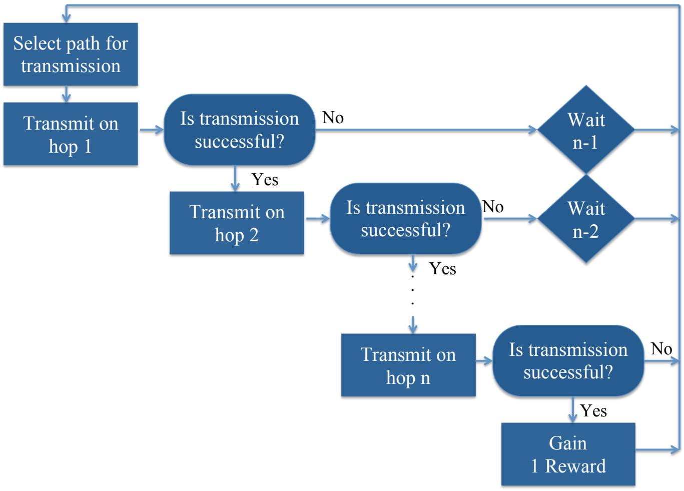
Select path for transmission
Transmit on hop 1
Is transmission successful?
No
Wait n-1
Yes
Transmit on hop 2
Is transmission successful?
No
Wait n-2
Yes
Transmit on hop n
Is transmission successful?
No
Yes
Gain 1 Reward

Fig. 4. Flow diagram of the path selection problem for an n-hop path.

These steps are illustrated by a timing diagram showing messages propagating over a 3-hop path (Fig. 3) and a flow diagram (Fig. 4). The objective of the source, given this model, is to maximize the discounted expected number of successfully delivered messages (messages reaching the destination) over infinite decision times, as we consider delay sensitive communication.

Model Variations. Illustrating the generality of our work, we show in the following section how our formulation can naturally extend to some variations, namely by relaxing some assumptions. We examine particularly two possibilities: (1) model $\mathcal{A}$: when a message is lost, the source does not have to wait but can start a new message transmission directly in the following time step and (2) model $\mathcal{B}$: the available paths have different number of hops. Our theoretical results about myopic locking (Section V-A) and Whittle intractability (Section V-B) extend as well to both of these models, but due to space constraints we do not provide further analyses on this matter in this work.

IV. Problem Formulation

We denote by $S(t) = [S_1(t),\ S_2(t), ...,\ S_{\mathcal{K}}(t)]$ the set of states of the $\mathcal{K}$ available n-hop paths where:

$$S_k(t) = \{s_{k,1}(t),\ s_{k,2}(t), ...,\ s_{k,n}(t)\}$$

such that $s_{k,i}(t) \in \{G, B\}$ is the state of i^{th} hop along path k. Let $a(T_j) = [a_1(T_j),\ a_2(T_j), ...,\ a_{\mathcal{K}}(T_j)] \in [0,1]^{\mathcal{K}}$ be the vector of actions taken at decision time T_j, where $a_k(T_j) = 1$

[3] Cases where the link quality might change, e.g. due to daily variations of interference on wireless hops, can be easily incorporated by changing the system parameters, which can be determined [19].

[4] The exact nature of the packet-drop detection mechanism is not of interest in this work, which is only a first step towards a solution of the general problem. Delayed and incorrect detection for a more general problem are beyond the scope of this work.

($a_k(T_j) = 0$) means transmitting (not transmitting resp.) over the k^{th} path at decision time T_j. Thus, $a_k(T_j) = 1$ implies:

$$a_{k,i}(t)|_{t=j\cdot n+i-1} = 1 \quad \forall i \in [1, n] \tag{1}$$

subject to:

$$\sum_{i=1}^{n} a_{k,i}(t)|_{t=t} = 1 \tag{2}$$

and $a_k(T_j) = 0$ implies:

$$a_{k,i'}(t)|_{t=j\cdot n+i-1} = 0 \quad \forall i', i \in [1, n] \tag{3}$$

Equations (1) and (2) mean that, when path k is selected for transmission at decision time T_j, every hop along path k is used only once in the time interval $[j \cdot n, (j+1) \cdot n - 1]$, such that the first hop along path k is used first, then the second hop, etc. Equation (3) means that when path k is not selected for transmission at decision time T_j, none of its hops are used in the time interval $[j \cdot n, (j+1) \cdot n - 1]$. The action vector $a(T_j)$ corresponds to the routing decision taken at decision time T_j.

A. Path State Information

Since the source operates under partial information of the state the hops are in, and since not all states can be observed, this problem can be transformed into a partially observable Markov decision process (POMDP) with all past and current state information contained in a sufficient statistic known as the belief [26]. This hop belief is the conditional probability over the hop state space. In our problem, we assume independent paths, with stochastically independent hops. Accordingly, we maintain independently for each hop, a belief,

$$w_{k,i}(t) \ \forall k \in [1, \mathcal{K}], i \in [1, n]$$

where

$$w_{k,i}(t)$$

is the conditional probability that the relative hop is in the reliable state, given all previous feedback obtained for that hop. Initially, the hop belief is set to the stationary probability, $w_{k,i}(t)|_{t=0} = \frac{\alpha_{k,i}}{\alpha_{k,i}+\beta_{k,i}}$. Afterwards, and at every time unit, this belief is updated independently for each hop as follows:

$$w_{k,i}(t+1) = \begin{cases} 1 - \beta_{k,i} & a_{k,i}(t) = 1, s_{k,i}(t) = G, \\ \alpha_{k,i} & a_{k,i}(t) = 1, s_{k,i}(t) = B, \\ \tau(w_{k,i}(t)) & a_{k,i}(t) = 0. \end{cases}$$

where: $\tau(w_{k,i}(t)) = (1 - \beta_{k,i})w_{k,i}(t) + \alpha_{k,i}(1 - w_{k,i}(t))$.

The source is the sole place deciding which path to use for transmission. It should thus account for the states of the hops that the message (to be sent) might see when this message reaches these specific hops. We represent this information by a belief vector Ω_k:

$$\begin{aligned} \Omega_k(t) &= [w_{k,1}(t),\ w_{k,2}(t+1),\ w_{k,3}(t+2), ..., w_{k,n}(t+n-1)] \\ &= [w_{k,1}(t),\ \tau(w_{k,2}(t)),\ \tau(\tau(w_{k,3}(t))), ..., \tau^{n-1}(w_{k,n}(t))], \end{aligned} \tag{4}$$

where: $\tau^x(w_{k,i}(t)) = \underbrace{\tau(\tau(...\tau(w_{k,i}(t))))}_{x \text{ times}}$, and $\tau^0(w_{k,i}(t)) = w_{k,i}(t)$.

The recursive call of function τ is done relative to the position of the hop on that path, as a message needs 1 time unit to traverse a hop. The problem considers $\mathcal{K}$ paths to select from; the source node thus keeps belief vectors of all paths, denoted by $P(t) = [\Omega_1(t),\ \Omega_2(t), ...,\ \Omega_{\mathcal{K}}(t)]$.

B. Expected Discounted Reward

We assume that every successful message delivery corresponds to a 1 unit of reward. The expected total discounted reward averages the accumulated discounted rewards obtained for a sequence of actions/decisions over time. The maximization of this function thus constitutes the objective sought. Denote by π the set of all actions vectors, i.e., $a(T)$ $\forall T$. Consequently, π constitutes the routing policy. Let $R_{a(T_j)}$ be the reward obtained relative to the action vector $a(T_j)$ at decision time T_j. The expected total discounted reward over infinite decision times, given an initial belief vector P is expressed by:

$$E_\pi \left[\sum_{T_j \in T} \gamma^{\ (\frac{T_j}{n})} R_{a(T_j)} | P \right] \tag{5}$$

subject to the constraint: $\sum_{k=1}^{\mathcal{K}} a_k(T_j) = 1$, where $\gamma : 0 < \gamma < 1$ is the discounted factor. The constraint implies that, at any decision time, exactly one path is used for transmission.

Value Function. Denote by $V_\gamma(P)$ the value function, i.e. the maximum expected total discounted reward obtained by the optimal policy under the set of initial belief vectors P. Then:

$$V_\gamma(P) = \max_k \{V_\gamma(P; a_k = 1)\} \quad \forall k \in \mathcal{K}$$

where $V_\gamma(P; a_k = 1)$ denotes the total expected discounted reward from selecting path k for transmission at first followed by the optimal policy in future decision times. The expression for $V_\gamma(P; a_k = 1)$ can be obtained according to Bellman's equation [5], [33]. For clearer illustration, we develop the expression of $V_\gamma(P; a_k = 1)$ for the case of $\mathcal{K} = 2$ and $n = 2$, where $P = [\Omega_1, \Omega_2]$ is the set of belief vectors such that $\Omega_k = [w_{k,1}, \tau(w_{k,2})]$.

$$\begin{aligned} V_\gamma(P(t); a_1 = 1) = w_{1,1}\tau(w_{1,2}) & \\ + \gamma[w_{1,1}\tau(w_{1,2})V_\gamma(\tau(1 - \beta_{1,1}), \tau(1 - \beta_{1,2}), \tau^2(w_{2,1}), \tau^3(w_{2,2})) & \\ + w_{1,1}(1 - \tau(w_{1,2}))V_\gamma(\tau(1 - \beta_{1,1}), \tau(\alpha_{1,2}), \tau^2(w_{2,1}), \tau^3(w_{2,2})) & \\ + (1 - w_{1,1})V_\gamma(\tau(\alpha_{1,1}), \tau^3(w_{1,2}), \tau^2(w_{2,1}), \tau^3(w_{2,2}))] & \end{aligned} \tag{6}$$

$$\begin{aligned} V_\gamma(P(t); a_2 = 1) = w_{2,1}\tau(w_{2,2}) & \\ + \gamma[w_{2,1}\tau(w_{2,2})V_\gamma(\tau^2(w_{1,1}), \tau^3(w_{1,2}), \tau(1 - \beta_{2,1}), \tau(1 - \beta_{2,2})) & \\ + w_{2,1}(1 - \tau(w_{2,2}))V_\gamma(\tau^2(w_{1,1}), \tau^3(w_{1,2}), \tau(1 - \beta_{2,1}), \tau(\alpha_{2,2})) & \\ + (1 - w_{2,1})V_\gamma(\tau^2(w_{1,1}), \tau^3(w_{1,2}), \tau(\alpha_{2,1}), \tau^3(w_{2,2}))] & \end{aligned} \tag{7}$$

The derivation details for (6) and (7) can be found in Appendix A. Following similar steps (without the need to show more details), we illustrate the formulation of $V_\gamma(P; a_k = 1)$ under the two model variants $\mathcal{A}$ and $\mathcal{B}$. We assume that the discounted factor is applied per decision, i.e., the power of the discounted factor increased by 1 with every new transmission regardless of the elapsed transmission time.

Model $\mathcal{A}$: We also consider the case of $\mathcal{K} = 2$ and $n = 2$, where $P = [\Omega_1, \Omega_2]$ is the set of belief vectors such that

$\Omega_k = [w_{k,1}, \tau(w_{k,2})]$.

$$\begin{aligned}V_\gamma(P(t); a_1 = 1) = {}& w_{1,1}\tau(w_{1,2})\\ &+ \gamma[w_{1,1}\tau(w_{1,2})V_\gamma(\tau(1-\beta_{1,1}), \tau(1-\beta_{1,2}), \tau^2(w_{2,1}), \tau^3(w_{2,2}))\\ &+ w_{1,1}(1-\tau(w_{1,2}))V_\gamma(\tau(1-\beta_{1,1}), \tau(\alpha_{1,2}), \tau^2(w_{2,1}), \tau^3(w_{2,2}))\\ &+ (1-w_{1,1})V_\gamma(\alpha_{1,1}, \tau^2(w_{1,2}), \tau(w_{2,1}), \tau^2(w_{2,2}))]\end{aligned}$$

$$\begin{aligned}V_\gamma(P(t); a_2 = 1) = {}& w_{2,1}\tau(w_{2,2})\\ &+ \gamma[w_{2,1}\tau(w_{2,2})V_\gamma(\tau^2(w_{1,1}), \tau^3(w_{1,2}), \tau(1-\beta_{2,1}), \tau(1-\beta_{2,2}))\\ &+ w_{2,1}(1-\tau(w_{2,2}))V_\gamma(\tau^2(w_{1,1}), \tau^3(w_{1,2}), \tau(1-\beta_{2,1}), \tau(\alpha_{2,2}))\\ &+ (1-w_{2,1})V_\gamma(\tau(w_{1,1}), \tau^2(w_{1,2}), \alpha_{2,1}, \tau^2(w_{2,2}))]\end{aligned}$$

Model $\mathcal{B}$: We develop the expression of $V_\gamma(P; a_k = 1)$ for $\mathcal{K} = 2$. We consider the first path to have one hop and the other to have two hops. $P = [\Omega_1, \Omega_2]$ is thus the set of belief vectors such that $\Omega_1 = [w_{1,1}]$ and $\Omega_2 = [w_{2,1}, \tau(w_{2,2})]$.

$$\begin{aligned}V_\gamma(P(t); a_1 = 1) = {}& w_{1,1}\\ &+ \gamma[w_{1,1}V_\gamma(1-\beta_{1,1}), \tau(w_{2,1}), \tau^2(w_{2,2}))\\ &+ (1-w_{1,1})V_\gamma(\alpha_{1,1}, \tau^3(w_{1,2}), \tau(w_{2,1}), \tau^2(w_{2,2}))]\end{aligned}$$

$$\begin{aligned}V_\gamma(P(t); a_2 = 1) = {}& w_{2,1}\tau(w_{2,2})\\ &+ \gamma[w_{2,1}\tau(w_{2,2})V_\gamma(\tau^2(w_{1,1}), \tau(1-\beta_{2,1}), \tau(1-\beta_{2,2}))\\ &+ w_{2,1}(1-\tau(w_{2,2}))V_\gamma(\tau^2(w_{1,1}), \tau(1-\beta_{2,1}), \tau(\alpha_{2,2}))\\ &+ (1-w_{2,1})V_\gamma(\tau(w_{1,1}), \alpha_{2,1}, \tau^2(w_{2,2}))]\end{aligned}$$

$V_\gamma(P; a_k = 1)$ can be split into two main components: one which represents the expected immediate reward relative to selecting path k and a second representing the discounted future reward resulting from to choosing path k at the first decision.

$$V_\gamma(P; a_k = 1) = \underbrace{w_{k,1}\tau(w_{k,2})}_{\text{immediate expected reward}} + \underbrace{\gamma[...]}_{\text{discounted future reward}}$$

Each term (V_γ), in the future reward, is a function of a joint set of beliefs each spanning the set of real values in the interval $[0, 1]$. The dimensions of V_γ grow with $n\mathcal{K}$ and thus the required computations increase immensely as $\mathcal{K}$ and n increase. Solving the Bellman equation pertaining to problem (5), and hence obtaining the optimal policy, with a general purpose solver for POMDP, becomes rapidly intractable (even for the simpler case when increasing $\mathcal{K}$ for $n = 1$). Therefore, efficient near-optimal solution methods are sought.

V. Index Policies: Formulation and Analysis

Index policies are selection protocols that assign an *index*, to each of the $\mathcal{K}$ paths. This index evaluates how rewarding it is to select a path under a particular state. At every decision time, an index policy selects the path with the highest index. Some path indices are strongly decomposable, i.e. can be computed separately for each path, without regard of the states of other paths. This reduces the complexity of the problem as compared to a full POMDP solution. We examine two such index policies for multi-hop paths: (i) Myopic policy and (ii) Whittle index.

A. Myopic Performance

The myopic policy assigns the immediate expected reward of a path as its index. This significantly reduces the amount of computation required as any possible effect of the future discounted reward on decision making is disregarded. It has been shown for stochastically identical single hop paths (see Section II) that such a myopic strategy guarantees an optimal solution [30], [31]. However, to the best of our knowledge, not much has been said about this policy for the case of non-identical hops with equal rewards, and more importantly, for multi-hop paths.

1) Entirely Memory-less Hops: We first show a case of non-identical hops, where future rewards do not contribute to decision making in multi-hop paths: in this case the myopic policy is optimal. The Markovian process governing the evolution of the state of hops (described in Section I) becomes entirely memory-less when the probability of being in a state at time t is the same, regardless whether the state at $t-1$ was reliable or unreliable, i.e., $1 - \beta_{k,j} = \alpha_{k,j}$. The belief as a result remains constant at all times

$$w_{k,j} = 1 - \beta_{k,j} = \alpha_{k,j} = \tau(w_{k,j}). \tag{8}$$

Proposition 1. *For a set of $\mathcal{K}$ paths, each consisting of n entirely memory-less hops, the myopic policy is optimal.*

Proof: The value function of the corresponding general case can be written as follows: $V_\gamma(P) = \max_{a \in a(T)} [R(P, a) + \gamma \sum_{P'} Pr(P'|a, P)V_\gamma(P')]$, where P' is the belief vector at the following decision time. Regardless of what action is chosen, the value of P' will always be the same. This follows directly from (8) since all hop beliefs will remain constant. Hence $V_\gamma(P') = V_\gamma(P)$ and $Pr(P'|a, P) = 1$ and therefore $V_\gamma(P) = \frac{\max_{a \in a(t)}[R(P,a)]}{1-\gamma}$. This proves that the myopic policy is indeed optimal. ■

Furthermore, we can conclude from (8) that a single path may have the highest expected immediate reward at all times. The optimal policy in this case transmits over one path only at all decision times.

2) Positively Correlated Hops: We study now a more general case where hops are positively correlated, i.e. $1 - \beta_{k,i} > \alpha_{k,i}\ \forall k \in \mathcal{K}$. We show that the optimality of the myopic performance is not guaranteed. More importantly, we identify a condition under which the myopic policy gets locked, meaning that it avoids the selection of certain paths regardless of how reliable they could be. We first consider single hop paths only, i.e., one transmission hop per path (the hop notation will be omitted). The results are generalized later in this section to include n-hop paths. The belief of a single hop path has the following two important characteristics [22], [34]:

$\alpha_k < \tau(w_k) < 1 - \beta_k$.

$$\tau^t(w_k) \text{ monotonically converges to } w_0^k = \frac{\alpha_k}{\alpha_k + \beta_k} \text{ as } t \to \infty. \tag{9}$$

Lemma 1. *The stationary probability of a positively correlated single hop path satisfies $\alpha_k < w_0^k < 1 - \beta_k$.*

Proof: This follows directly from the positive correlation assumption ($\alpha_k + \beta_k < 1$). ■

Theorem 1. *If a single hop path, k, exists such that for any other path $k' \in \mathcal{K}$, $w_0^k < \alpha_{k'}$, then the myopic policy will never select path k for transmission after the first time k is observed in the unreliable state.*

Proof: When a path k is selected and observed to be in an unreliable state, its belief takes the value $w_k = \alpha_k$

(refer to the update function Section IV). Hops are positively correlated, thus by *Lemma* 1 we have: $\alpha_k < w_0^k < \alpha_{k'}$. However $\min\{w_{k'}\} = \alpha_{k'}$, so by (9) $w_{k'} > w_k$ at all times, hence concluding the proof. ∎

Corollary 1. *If the beliefs are initialized to their stationary probabilities*[5]*, then path k will never be selected by the myopic policy for transmission.*

IMPORTANT: It is crucial to note that *Theorem* 1, on its own, does not necessarily indicate that the myopic performance is not good. In fact, *Theorem* 1 also applies to the entirely memory-less case (Section V-A1) where the myopic policy is indeed optimal. The significance of *Theorem* 1 on the quality of the myopic routing decisions is determined by the stochastic properties of the neglected paths. In particular, hops with a small α_k and β_k are the most influential on the myopic performance. Such hops have a low switching probability and tend to stay in their current state for long periods, a behavior which resembles that of power line communication hops [5]. Missing transmission on these hops when their current state is reliable is expected to affect the quality of the routing decisions (this claim is confirmed in Section VII). A simple example of two single hop paths, illustrated below, conveys the effects. Consider a source with two paths: $Path_1$: $\alpha_1 = 0.6$; $1 - \beta_1 = 0.65$ (frequently switching resembling wireless hops) and $Path_2$: $\alpha_2 = 0.1$; $1 - \beta_2 = 0.93$ (slow switching resembling power line hops). A source selecting paths according to the myopic policy, will never transmit on $Path_2$ after it observes it in an unreliable state for the first time[6]. Therefore it does not make use of the fact that $Path_2$ can return to the reliable state at a later time. Transmitting on $Path_1$ which switches more frequently will be less rewarding than transmitting on $Path_2$ when it is in the reliable state. Thus neglecting $Path_2$ forever is expected to yield less reward.

Theorem 2. *If an n-hop path k exists such that for any path $k' \in \mathcal{K}$:*

$$w_0^{k,f} \prod_{h=1}^{f-1} \tau^{n-1}(1-\beta_{k,h}) \prod_{l=f+1}^{n} \tau^{n+l-1}(1-\beta_{k,l}) < \prod_{r=1}^{n} \tau^{r-1}(\alpha_{k',r})$$

for any $f \in [1, n]$, then path k will never be selected by the myopic policy after the first time its f^{th} hop is observed in an unreliable state.

Details deriving Theorem 2 can be found in Appendix B. We thus conclude that the myopic routing decisions could lead to performances that are not optimal, where the significance of the performance gap between the myopic and optimal solution is determined by the stochastic properties of the available paths.

B. Whittle Index: A Path Formulation

n this section we discuss the *Whittle* index [28]. We generalize and formulate this index for multi-hop paths. The Whittle index of a path depends merely on the properties of that particular path and not of other paths. Accordingly, it is enough to consider a single n-hop path. Given a single path, at each decision time the source can make one of two possible actions (i) use that path for transmission (make it active) or idle that path (make it passive). An optimal policy would partition the path state space into a passive and an active set where it is optimal to idle the path or use it for transmission respectively. The Whittle index measures the attractiveness of transmitting over a path under a subsidy, λ. That is, we consider a multi-hop path where a constant subsidy[7], λ, is obtained whenever the path is idled. Clearly this subsidy λ affects how the state space is optimally partitioned between the active and the passive set. States which remain active under a larger subsidy are thus more attractive to the source. Based on this intuition, the minimum subsidy required to move a given state from the active set to the passive set constitutes a measure of how attractive that state is [22].

More precisely, we denote by $V_{\gamma,\lambda}(P)$ the value function corresponding to the maximum expected total discounted reward that can be obtained from a single path with subsidy λ and belief vector P (we drop the path index from the notation). Denote by $V_{\gamma,\lambda}(P; a)$ the total expected discounted reward from taking action a at the first decision time followed by the optimal policy in the future. Thus:

$$V_{\gamma,\lambda}(P) = \max\{V_{\gamma,\lambda}(P; a = 0), V_{\gamma,\lambda}(P; a = 1)\}. \qquad (10)$$

As an illustration we consider a 2-hop path under a given subsidy λ. The value functions relative to taking the active and passive actions on this path in the first decision can be derived as in Section V-A and are respectively written as:

$$\begin{aligned} V_{\gamma,\lambda}(P; a = 1) = {} & w_1\tau(w_2) \\ & + \gamma[w_1\tau(w_2)V_{\gamma,\lambda}(\tau(1-\beta_1), \tau(1-\beta_2)) \\ & + w_1(1-\tau(w_2))V_{\gamma,\lambda}(\tau(1-\beta_1), \tau(\alpha_2)) \\ & + (1-w_1)V_{\gamma,\lambda}(\tau(\alpha_1), \tau^3(w_2))]. \end{aligned} \qquad (11)$$

$$V_{\gamma,\lambda}(P; a = 0) = \lambda + \gamma V_{\gamma,\lambda}(\tau^2(w_1), \tau^3(w_2)). \qquad (12)$$

Definition 1. *The passive set $\mathcal{P}(\lambda)$ is the set of path states for which it is optimal to make the path passive under subsidy λ.*

Following our belief formulation in Section IV-A, a path state is represented through the belief vector Ω. Accordingly, we define the passive set for the 2-hop path as: $\mathcal{P}(\lambda) = \{[w_1, \tau(w_2] : V_{\gamma,\lambda}(P; a = 0) \geq V_{\gamma,\lambda}(P; a = 1)\}$ and generalize it for an n-hop path as:

$$\mathcal{P}(\lambda) = \{[w_1, \tau(w_2), ..., \tau^{n-1}(w_n)] : V_{\gamma,\lambda}(P; a = 0) \geq V_{\gamma,\lambda}(P; a = 1)\}. \qquad (13)$$

Whittle index for a path would be meaningful if it results in some consistency when making paths passive. In other words, a path made passive under subsidy λ should also be made passive under a subsidy $\lambda' > \lambda$. We thus define *indexability*:

Definition 2. *A path is said to be indexable if its passive set $\mathcal{P}(\lambda)$ increases from $\emptyset$ to the whole state space of $[0, 1]^n$ as λ increases from $-\infty$ to $+\infty$.*

If the path is indexable, the Whittle index is the infimum subsidy λ which makes the passive and active actions equally

[5] This initialization of beliefs to the their stationary value is assumed in most of the previous literature [22], [30], [31].

[6] Assuming the initial belief values allow it to be selected in the first place.

[7] The subsidy, λ, can be thought of as a counter reward for not using the path for transmission.

rewarding and is expressed as:

$$W(P) = \inf_{\lambda}\{\lambda : V_{\gamma,\lambda}(P; a=0) = V_{\gamma,\lambda}(P; a=1)\}. \quad (14)$$

In other words, the Whittle index is the infimum value of the subsidy λ which makes the source indifferent between using a path for transmission or idling it. A larger index indicates that the path is more attractive, in the sense that it requires a higher subsidy to be made passive. A source can choose at every decision time the path with the highest Whittle index for transmission. It is important to notice though, that the dimensions of $V_{\gamma,\lambda}$ in (11) and (12) grow with the number of hops n and the state space hence expands to $[0,1]^n$. Consequently solving for Whittle index (14) as n increases becomes intractable. This implies that computing the Whittle index efficiently for an n-hop path may not be feasible.

VI. Harmonic Discounted Index (HDI)

Given the intractability of the Whittle index for a multi-hop path, the goal is to design a tractable path metric that reflects the degree of attractiveness of transmitting over a path at a given point in time. We first advocate the intuition behind our path metric, after which we formally present it. A hop can be correlated with a simple conducting wire. A poorly conducting wire, which renders a propagating signal undetectable by a receiver, is equivalent to a "poorly attractive" (unreliable) hop which leads to the loss of the message being transmitted. The attractiveness of a hop can thus be directly correlated with a conductance measure. Recall that hops along a path are assumed to be stochastically independent, i.e., every hop independently exists in the good or bad state relative to its own 2-state Markov chain. Accordingly, every hop would constitute an independent wire which has its own conductance. A multi-hop path, in this prospect, becomes a sequence of multiple conducting wires connected in series. The metric embodying the attractiveness of transmitting over the path hence translates to the equivalent conductance of the series combination. We start first by determining the *Hop Conductance*.

A. Hop Conductance

We define the hop conductance as the measure of the attractiveness of transmitting on a hop along a path. This attractiveness entitles (i) the attractiveness of transmitting on the medium constituting the hop and (ii) the feedback attractiveness relative to the position of this hop of the path. We explain in details these two factors in the following Sections.

1) Attractiveness of Hop Transmission Medium: We consider each hop on its own regardless of all other hops. This effectively transforms a single n-hop path to n separate 1-hop paths. We consider that the source can transmit on these hops independently, hence reducing the set of decision times T to $t = \{0, 1, 2, ..., \infty\}$. The set of belief vectors of an n-hop path $\Omega = [w_1, \tau(w_2), ..., \tau^{n-1}(w_n)]$ transforms under this decomposition to $[\Omega_1, \Omega_2, ..., \Omega_n]$ where $\Omega_i = w_i \ \forall i \in [1,n]$ since every path now has only 1 hop. Given this decomposition, we consider next a single hop under the subsidy concept and drop the index i from the notation. We measure the attractiveness of transmitting on this hop by calculating its corresponding Whittle index. Accordingly, $V_{\gamma,\lambda}(w)$, the maximum expected total discounted reward that can be obtained from a hop with subsidy λ is: $V_{\gamma,\lambda}(w) = \max\{V_{\gamma,\lambda}(w; a=0), V_{\gamma,\lambda}(w; a=1)\}$. $V_{\gamma,\lambda}(w; a)$ denotes the total expected discounted reward from taking action a as the first decision followed by the optimal decisions in the future and can be written as

$$V_{\gamma,\lambda}(w; a=1) = w + \gamma[wV_{\gamma,\lambda}(1-\beta) + (1-w)V_{\gamma,\lambda}(\alpha)],$$

$$V_{\gamma,\lambda}(w; a=0) = \lambda + \gamma[V_{\gamma,\lambda}(\tau(w))].$$

The passive set $\mathcal{P}(\lambda)$ for a single hop reduces to

$$\mathcal{P}(\lambda) = \{w : V_{\gamma,\lambda}(w_i; a=0) \geq V_{\gamma,\lambda}(w_i; a=1)\}. \quad (15)$$

The passive set in (13) describes a property for a whole n-hop path, while in (15) it describes a property for a single hop and thus constitutes a per-hop description. In other words, (15) is a decomposition of (13) that follows naturally from the decomposition of the belief vector Ω of the n-hop path. A single hop is said to be indexable if $\mathcal{P}(\lambda)$ increases from $\emptyset$ to the state space of $[0,1]$ as λ increases from $-\infty$ to $+\infty$. If the hop is indexable, then its corresponding Whittle index is $W(P) = \inf_{\lambda}\{\lambda : V_{\gamma,\lambda}(w; a=0) = V_{\gamma,\lambda}(w; a=1)\}$.

Important: The Whittle index under this formulation admits a very efficient way of being computed. In fact, a closed form expression of Whittle index under this single hop formulation is established in [22]. Thus, we obtain n measures of attractiveness, $W_i \ \forall i \in [1,n]$, for each of the n mediums constituting the individual hops with negligible overhead.

2) Attractiveness of Hop Feedback: The amount of information revealed to the source relative to transmitting on a path affects the source's later decisions. When transmitting over some path, losing a message at any of its underlying hops will yield the same result of no immediate reward. However the amount of information revealed to source is not the same. Information about the states of the hops on a path are obtained up to the hop leading to message loss (inclusive) (refer to Section III). The amount of obtained information, hence increases as a message traverses more hops of a path even if it is destined to failure. It is important to note that although this information is useless for the current reward, it may be of fundamental value affecting the future rewards (since the obtained information affects subsequent decisions). Consider an example of two 3-hop paths, each decomposed into three 1-hop paths with the following indices: $Path_1$: $W_1 = 0.2$, $W_2 = 0.6$, $W_3 = 0.94$ and $Path_2$: $W_1 = 0.94$, $W_2 = 0.6$, $W_3 = 0.2$. Despite the fact that $Path_1$ and $Path_2$ consist of hops with similar indices, the fact that these hops are positioned differently along $Path_1$ and $Path_2$, makes these paths not equivalent. In fact, one can notice that the amount of information revealed to the source is indeed different and is expected to be higher if $Path_2$ is favored over $Path_1$, especially that we wait n time slots regardless of the destiny of the transmission[8]. The feedback attractiveness of a hop becomes less significant as the hop is further from the source, as obtaining feedback relative to that hop is prone to more uncertainty. We account for this feedback attractiveness by embedding a discounting attractiveness index, which serves to decrease the contribution of W_is of more distant hops:

[8] The only inefficiency of selecting $Path_2$ over $Path_1$ may be the extra transmission energy. But as indicated in Section I, energy is not a focus in this work as we assume main-powered devices.

Definition 3. *Let DI_i be an index measuring the hop conductance of i^{th} hop within an n-hop path.*

$$DI_i = \delta^{i-1} W_i,$$

where W_i is the Whittle index of the i^{th} hop (not the Whittle index of the whole path) and $0 < \delta < 1$ is a discounting factor[9].

B. Path Conductance

Given an n-hop path with the associated hop conductances, the metric measuring the attractiveness of transmitting over a path reduces thus to the overall path conductance. The path conductance in an n-hop path is the equivalent conductance of a series combination of hop conductances.

Definition 4. *The path conductance is equivalent to $\frac{1}{n}^{th}$ of the harmonic mean of the n individual hop conductances (DI_i) associated with underlying hops.*

The harmonic mean of a set of values tends strongly toward the smallest values in that set. It has a tendency to increase the impact of small values and alleviate the influence of large outliers. In paths, the smaller the individual hop conductances ($DI_i(s)$) are, the less attractive they are for transmission. Therefore the harmonic mean of these individual measures is most influenced by the least attractive hops along a path. This can be fairly justified by the fact that a single unreliable hop across the whole path is enough to make the whole path bad (yielding no reward) regardless of how many and how good other hops on that path are. We formally define the path metric:

Definition 5. *Let* Harmonic Discounted Index *(HDI) be the measure the attractiveness of transmitting over a path*

$$HDI = \left[\sum_{i=1}^{n} \left(\frac{1}{DI_i}\right)\right]^{-1}. \tag{16}$$

Theorem 3. *The HDI metric circumvents the non-optimal myopic locking and can be computed in* $\mathrm{O}(\mathcal{K}n)$.

Proof:
Computation Complexity. The $\mathrm{O}(\mathcal{K}n)$ computation follows directly from (i) (16) and (ii) the fact that the DI indices can be determined in $\mathrm{O}(1,)$ as a result of the established closed expressions in [22], [35].
Circumvention. For better illustration, we show how HDI circumvents the non-optimal myopic locking in the simple case of 1-hop paths. Avoiding the myopic locking in this simple case, leads Theorem 1 being not satisfied. As Theorem 2 is an extension of Theorem 1, having the latter unsatisfied implies that the former will not hold as well.

Under the simple case of 1-hop path, our HDI metric reduces to the simple Whittle index of a hop, which admits an known closed form [22], [35]. Theorem 1 says that if $w_0^k < \alpha_{k'}$, then the myopic policy always selects path k' as $w_k < w_{k'}$ will be satisfied at all times. We alternatively show that if $w_0^k < \alpha_{k'}$, then the Whittle index of k, noted by $W_k(w)$, is not always less than the Whittle index of path k', noted by $W_{k'}(w')$. Consequently, if $W_k(w)$ can be greater than $W_{k'}(w')$, then path k can be selected by the source under HDI: Thus circumventing the myopic locking.

From the closed forms of the Whittle index in [35] we have: $W_{k'}(\alpha_{k'}) = \alpha_{k'}$ and $W_k(w_0^k) = \frac{w_0^k}{1-\gamma(1-\beta_k - w_0^k)}$. It is clear that $W_k(w_0^k) = \frac{w_0^k}{1-\gamma(1-\beta_k - w_0^k)} > w_0^k$. This implies that for $w_0^k < \alpha_{k'}$ the Whittle index may still allow path k to selected for transmission circumventing the myopic locking. ■

VII. Experimental Evaluation

This section describes the experimental setup and illustrates performance evaluations of our HDI metric when embedded within an index policy in a variety of simulation scenarios. We evaluate an index policy which transmits at every time on the path with the highest index, where we vary this index between different alternatives such as the myopic index, HDI index and other indices based on different ways of combining hop conductances. The alternative index policies are briefly listed below:
Mnlog index: This policy computes the $\log$ of the hop conductances along a path and selects at every decision time the path with the highest mean of logs, i.e., $\max_{\mathcal{K}}\{\frac{1}{n}\sum_{i=1}^{n} \log(\delta^{i-1}W_i)\}$. Such a policy, similar to our HDI metric, reduces the impact of the good hops on the overall path quality.
Min: The policy thus selects the path which has the highest minimum hop conductance, i.e., $\max_{\mathcal{K}}\{\min_{i\in[1,n]} \delta^{i-1}W_i\}$.
Sum: This policy selects the path which has the highest sum of hop conductances, i.e., $\max_{\mathcal{K}}\{\sum_{i=1}^{n} \delta^{i-1}W_i\}$.
Prod: In this case, the overall path conductance is determined by computing the product of individual hop conductances. The path with the highest product is selected, i.e., $\max_{\mathcal{K}}\{\prod_{i=1}^{n} W_i\}$.
HI: In this policy the feedback attractiveness factor, δ_i, is disregarded. The hop conductance is thus equivalent to the hop's transmission attractiveness alone. The path which has the highest harmonic mean is selected, i.e., $\max_{\mathcal{K}}\{\left[\sum_{i=1}^{n}(\frac{1}{W_i})\right]^{-1}\}$.

In all simulations, the transition probabilities of hops are generated uniformly at random within given bounds (specified per case) and obeying the positive correlation assumption. The discounted parameters are fixed to the values $\gamma = 0.95$ and $\delta = 0.95$. For every set of randomly generated paths, 10^4 runs are repeated, where in each run the discounted reward representing the successful message transmissions is accumulated for a horizon of 10^4 decision times. The reported reward is the mean value of the accumulated discounted rewards over all runs, scaled by $1 - \gamma$.

A. Exploring the Transition Space of Hops

We first compare the performance of the HDI metric against flooding, i.e., transmitting every message on all available paths. Flooding clearly is an upper bound on the optimal solution. We run our simulations over a wide scope of transition probabilities. In particular, we divide the search space in $[0, 1]$ into eight categories $\{L1, L2, L3, L4, H1, H2, H3, H4\}$,

[9] In Section VII-A, we specifically evaluate the impact of this discounting attractiveness factor, δ, showing the performance gain of HDI metric in (16) over an HDI variant without δ.

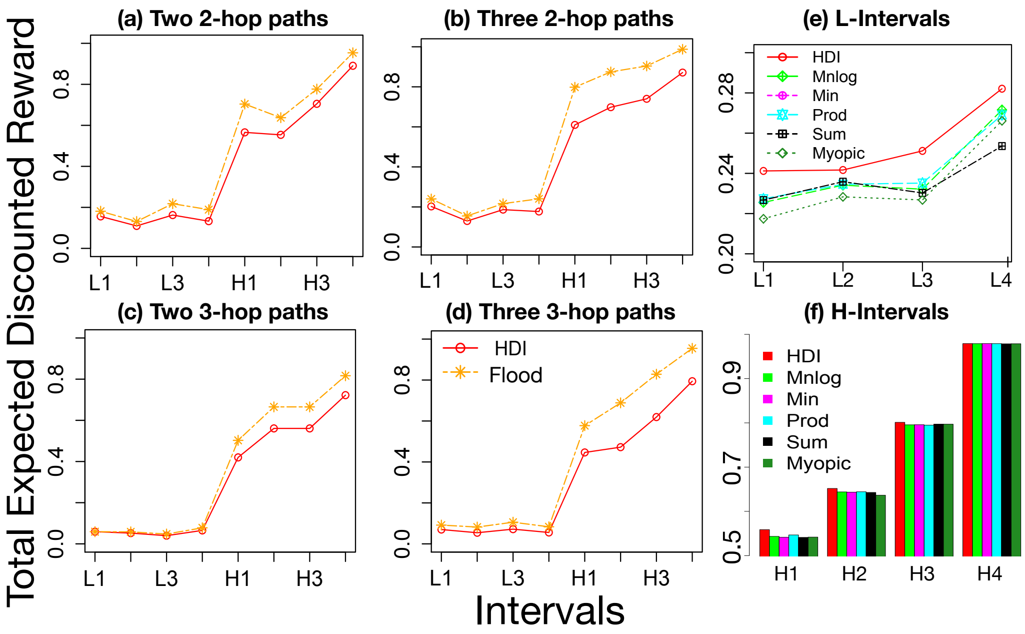

Fig. 5. Performance of HDI index policy over L and H-Intervals.

ranging from hops that slowly switch between the reliable and unreliable state to those that switch very frequently. Our Results, Fig. 5 $(a - d)$, show performance evaluations in networks consisting respectively of two $2-$hop paths, three $2-$hop paths, two $3-$hop paths and three $3-$hop paths.

	L1	**L2**	**L3**	**L4**
α	$[0, \beta]$	$[0.1, \beta]$	$[0.2, \beta]$	$[0.3, \beta]$
β	$[0.1, 0.2[$	$[0.2, 0.3[$	$[0.3, 0.4[$	$[0.4, 0.5[$

	H1	**H2**	**H3**	**H4**
α	$[0, 1-\beta]$	$[0.2, 1-\beta]$	$[0.3, 1-\beta]$	$[0.4, 1-\beta]$
$1-\beta$	$[0.6, 1[$	$[0.7, 1[$	$[0.8, 1[$	$[0.9, 1[$

The performance of our HDI metric is close to that of flooding for the ranges $L1$ through $L4$, indicating a close to optimal performance in these respective ranges. However for ranges $H1$ through $H4$, it can be seen that the performance gap grows bigger[10]. This can be attributed to that fact that hops in these ranges tend to switch rapidly between states, spending more time in the reliable state. In these cases uncertainty increases, making decisions between exploitation versus exploration prone to more randomness. Flooding, in comparison with a policy which selects one path only, explores all other potentials, who in this case, have a high probability of being good. We also evaluate the performance of HDI versus the alternative policies by separating the search space to L and H intervals. It can be seen (Fig. 5 (e, f)) that HDI outperforms all alternatives over all ranges. This improvement nonetheless gradually decreases as hops go higher in the H interval.

B. Smart Grid Sensor Network with Frequent and Slow Switching Hops.

Typical smart grid sensory networks contain heterogeneous hops of wireless and power line communication hops [3]–[5]. We simulate such typical smart grid scenario by combining frequent and slow switching hops [5], [17], [18]. We consider a network having $\mathcal{K}$ independent 2-hop paths available for transmission. Hops along a path are generated uniformly at random as either slow switching (L1 range) or fast switching (ranges H3 and H4). We illustrate the performance measures for different number of available paths varying between $\mathcal{K} = \{2, 3, 4, ..., 10\}$. Our results, Fig. 6 (a), show that the HDI selection policy outperforms the myopic policy for all numbers of available paths. We further strengthen the significance of the improvement obtained by our HDI metric by limiting the number of available paths to 2 and comparing the performance with that of flooding for a number of hops per path, spanning $n = \{2, 3, 4\}$. Our results, Fig. 6 (b), show that despite the narrow margin for improvement, our HDI metric succeeds in making a positive improvement over the myopic performance, showing a close-to-optimal performance.

For the two networking scenarios in this section, we also show the performance gain relative to the feedback attractiveness factor δ. The performance evaluations confirming its benefit are shown in Fig 7.

C. Myopic Performance Under Locking.

n this section we confirm (i) the deterioration of myopic performance under locking and (ii) the ability of HDI metric to circumvent it. We generate a first set of $\mathcal{K}$ independent n-hop paths where hops are randomly generated satisfying $0.7 < 1 - \beta < 0.85$ and $0.6 < \alpha < 1 - \beta$. Such hops switch frequently between states, representing a behavior similar to that of wireless channels. A second set of $\mathcal{K}'$ n-hop paths (of the same size as $\mathcal{K}$) are also created. Every path in $\mathcal{K}'$ satisfies *Theorem* 2 with some path $k \in \mathcal{K}$, i.e. these paths will be neglected by the myopic policy. Paths in set $\mathcal{K}'$ are chosen to be slow switching, similar to the behavior of power line communication hops [5]. A source can transmit

[10] This gap is noticed to grow bigger as well when the number of available paths increases, which is perfectly explainable by the nature of flooding.

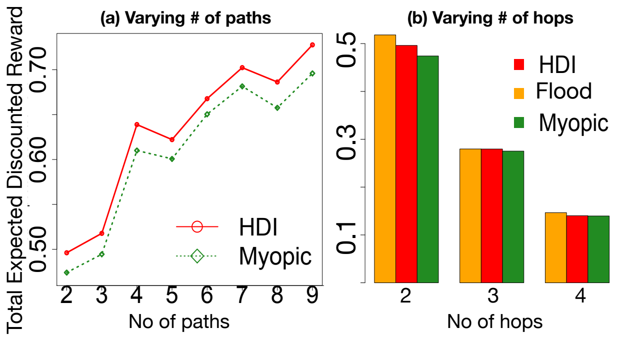

Fig. 6. Performance of selection policies with slow and fast switching hops.

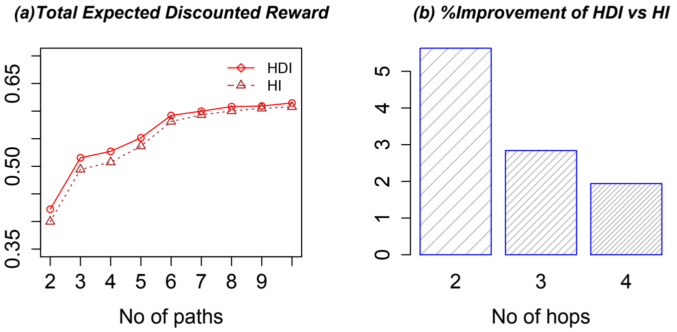

Fig. 7. Contribution of feedback attractiveness factor δ_i.

on any path within these two sets. Simulations are carried for $(|\mathcal{K}| + |\mathcal{K}'|) = \{4, 6, 8, ..., 20\}$ available paths and for $n = \{1, 2, 3, 4\}$ hops/path. Results in Fig. 8 (a-c) show a noticeable deterioration in the myopic performance where our HDI metric manages to reach an improvement of $\sim 20\%$ over myopic.

Paths with different number of hops. We slightly modify our simulation to allow paths with different numbers of hops. In particular, every path is generated with a size of $n = \{2, 3, 4\}$, chosen uniformly at random. Our results, Fig. 8 (d), show that the myopic deterioration extends to such cases where our HDI metric benefits from any available "short" good paths and maintains $\sim 20\%$ improvement over myopic.

We also run evaluate our HDI index against a per-hop myopic (greedy) index. Results show that a per-hop technique performs worse that source routing myopic (this is expected as per-hop neglects the impact of hops further down a path).

VIII. Conclusion

This paper presents a formulation of the path selection problem as a partially observable Markov decision process, for optimizing source routing decisions over multiple time-varying paths. We show that, while the greedy myopic policy is easy to be computed and optimal for stochastically non-identical paths with memory-less hops, the myopic policy can lead to bad performances. More precisely, it might get stuck in bad states where it avoids transmission on certain potentially good paths. Furthermore we devise a generalization of the Whittle index, known by previous literature for its good performance. However, we show that this index becomes intractable for multi-hop paths as the number of hops increases. We present HDI, a new "*efficient to compute*" path selection metric that circumvents the non-optimality of myopic locking. We evaluate experimentally the performance of an index-based HDI policy in various simulation scenarios. We illustrate that the HDI policy outperforms other alternatives index policies. Some directions for future work involve addressing more challenging versions of the path selection problem, e.g. under delayed or unreliable packet-drop mechanisms.

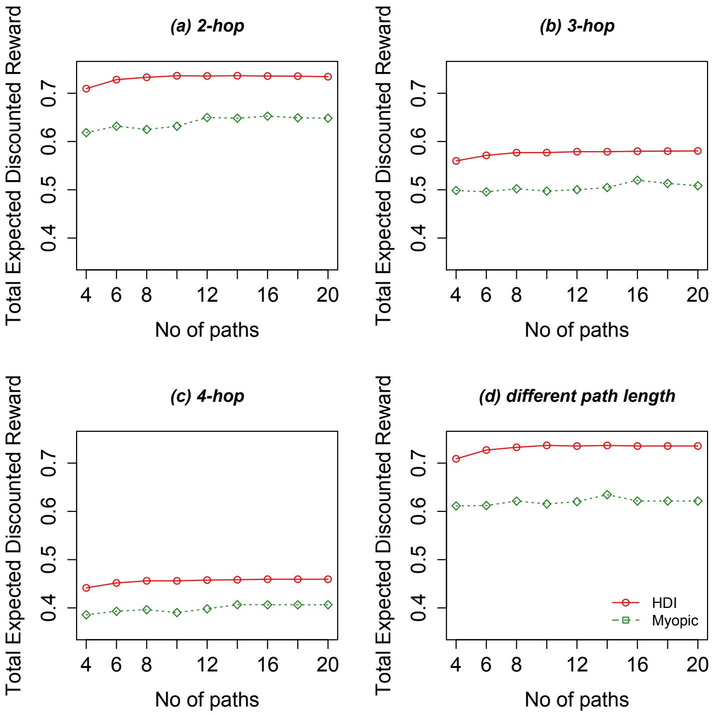

Fig. 8. Performance of HDI versus Myopic under locking.

References

[1] J. Vasseur, "Terminology in low power and lossy networks," Cisco Systems, Inc, Tech. Rep., 2013. [Online]. Available: http://tools.ietf.org/html/draft-ietf-roll-terminology-12

[2] J. Ko, A. Terzis, S. Dawson-Haggerty, D. Culler, J. Hui, and P. Levis, "Connecting low-power and lossy networks to the internet," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 96–101, April 2011.

[3] G. Bumiller, L. Lampe, and H. Hrasnica, "Power line communication networks for large-scale control and automation systems," *IEEE Communications Magazine*, vol. 48, no. 4, pp. 106–113, 2010.

[4] DLC+VIT4IP, "D1.1 scenarios and requirements specification," DLC-VIT4IP, Tech. Rep., 2010. [Online]. Available: http://www.dlc-vit4ip.org/wb/media/Downloads

[5] D. Dzung and Y.-A. Pignolet, "Dynamic selection of wireless/powerline links using markov decision processes," *IEEE Conference on Smart Grid Communication*, 2013.

[6] T. Winter, P. Thubert, A. Brandt, J. Hui, R. Kelsey, P. Levis, K. Pister, R. Struik, J. Vasseur, and R. Alexander, "Rpl: Ipv6 routing protocol for low-power and lossy networks," Internet Engineering Task Force (IETF), Tech. Rep., 2012. [Online]. Available: http://tools.ietf.org/html/rfc6550

[7] D. M. D. Johnson, Y. Hu, "The dynamic source routing protocol (dsr) for mobile ad hoc networks for ipv4," Internet Engineering Task Force (IETF), Tech. Rep., 2007. [Online]. Available: http://tools.ietf.org/html/rfc4728

[8] D. S. J. De Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," *Wirel. Netw.*, vol. 11, no. 4, pp. 419–434, Jul. 2005.

[9] E. N. Gilbert *et al.*, "Capacity of a burst-noise channel," *Bell Syst. Tech. J*, vol. 39, no. 9, pp. 1253–1265, 1960.

[10] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J*, vol. 42, no. 5, pp. 1977–1997, 1963.

[11] L. Johnston and V. Krishnamurthy, "Opportunistic file transfer over a fading channel: A pomdp search theory formulation with optimal threshold policies," *Wireless Communications, IEEE Transactions on*, vol. 5, no. 2, pp. 394–405, Feb 2006.

[12] D. Zhang and K. Wasserman, "Transmission schemes for time-varying wireless channels with partial state observations," in *INFOCOM 2002*, vol. 2, June 2002, pp. 467–476.

[13] A. Laourine and L. Tong, "Betting on gilbert-elliot channels," *Wireless Communications, IEEE Transactions on*, vol. 9, no. 2, pp. 723–733, February 2010.

[14] D. Dzung, R. Guerraoui, D. Kozhaya, and Y. A. Pignolet, "To transmit now or not to transmit now," in *SRDS*, 2015, pp. 246–255.

[15] C. Tang and P. K. McKinley, "Modeling multicast packet losses in wireless lans," in *Proceedings of the 6th ACM International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWIM '03. New York, NY, USA: ACM, 2003, pp. 130–133.

[16] J. pierre Ebert, A. Willig, D. ing Adam Wolisz, and T. Berlin, "A gilbert-elliot bit error model and the efficient use in packet level simulation," Tech. Rep., 1999. [Online]. Available: http://www2.tkn.tu-berlin.de/publications/papers/tkn_report02.pdf

[17] R. Rao, S. Akella, and G. Guley, "Power line carrier (plc) signal analysis of smart meters for outlier detection," in *Smart Grid Communications, 2011 IEEE International Conference on*. IEEE, 2011, pp. 291–296.

[18] J.-P. Vasseur and A. Dunkels, *Interconnecting smart objects with ip: The next internet*. Morgan Kaufmann, 2010.

[19] G. Hasslinger and O. Hohlfeld, "The gilbert-elliott model for packet loss in real time services on the internet," in *Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008 14th GI/ITG Conference* -, March 2008, pp. 1–15.

[20] K. Glazebrook and H. Mitchell, "An index policy for a stochastic scheduling model with improving/deteriorating jobs," *Naval Research Logistics*, vol. 49, no. 7, pp. 706–721, 2002.

[21] "Whittle's index policy for a multi-class queueing system with convex holding costs," *Mathematical Methods of Operations Research*, vol. 57, no. 1, 2003.

[22] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theor.*, vol. 56, no. 11, Nov. 2010.

[23] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of Applied Probability*, vol. 27, no. 3, pp. pp. 637–648, 1990.

[24] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," *J. ACM*, vol. 58, no. 1, pp. 1–50, Dec. 2010.

[25] W. Jiang, J. Tang, and B. Krishnamachari, "Optimal power allocation policy over two identical gilbert-elliott channels," in *Communications (ICC), 2013 IEEE International Conference on*, June 2013, pp. 5893–5897.

[26] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Athena Scientific, 2000.

[27] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, pp. 99–134, 1998.

[28] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, pp. 287–298, 1988.

[29] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Math. Oper. Res.*, vol. 24, no. 2, pp. 293–305, Feb. 1999.

[30] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multichannel opportunistic access: structure, optimality, and performance," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431–5440, 2008.

[31] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theor.*, vol. 55, no. 9, pp. 4040 –4050, sept. 2009.

[32] S. Guha and K. Munagala, "Approximation algorithms for partial-information based stochastic control with markovian rewards," *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pp. 483–493, 2007.

[33] D. Braziunas, "Pomdp solution methods," University of Toronto, Tech. Rep., 2003.

[34] N. Nayyar, Y. Gai, and B. Krishnamachari, "On a restless multi-armed bandit problem with non-identical arms," *49th Annual Allerton Conference on Communication, Control, and Computing*, pp. 369 –376, sept. 2011.

[35] J. Ny, M. Dahleh, and E. Feron, "Multi-uav dynamic routing with partial observations using restless bandit allocation indices," *American Control Conference, 2008*, pp. 4220–4225, 2008.

Appendix A
Derivation of the Value Function

The general Bellman equation can be written as:

$$V_\gamma(P) = \max_{a \in a(T)} \left[R(P,a) + \gamma \sum_{P'} Pr(P'/a,P) V_\gamma(P') \right] \tag{17}$$

Activating first path

Activating the first path by taking action $a = \{a_1 = 1, a_2 = 0\}$ is represented by:

$$\begin{aligned} V_\gamma(P; a_1 = 1) &= R(P, a_1 = 1) \\ &\quad + \gamma \sum_{P'} Pr(P'|a_1 = 1, P) V_\gamma(P') \end{aligned}$$

$R(P, a_1 = 1)$ is the expected reward obtained from activating the first path under belief vector $P = [\Omega_1, \Omega_2]$. The expected reward obtained from activating the first path with hops beliefs $\Omega_1 = [w_{1,1}, \tau(w_{1,2})]$ is the probability that all hops of a path are reliable (since otherwise the reward is 0). Hops along a path are stochastically independent. Therefore

$$R(P, a_1 = 1) = w_{1,1} \cdot \tau(w_{1,2})$$

Now we determine the set of elements in P', where P' is the updated belief vector relative activating the first path. First we state the following fact that follows from the model definition Section III:

Fact 1. *A source can detect a maximum of one unreliable hop per message transmission.*

Based on *Fact* 1, a source can observe the path in a total of $n+1$ states after being used for transmission (1 state where all hops are reliable and n other states where in each 1 hop is unreliable). $P' = [\Omega'_1, \Omega'_2]$ is updated relative to the observed path state (refer to Section IV-A) and thus $|P'| = n + 1$ (in this particular case $|P'| = 3$). The states that can be observed by the source are:

1) **First hop reliable and second hop reliable**. Assuming the first decision is taken at decision time T_j, then at time $t = j + 1$ the first hop is observed to be reliable. At time $t = j+2$ the second hop is observed to be reliable while the first hop is idled. Knowing that the hops of second path are idled for 2 time units, the individual hop beliefs at the next decision time $T_{j+1} = j + 2$ are:

$$w_{1,1} = \tau(1 - \beta_{1,1}); \ w_{1,2} = 1 - \beta_{1,2}$$

Therefore the updated belief vector

$$\begin{aligned} P_1' &= [w_{1,1},\ \tau(w_{1,2});\ w_{2,1},\ \tau(w_{2,2})] \\ &= [\tau(1-\beta_{1,1}),\ \tau(1-\beta_{1,2});\ \tau^2(w_{2,1}),\ \tau^3(w_{2,2})] \end{aligned}$$

and is observed with probability

$$Pr(P_1'/a_1 = 1, P) = w_{1,1} \cdot \tau(w_{1,2})$$

2) **First hop reliable and second hop unreliable**. This is similar to first case, except that the second hop is observed unreliable, So:

$$w_{1,1} = \tau(1-\beta_{1,1});\ w_{1,2} = \alpha_{1,2}$$

Therefore the updated belief vector

$$P_2' = [\tau(1-\beta_{1,1}),\ \tau(\alpha_{1,2});\ \tau^2(w_{2,1}),\ \tau^3(w_{2,2})]$$

and is observed with probability

$$Pr(P_2'/a_1 = 1, P) = w_{1,1} \cdot (1 - \tau(w_{1,2}))$$

3) **First hop unreliable**. Following a similar logic as that of the previous two cases, at time $t = j+1$ the first hop is observed to be unreliable and the message is lost revealing no information about the second hop. Consequently at the next decision time T_{j+1} (equivalent to $t = j+2$)the first hop would have idled once while the second hop would have idled twice. The individual hop beliefs at the next decision time, T_{j+1} are:

$$w_{1,1} = \tau(\alpha_{1,1});\ w_{1,2} = \tau^2(w_{1,2})$$

Therefore the updated belief vector

$$\begin{aligned} P_3' &= [w_{1,1},\ \tau(w_{1,2});\ w_{2,1},\ \tau(w_{2,2})] \\ &= [\tau(\alpha_{1,1}),\ \tau^3(w_{1,2});\ \tau^2(w_{2,1}),\ \tau^3(w_{2,2})] \end{aligned}$$

and is observed with probability

$$Pr(P_3'/a_1 = 1, P) = 1 - w_{1,1}$$

Thus:

$$\begin{aligned} V_\gamma(P; a_1 = 1) = {} & w_{1,1}\tau(w_{1,2}) \\ & + \gamma[w_{1,1}\tau(w_{1,2})V_\gamma(\tau(1-\beta_{1,1}), \tau(1-\beta_{1,2}), \tau^2(w_{2,1}), \tau^3(w_{2,2})) \\ & + w_{1,1}(1-\tau(w_{1,2}))V_\gamma(\tau(1-\beta_{1,1}), \tau(\alpha_{1,2}), \tau^2(w_{2,1}), \tau^3(w_{2,2})) \\ & + (1-w_{1,1})V_\gamma(\tau(\alpha_{1,1}), \tau^3(w_{1,2}), \tau^2(w_{2,1}), \tau^3(w_{2,2}))] \end{aligned}$$

The same logic can be applied to the obtain the expression of $V_\gamma(P; a_2 = 1)$.

Appendix B
Derivation of Theorem 2

Consider an n-hop path k'. The smallest value of its myopic index at any time is:

$$\min\left\{\prod_{r=1}^{n} \tau^{r-1}(w_{k',r})\right\} = \prod_{r=1}^{n} \min\{\tau^{r-1}(w_{k',r})\}$$

The minimum value of the belief of the i^{th} hop on path k' that the source can observe is $\tau^{i-1}(\alpha_{k',i})$. This verifies the (RHS) of the Theorem 2.

Observing the f^{th} hop in the unreliable state means:

1) All hops before f were reliable and will have beliefs:

$$\tau^{n-h}(1-\beta_{k,h}) \quad \forall h \in [1, f-1]$$

2) All hops after f where idled for n times, thus:

$$\tau^{n}(w_{k,l}) \quad \forall l \in [f+1, n]$$

3) the f^{th} hop is observed unreliable and idled for $n-f$ times:

$$\tau^{n-f}(\alpha_{k,f})$$

Given that:

$$P = [w_{k,1},\ \tau(w_{k,2}), ...,\ \tau^{n-1}(w_{k,n})]$$

The myopic index of path k after its f^{th} hop is observed unreliable is thus:

$$\tau^{n-1}(\alpha_{k,f}) \prod_{h=1}^{f-1} \tau^{n-1}(1-\beta_{k,h}) \prod_{l=f+1}^{n} \tau^{n+l-1}(w_{k,l})$$

To guarantee that path k will never get selected after its f^{th} hop is observed unreliable, it is sufficient that the maximum value of myopic index of path k, after observing the f^{th} hop unreliable, is smaller than the minimum myopic index of path k'.

$$\max\left\{\tau^{n-1}(\alpha_{k,f}) \prod_{h=1}^{f-1} \tau^{n-1}(1-\beta_{k,h}) \prod_{l=f+1}^{n} \tau^{n+l-1}(w_{k,l})\right\} < \prod_{r=1}^{n} \tau^{r-1}(\alpha_{k',r}) \tag{18}$$

The LHS of (18) can be written as:

$$\max\left\{\tau^{n-1}(\alpha_{k,f})\right\} \max\left\{\prod_{h=1}^{f-1} \tau^{n-1}(1-\beta_{k,h})\right\} \max\left\{\prod_{l=f+1}^{n} \tau^{n+l-1}(w_{k,l})\right\}$$

Since hops are positively correlated, $\tau^t(\alpha_{k,f}) \to w_0^{k,f}$ as $t \to \infty$:

$$\max\left\{\tau^{n-1}(w_{k,f})\right\} = w_0^{k,f}$$

$$\max\left\{\prod_{h=1}^{f-1} \tau^{n-1}(1-\beta_{k,h})\right\} = \prod_{h=1}^{f-1} \tau^{n-1}(1-\beta_{k,h})$$

and:

$$\begin{aligned} \max\left\{\prod_{l=f+1}^{n} \tau^{n+l-1}(w_{k,l})\right\} &= \prod_{l=f+1}^{n} \max\left\{\tau^{n+l-1}(w_{k,l})\right\} \\ &= \prod_{l=f+1}^{n} \tau^{n+l-1}(1-\beta_{k,l}) \end{aligned}$$