Local Relay Selection in Presence of Dynamic Obstacles in Millimeter Wave D2D Communication

Durgesh Singh, Arpan Chattopadhyay & Sasthi C. Ghosh

Abstract—Blockage due to obstacles in millimeter wave (mmWave) device to device (D2D) communication is a prominent problem due to their severe penetration losses. Potential user equipments (UEs) in vicinity of the source UE must be explored in order to select a new relay when the current link gets blocked. However, dynamic obstacles are not known in advance and thus may cause unpredictable fluctuations to D2D channel quality causing newly selected relay link also to be susceptible to blockage. This might cause frequent relay switching leading to call drops and high energy consumption. We have proposed the idea of reducing frequency in relay exploration and switching and thus average end-to-end delay (in seconds) at the expense of additional exploration time units (few milliseconds) during beam alignment. We seek to learn the uncertainty in D2D link qualities by modeling the problem as finite horizon partially observable Markov decision process (POMDP) framework locally at each UE. We have derived an optimal threshold policy which counts the number of successive acknowledgment successes/failures for making decisions of selecting or not selecting a given relay locally. Through extensive simulation, we validate our theoretical findings and demonstrate that our approach captures the trade-off between average exploration time and average end-to-end (E2E) delay in presence of dynamic obstacles.

Index Terms—Relay selection, Millimeter wave D2D communication, Exploration, Dynamic Obstacles, POMDP.

I. INTRODUCTION

Device to device (D2D) communication enables proximity devices or user equipments (UEs) to directly communicate with one another bypassing the base station (BS) [1]. High available bandwidth and short range transmission of millimeter wave (mmWave) is a lucrative choice for D2D communication [2]–[4]. Although very high propagation losses of mmWave are compensated by transmitting directional beams using antennas arrays, severe penetration loss of mmWave make it susceptible to very high blockage from various obstacles [2], [5], [6]. Obstacles may block the mmWave signal completely, thus needing an almost line of sight (LOS) path. Additionally, the presence of dynamic obstacles might deteriorate D2D channel condition rapidly & abruptly causing unprecedented link breakage and hence packet loss, delay and high energy consumption.

In case of blockage by dynamic obstacle, a new relay is chosen by exploring new relay links through beam-forming [7], [8] which are in the source UE’s vicinity. This is a directional search of new relay and has a considerable delay. A detailed techniques for beam-alignment and beam-management under various scenarios can be found in [9]. The new relaying UE must be chosen carefully during exploration time (ranging from few microseconds to 10 ms [7] for short range), because it may get blocked during data transmission time (100-1000 times higher than exploration time) due to presence of dynamic obstacles, even when the source and relay beams are perfectly aligned to achieve highest data rate as shown in figure [1] This might lead to frequent relay exploration and switching which causes increased delay, energy consumption and probably outage leading to call drops. In addition, exchange of channel state information with the BS might cause extra delay [10] since the D2D channel is not directly visible to the BS. Hence the decision to explore and select a relay must be made locally.

Fig. 1. Perfectly aligned beam of source-destination during exploration susceptible for blockage due to dynamic obstacles.

Most of the works [11]–[14] deal with static obstacles. However, dynamic obstacles are unpredictable which must be captured and their effects taken into account. The authors in [6] showed a significant drop in data rate when a pedestrian act as a blockage. The authors in [15], [16] used radar to capture the movement of UEs and obstacles. Vision cameras and machine learning (ML) techniques were used by the authors in [4], [6], [17], [18] for tracking obstacle’s spatio-temporal behavior. However all these require either expensive hardware or high processing time and energy usage. These might be apt at the BS but not at the UE which needs a local solution in an online fashion. The ML based solution in [17] might need re-training when there is change in the environment. Moreover, it requires that the link breakage event must follow some well defined pattern with some known distribution. However, due to the presence of obstacles, variations in link quality behavior is

arXiv:2011.01537v1 [cs.NI] 1 Nov 2020

Durgesh Singh and Sasthi C. Ghosh are with the Advanced Computing & Microelectronics Unit, Indian Statistical Institute, Kolkata 700108, India. Email: durgesh.ccet@gmail.com, sasthi@isical.ac.in.

Arpan Chattopadhyay is with the Department of Electrical Engineering, Indian Institute of Technology Delhi. Email: arpanc@ee.iitd.ac.in.

This work was supported by the faculty seed grant and professional development allowance (PDA) of IIT Delhi.
are summarized as follows: have been derived which maps the belief to a set of actions. Our major contributions in this paper take the decision of whether selecting or not selecting the loss and in-turn delay by sending additional exploration probe packets to learn the channel quality. Optimal threshold policies can be used in modeling the variations in unobserved D2D links. The authors in [23] utilized POMDP to derive an optimal policy which tells the UE locally when to go for exploration after suspending communication on current relay link after successive packet losses. Whereas, in contrast, this work deals with selecting an appropriate relaying UE during the exploration phase locally. The probabilistic model and cost structure are also accordingly different in this work.

In our work, we have investigated the idea of reducing frequency in relay switching and thus average end-to-end (E2E) delay at the expense of additional exploration time during beam alignment. We have modeled the problem as a finite horizon POMDP. The states are the D2D relay link qualities which are not observable at the current time instant. It can only be observed after receiving the acknowledgements (ACKs) of the probe packets which are sent in order to access a link. Even the ACKs can get lost due to presence of dynamic obstacles. Information about dynamic obstacles are not known at BS a priori and it can only be learned through ACKs of probe packets. The goal is to take decision of whether selecting or not selecting a relay which minimizes packet loss and in-turn delay by sending additional exploration probe packets to learn the channel quality. Optimal threshold policies have been derived which maps the belief to a set of actions. By exploiting the derived policy structure, we have obtained a stationary policy which tells the UE during exploration that after how many successive ACK successes or ACK failures to take the decision of whether selecting or not selecting the relay respectively. Theoretical analysis is validated through extensive simulation. Our major contributions in this paper are summarized as follows:

1) We modeled relay selection problem during the exploration time locally considering the presence of dynamic obstacles as a POMDP. This model can be applied to a scenario without explicitly knowing the dynamic obstacle’s distribution.

2) The model learns the channel quality in an online fashion using ACKs which can also get lost. The model does not require to undergo a training phase and hence suitable when there is a change in the environment.

3) We showed that the optimal policy is threshold type policy. This is a non-trivial result that required proof of several interesting intermediate results. Our optimal policy can be implemented locally at each UE, thereby facilitating distributed implementation.

4) The threshold policy for the problem is further reduced to counting the number of successive ACK successes or ACK failures, which is simple and easy to implement.

II. System Model

We are considering the device-tier of 5G D2D architecture mentioned in [24], where devices or UEs can communicate among themselves with or without the help from BS. The service region is discretized into various zones or grids as shown in figure 2(a) with one BS. Each zone may have many UEs and is assumed to have at least one D2D device which is ready to take part in D2D communication as a relay or source/destination node. A zone \( i \) containing the source UE may form connection to a UE of another zone \( j \in \cup^i \), where \( \cup^i \) is the viable relay zones of the zone \( i \) which is given by the BS. A viable relaying zone of zone \( i \) is one which is nearer to the zone containing the destination UE and is in the communication range of the zone \( i \). When the source UE in the fixed zone \( i \) forms a connection with another UE of zone \( j \in \cup^i \), then the link formed between zones \( i \) and \( j \) is termed as link \( j \). Link is formed between UEs of two zones when they are in communication range of each other and the received signal strength is sufficient for the required data rate. Each UE can communicate with one another on mmWave channels using directional antennas. The received signal strength (\( Q_{ij} \)) on zone \( j \) from zone \( i \) is modeled as [13]:

\[
Q_{ij} = \mu \cdot P \cdot G_i \cdot G_j \cdot PL_{ij}
\]

where, \( \mu \) is the shadowing random variable, \( P \) is the transmit power of the UE, \( G_i \) & \( G_j \) are transmit and receive beam-forming gains respectively. \( PL_{ij} \) is the distance dependent path loss function between zone \( i \) and zone \( j \).

Time is discretized as \((nN + l)\delta\) as shown in figure 2(b), where \( n \) belongs to set of non-negative integers, \( l \) takes integer values in \([0, N - 1]\), \( \delta \) is the smaller discretized time slot when the UEs transmit packets locally. It is assumed that \( \delta \) (for each \( l \in [0, N - 1] \)) is large enough to send one packet of size \( L \) bytes. Here, \( N \) is the number of time slots (of \( \delta \) duration) between two consecutive global decisions by the BS. Global decision by BS is made at time when \( nN + l \) is divisible by \( N \). At this time instant BS takes the channel state information from all UEs in the service region and gives the decision of best relaying UE of a given zone for a given source UE. Hence, in between two consecutive time instants when BS can make global decision, a UE can send at-most \( N \) packets of size \( L \) to another UE. Note that at time \( l = 0 \), the UE chooses the relay link suggested by the BS and at time \( l \in \{1, 2, \ldots, N - 1\} \), UEs do not get channel state information from the BS. At global time instants, BS makes decision to UEs, i) the best relaying UE for a given source UE and ii) viable relaying zones \( \cup^i \) for given source zone \( i \), hence the zone \( i \) may choose an appropriate zone for relaying data from the set \( \cup^i \) by undergoing exploration.

Exploration: When the current link quality is not good enough then source UE locally explores for an appropriate relaying UE from given viable set. One of the basic procedure in the exploration phase involves sequentially searching space in all directions to align transmitter and receiver beams (beam alignment). The quality of link is piggy-back to the sending UE. This overall process of beam alignment is termed as the exploration phase. Any of the state of art approaches
Exploration zone

This exploration by source UE in zone on the effects of dynamic obstacle's blockage on a given D2D considered as a black box. It is assumed that the beams are W e have considered an abstraction of this phase and thus it is mentioned in [9] can be applied for the exploration phase. We have considered an abstraction of this phase and thus it is considered as a black box. It is assumed that the beams are perfectly aligned after the exploration phase and thus we focus on the effects of dynamic obstacle’s blockage on a given D2D link. This exploration by source UE in zone is done for UEs belonging only in the set $U^i$ to find out the best relaying zone for that time instant. Note that the UE is using directional mmWave antennas for exploring the neighbors and hence explorations cause some significant delay. Each exploration time duration is denoted by $\epsilon < \delta$ as shown in figure 2b). In a given duration $\delta$, we can perform a number of explorations. Each of the time instants during exploration time of duration $\epsilon$ is denoted by $m \in \{0, 1, \ldots, V - 1\}$, where $V = \lceil \delta / \epsilon \rceil$. The overall exploration time is bounded by the maximum value $M$ which can be more than $V$. It is assumed that once exploration is complete, switching takes negligible time.

There are static and dynamic obstacles in the service region. There is no external facility like radars or vision cameras available at BS to track them. The behavior of dynamic obstacles are not known a priori and need to be learned from the received ACKs of probe packets during exploration phase in an online fashion. Since mmWaves are highly susceptible to obstacles and suffer from severe penetration losses, we assume that even a single moving or static obstacle may break an already established D2D link and can cause packet loss.

III. Problem Formulation as POMDP

Zone $i$ containing source UE forms a link with an UE (relay or destination) of some other zone $j \in U^i$. Global decision for the best relay is given by the BS at the time instant $n \cdot N \delta$ to relay data packet till $(n + 1) \cdot N \delta$ time instant. The quality of link given by the BS may deteriorate due to presence of dynamic obstacles resulting in link outage and a new relay link needs to be explored from the viable set $U^i$ locally. A newly found relay link might provide very good data rate initially during exploration phase, but it may gradually deteriorate in future data transmission time because of dynamic obstacles. This problem is subtle due to the fact that exploration time is very small (ranging from few microseconds to 10 ms [7]) in contrast to the data transmission time which is in order of seconds (usually 100-1000 times more than search time). The main idea is to reduce the E2E delay on the cost of sending additional probe packets till decision of whether selecting or not selecting a given relay for data transmission considering dynamic obstacles can be learned. This process is limited up to maximum $M$ exploration time units where each exploration time duration is of $\epsilon$ time. If the given relay is not selected within this time then the exploration process again starts for another possible relaying UE in a zone belonging to $U^i$.

We model this problem of relay selection during exploration as that of POMDP considering uncertainty in relay link quality. We will define the POMDP as follows: For all the possible links $j \in U^i$, the state is written as $y^j_m \in \{0, 1\}$ for exploration time instant $m$. Values 1 and 0 signify if relay link $j$ is in good or bad state respectively. Good and bad notion denotes that whether link will be formed successfully or not respectively. State represented by $G$ and $\bar{G}$ respectively denote the good and bad state as shown in figure 3. The action set is defined as A: do not select relay link, select relay link, decision cannot be made thus continue sending probe packets. The first action signifies that the current link is bad, hence stop exploring it and start a new exploration. The second action signifies that the current link being explored is of good quality, hence choose it for data transmission. The third action signifies that the decision of relay selection cannot be made and exploration should continue for another $\epsilon$ unit of time to learn the relay link quality. The action is denoted by $a^j_m$ for link $j$ at time instant $m$. The link quality is observed by the ACKs of the received probe packet signal. The ACK test for link $j$ at time $m$ is denoted as $w^j_m \in \{0, 1\}$ for a given state $y^j_m$, and action $a^j_m$, may also be uncertain due to the unpredictable channel condition. In figure 3 A and $\vec{A}$ denotes whether ACKs are received successfully or not respectively.

Figure 3 represents the probabilistic structure of the problem as shown below.

\[
P(w^j_m = 1 | y^j_m = 1) = k; P(w^j_m = 0 | y^j_m = 1) = 1 - k
\]

\[
P(w^j_m = 1 | y^j_m = 0) = g; P(w^j_m = 0 | y^j_m = 0) = 1 - g
\]

However, note that the ACK piggyback in current time instant will show quality of the channel for probe packet sent in previous time instant. Hence a transition probability of $g$ is introduced above. The probabilistic structure assumed for the system state transition is given below. We have assumed that $q > s$ and $k > g$ which signify respectively that the probability of a link becoming good from previous good state is higher than that of previous bad state and probability of successful ACK from good state is higher than that of bad state.

\[
P(y^j_{m+1} = 1 | y^j_m = 1) = q; P(y^j_{m+1} = 0 | y^j_m = 1) = 1 - q
\]

\[
P(y^j_{m+1} = 1 | y^j_m = 0) = s; P(y^j_{m+1} = 0 | y^j_m = 0) = 1 - s
\]

For a given relaying zone $j$, let us define the information vector available locally to the zone $i$ till time instant $m$ as $H^j_m = (w^j_0, w^j_1, \ldots, w^j_m)$. The sufficient statistics or belief POMD (chapter 5) for this problem is defined locally for relaying zone $j$ as:

\[
r^j_m = P(y^j_m = 1 | H^j_m)
\]
This equation gives the probability that a relaying link/zone is in good state given the previous history of information. The estimator function for the local system can be defined as:
\[ r^j_{m+1} = \Phi'(r^j_m, w^j_{m+1}). \]  
(3)

Using Baye’s rule we get,
\[ r^j_{m+1} = \begin{cases} \frac{(qr^j_m + (1-r^j_m)s)k}{H}r^j_m + (1-q)(1-s)g & \text{if } w^j_{m+1} = 1 \\ \frac{(qr^j_m + (1-r^j_m)s)(1-k)}{H}r^j_m + (1-q)(1-s)(1-g) & \text{if } w^j_{m+1} = 0 \end{cases} \]  
(4)

where,
\[ H = (qr^j_m + (1-r^j_m)s)k + (r^j_m(1-q) + (1-r^j_m)(1-s))g \]
\[ J = (qr^j_m + (1-r^j_m)s)(1-k) + (r^j_m(1-q) + (1-r^j_m)(1-s))(1-g) \]

**Cost structure:** The cost for stopping the exploration for both cases when link cannot be and can be formed is 0. However, if the link which seems to be good and selected, may become bad in upcoming time instants and thus causes packet loss. So it will incur \( D_1 \) cost to compensate for the packet loss. Similarly, the link which was bad and not selected could have been a good relay link in upcoming time instants. Then it will incur some cost \( D_2 \) in order to compensate for further exploration. If decision of relay selection cannot be made in current time instant, then probe packets are sent for another \( \epsilon \) time unit to continue the exploration. The cost incurred here is \( c_e \). This will go on till time \( M \) which is the upper bound on the exploration time for a single link. The objective is to derive a decision criterion, whether to select or not select the given relay link, or if this decision cannot be made then continue learning the current relay link quality. The expected cost is formulated as a dynamic program. At the end of the \((M-1)\)th period, the expected cost is:
\[ K^j_{M-1}(r) = \min\{rD_1, (1-r)D_2\} \]  
(5)

where \( r \) is a variable denoting boolean value. For the time instant \( m = M - 2 \), we have,
\[ K^j_{M-2}(r) = \min\{rD_1, (1-r)D_2, c_e + E[K^j_{M-1}(r)]\} \]  
(6)

where \( E[\cdot] \) is the mathematical expectation over the observations. Here the first term in minimization expression denotes the expected cost \( (rD_1 + (1-r)0) \) incurred when the learning of the current link being explored is stopped since it was in the bad state. The second term denotes the expected cost for selecting the current relay link being explored. The third cost is the expected cost for continuing exploring the current relay link for another round of \( \epsilon \) unit of time. We can write a general expression as:
\[ K^j_m(r) = \min\{rD_1, (1-r)D_2, c_e + E[K^j_{m+1}(\Phi'(r, w))]\} \]  
(7)

where \( w \in \{0, 1\} \) is a variable denoting ACKs failure/success. Note that the notation \( w^j_m \) defined earlier denotes the ACK failure/success of link \( j \) at time instant \( m \), whereas \( w \) is used to denote a general variable for ACK failure/success. We will derive the optimal policy for this problem in section [IV].

### IV. Derivation of the Optimal Policy in Exploration Phase

#### A. Properties of \( K^j_m(r) \)

The general expression for time instant \( m \) as mentioned in equation (7), can be written equivalently as:
\[ K^j_m(r) = \min\{rD_1, (1-r)D_2, E_m^j(r)\} \]  
(8)

where,
\[ E_m^j(r) = c_e + P(w = 1|r)K^j_{m+1}(\Phi'(r, 1)) \]
\[ + P(w = 0|r)K^j_{m+1}(\Phi'(r, 0)) \]  
(9)

For notation simplicity we will now remove the superscript \( j \) from each of the respective notations, e.g., we will write \( E_m^j(\cdot) \) as \( E_m(\cdot) \). Hence \( E_m^j(r) \) can now be denoted as \( E_m(r) \). Now equation (9) can be further reduced as:
\[ E_m(r) = c_e + ((qr + (1-r)s)k + (r(1-q) + (1-r)(1-s))g)K_{m+1}(\Phi'(r, 1)) + (1 - ((qr + (1-r)s)k + (r(1-q) + (1-r)(1-s))g))K_{m+1}(\Phi'(r, 0)) \]  
(10)

The expected cost for the base case at the end of \((M-1)\)th period is mentioned in equation (5). We now show some of the properties of \( E_m(r) \).

**Proposition 1:** \( E_m(r) \) is piece-wise linear and concave in \( r \).

**Proof:** Proof given in Appendix [A]

**Proposition 2:** \( \forall r \in [0, 1], E_{m-1}(r) \leq E_m(r) \leq E_{m+1}(r) \)

**Proof:** Proof given in Appendix [B]

### B. Policy Structure

The structure of an optimal policy for our POMDP problem is provided in the following theorem.

**Theorem 1:** The optimal policy for exploration in POMDP problem is a threshold policy. At any time instant \( m \in \{0, 1, \cdots, M-1\} \), the optimal action is to stop exploration on current relay link and start exploring other relay links and choose it for data transmission if \( r_m \leq \alpha_m \), or stop exploration on current relay link and choose it for data transmission if \( r_m \geq \beta_m \), otherwise continue sending probe packets once more on the current relay link which is being explored to check the link quality if \( \alpha_m < r_m < \beta_m \). Here \( \alpha_m, \beta_m \) are appropriate constants. Also, we can say for the thresholds: \( \alpha_0 \geq \cdots \geq \alpha_{M-1} \geq \cdots \geq \alpha_1 \) and similarly \( \beta_0 \leq \cdots \leq \beta_{M-1} \leq \cdots \leq \beta_1 \).

**Proof:** Proof given in Appendix [C]

Using above result, we can say as \( m \to \infty \), \( \alpha_m \) and \( \beta_m \) converges to some scalar \( \alpha \) and \( \beta \) respectively. This is because
\( \beta_m \) is bounded below and \( \alpha_m \) is bounded above. Hence, for very large horizon length \( M \), the optimal policy can be approximated by a stationary threshold policy with a time-invariant threshold \( \pi \) and \( \beta \). In this case we can say that at time instant \( m \), if \( r_m \leq \pi \) then stop exploration on current link and start exploring other links from \( U^* \), if \( r_m \geq \beta \) then stop exploration on current link and choose it for data transmission, otherwise if \( \pi < r_m < \beta \) then continue sending probe packet once more on the current link which is being explored to check the link quality.

It is easy to check that \( \Phi(r, w) \) is a non-decreasing function in \( r \). Let us denote \( r_0 \) as the prior belief. We can say that, when prior belief satisfies \( r_0 > \Phi(r_0, w) = r_1 \), then \( r_2 = \Phi(r_1, 0) < \Phi(r_0, 0) = r_1 \) and \( r_2 = \Phi(r_1, 1) < \Phi(r_0, 1) = r_1 \). Proceeding in this way, we can show that \( r_m \) strictly decreases with \( m \) whenever we observe several successive ACK failures/successes. Similarly, when the prior belief satisfies \( r_0 < \Phi(r_0, w) = r_1 \), then we can say that \( r_2 = \Phi(r_1, 0) > \Phi(r_0, 0) = r_1 \) and \( r_2 = \Phi(r_1, 1) > \Phi(r_0, 1) = r_1 \). Proceeding in this way, we can show that \( r_m \) strictly increases with \( m \) whenever we observe several successive ACK failures/successes. For getting \( x \) successive ACK failures, we can define recursively a probability \( \pi_x \) as: \( \pi_1 = \Phi(r_0, A), \pi_2 = \Phi(\pi_1, A), \ldots, \pi_x = \Phi(\pi_{x-1}, A) \) with \( \pi_0 = \pi_0 \). Similarly, for getting \( x \) successive ACK successes, we can define recursively a probability \( \pi'_x \) as: \( \pi'_1 = \Phi(r_0, A), \pi'_2 = \Phi(\pi'_1, A), \ldots, \pi'_x = \Phi(\pi'_{x-1}, A) \) with \( \pi'_0 = \pi_0 \). Let \( c \) and \( d \) be the smallest integer such that \( \pi_c \leq \pi \) and \( \pi'_d \geq \beta \) respectively. We can further simplify the stationary threshold policy as follows.

**Corollary 1.** Using theorem 2 we can simplify the optimal policy further as follows. Let \( c \) and \( d \) be the smallest integer such that \( \pi_c \leq \pi \) and \( \pi'_d \geq \beta \) respectively. If there are \( c \) successive ACK failures, stop exploring on current relay link being explored and start exploring some other relay link. If there are \( d \) successive ACK successes, stop exploring on current relay link being explored and choose it for data transmission. Otherwise, decision cannot be made and more probe packets are needed to be sent to learn the link quality.

**V. Experiments and Results**

Service region of dimension \( 100 \text{ m} \times 100 \text{ m} \) is divided into grid zones each of dimension \( 10 \text{ m} \times 10 \text{ m} \). We have taken \( \delta = 100 \text{ ms} \) and \( \epsilon = 1 \text{ ms} \). UEs are using directional transmitter and receiver antennas for \( 60 \text{ GHz} \) frequency with \( G_r = G_t = 6 \text{ dBi} \). LOS path loss exponent is 2.5 and zero mean log-normal shadowing random variable with standard deviation 3.5. Thermal noise density is \( -174 \text{ dBm/Hz} \) and devices are using 24 dBm transmit power. Capacity of each link \( (i, j) \) is \( B \log_2(1+S_{ij}) \text{ bits/sec} \), where \( B = 20 \text{ MHz} \) and \( S_{ij} \) is the received signal to noise ratio. The packet length is fixed and of size 65535 bytes. A source UE in a grid zone \( i \) can transmit data to a maximum of 24 grids surrounding it (\( |U^*| \leq 24 \)). There are maximum of 16 static and \( D \) dynamic obstacles present in these 24 grids surrounding the source UE, where \( D \in \{0, 4, 8, 12, 16 \} \). Each static obstacle is placed uniformly in the service region. Each dynamic obstacle is moving randomly and independently of each other and following a simple blockage model such that with probability 0.5 it will block a given link otherwise it will not block the link. A single source-destination pair is assumed for simplicity and all other devices in a given zone may act as relay. We have written our own C++ custom code and run them on a GNU 4.8 compiler on Intel core i7 machine and average of 1000 runs are taken. The average exploration time and E2E delay for sending 100 packets are the main parameters considered. We have compared our results with the received signal strength (RSS Based) approach which is most commonly used for relay selection during exploration after transmitter and receiver beams are perfectly aligned.

Value of \( M \) is derived experimentally in figure 3. This value should not be too small, otherwise there won’t be sufficient exploration to lead to a decision. Also \( M \) should not be too big, otherwise a lot of time would be wasted in exploration. Figure 4 shows the percentage of time the decision could not be made on varying values of \( M \) for different \( D \). It is clear that for almost all values of \( D \), decision cannot be made when \( M \) is 1 and 2. For \( D = 16 \) and \( M = 3 \), for around 3.5% of cases decision cannot be made a shown in figure 5. This figure shows that with dynamic obstacles, no. of times the decision of relay selection cannot be made, also increases. We have chosen \( M = 4 \) for performing further simulations.

Figure 6 shows the effect of dynamic obstacles on percentage of times new relays are explored and switched. With dynamic obstacles, the probability of link breakage increases and thus number of times new relays are explored and switched also increases. Our approach outperforms the RSS based approach because it chooses the link by learning dynamic obstacle’s presence while RSS based approach greedily chooses a link based on best RSS values which might have higher chance of blockage. Figure 7 shows following results: first with dynamic obstacles, average E2E delay is increased which follows from previous result of figure 6. Since higher number of dynamic obstacles will have higher chance of breaking a link and thus causes increase in E2E delay. Similarly, following the argument of previous result of figure 6 our approach outperforms RSS based approach. Secondly, with dynamic obstacles, average exploration time is increased because higher link breakage with dynamic obstacles would cause more exploration and switching of relays. Here RSS based approach performs better since it always takes one time unit of exploration, whereas our approach takes additional exploration time units to learn the quality of link considering dynamic obstacles. Thirdly, it can be seen from above two results that, average exploration time is higher in our approach at the cost of reducing the cost of average E2E delay significantly compared to the RSS based approach. This describes the trade-off of additional exploration time over the average E2E delay.

**VI. Conclusion**

The problem of selecting a given relay during exploration is investigated taking into account presence of dynamic obstacles. We have modeled this problem as a finite horizon
POMDP framework at each UE. Using this model, an optimal threshold policy is derived for each UE which is then simplified to a stationary policy. This policy governs the source UE to take decisions based on successive ACK failures or success on the current relay link under exploration. Through simulations, the trade-off between average exploration time and E2E delay is shown, where our approach captures the effects of dynamic obstacles significantly compared to the RSS based approach.

REFERENCES

[1] D. Singh and S. C. Ghosh, “Mobility-aware relay selection in 5G d2d communication using stochastic model,” IEEE Transactions on Vehicular Technology, vol. 68, no. 3, pp. 2837–2849, 2019.
[2] J. Qiao, X. S. Shen, J. W. Mark, Q. Shen, Y. He, and L. Lei, “Enabling device-to-device communications in millimeter-wave 5G cellular networks,” IEEE Communications Magazine, vol. 53, pp. 209–215, January 2015.
[3] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, “A tutorial on beam management for mmwave cellular networks with proactive performance prediction using camera images and deep reinforcement learning,” IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 2, pp. 802–816, 2020.
[4] M. Soleimani, R. Parada, S. Tomasin, and M. Zorzi, “Fast initial access for mmwave 5G systems with hybrid beamforming using online statistics learning,” IEEE Communications Magazine, vol. 57, no. 9, pp. 132–137, 2019.
[5] M. Johnston and E. Modiano, “Wireless scheduling with delayed csi: When distributed outperforms centralized,” IEEE Transactions on Mobile Computing, vol. 17, no. 11, pp. 2703–2715, 2018.
[6] T. Bai and R. W. Heath, “Coverage and rate analysis for millimeter-wave cellular networks,” IEEE Transactions on Wireless Communications, vol. 14, pp. 1100–1114, Feb 2015.
[7] S. Biswas, S. Vuppala, J. Xue, and T. Ratnarajah, “An analysis on relay assisted millimeter wave networks,” in 2016 IEEE International Conference on Communications (ICC), pp. 1–6, May 2016.
[8] N. Wei, X. Lin, and Z. Zhang, “Optimal relay probing in millimeter-wave cellular systems with device-to-device relaying,” IEEE Transactions on Vehicular Technology, vol. 65, pp. 10218–10222, Dec 2016.
APPENDIX A
PROOF OF PROPOSITION[1]
We will prove this by first showing that $K_m(r)$ is piece-wise linear and concave for each $m$ using induction. Then we prove our proposition. For time instant $(M-1)$, we have,
\[ K_{M-1}(r) = \min \{ rD_1, (1-r)D_2 \} \]
which is piece-wise linear and concave. For time instant $(M-2)$, we have,
\[ K_{M-2}(r) = \min \{ rD_1, (1-r)D_2, E_{M-2}(r) \} \]
where,
\[ E_{M-2}(r) = c_e + \min \left \{ \left( qr + s(1-r) \right) kD_1, \left( r(1-q) + (1-s)(1-r) \right) gD_2 \right \} + \min \left \{ \left( qr + s(1-r) \right)(1-k)D_1, \left( r(1-q) + (1-s)(1-r) \right)(1-g)D_2 \right \} \]
This overall equation is also piece-wise linear and concave.

Assuming $K_{m+1}(r)$ is piece-wise linear and concave in $r$, we can say that for some suitable scalars, $\eta_1, \eta_2, \cdots, \eta_n$ and $\beta_1, \beta_2, \cdots, \beta_n$, $K_{m+1}(r)$ can be written as:
\[ K_{m+1}(r) = \min \{ \eta_1 + \beta_1 r, \eta_2 + \beta_2 r, \cdots, \eta_n + \beta_n r \} \quad (11) \]
We can write,
\[ K_m(r) = \min \{ rD_1, (1-r)D_2, E_m(r) \} \]
Expanding above using equation (10), we get:
\[ K_m(r) = \min \left \{ rD_1, (1-r)D_2, c_e + \left( (rq + (1-r)s)k + (r(1-q) + (1-r)(1-s))g \right) x \right \} + \min \left \{ \left( rq + (1-r)s \right) k \left( 1 - \left( (r(1-q) + (1-r)(1-s))g \right) \right) \right \} \]
Let us substitute $Y$ for $\left( (rq + (1-r)s)k + (r(1-q) + (1-r)(1-s))g \right)$ to simplify the calculations. Now using equation (11), the equation (12) is reduced as:
\[ K_m(r) = \min \left \{ rD_1, (1-r)D_2, c_e + \left( \eta_1 + \beta_1 \frac{(qr + s(1-r))k}{Y}, \eta_2 + \beta_2 \frac{(qr + s(1-r))k}{Y}, \cdots, \eta_n + \beta_n \frac{(qr + s(1-r))k}{Y} \right \} + \left( 1 - Y \right) \min \left \{ \eta_1 + \beta_1 \frac{(qr + s(1-r))(1-k)}{1-Y}, \eta_2 + \beta_2 \frac{(qr + s(1-r))(1-k)}{1-Y}, \cdots, \eta_n + \beta_n \frac{(qr + s(1-r))(1-k)}{1-Y} \right \} \quad (13) \]
We can further reduce above equation as:
\[ K_m(r) = \min \left \{ rD_1, (1-r)D_2, c_e + \min \{ \left( Y \eta_1 + \beta_1 (qr + s(1-r))k \right), \cdots, \left( Y \eta_n + \beta_n (qr + s(1-r))k \right) \} + \min \{ \left( 1 - Y \right) \eta_1 + \beta_1 (qr + s(1-r))(1-k), \cdots, \left( 1 - Y \right) \eta_n + \beta_n (qr + s(1-r))(1-k) \} \} \quad (14) \]
This is again piece-wise linear and concave in $r$. Thus the induction is complete.

Now we will show that $E_m(r)$ is also piece-wise linear and concave in $r$:
\[ E_m(r) = c_e + \left( Y \eta_1 + \beta_1 \frac{(qr + s(1-r))k}{Y}, \eta_2 + \beta_2 \frac{(qr + s(1-r))k}{Y}, \cdots, \eta_n + \beta_n \frac{(qr + s(1-r))k}{Y} \right \} + \left( 1 - Y \right) \min \left \{ \eta_1 + \beta_1 \frac{(qr + s(1-r))(1-k)}{1-Y}, \eta_2 + \beta_2 \frac{(qr + s(1-r))(1-k)}{1-Y}, \cdots, \eta_n + \beta_n \frac{(qr + s(1-r))(1-k)}{1-Y} \right \} \quad (15) \]
The first term $c_e$ is constant. For the next two terms $(Y)K_{m+1}(\Phi'(r, 1))$ and $(1-Y)K_{m+1}(\Phi'(r, 0))$, by expanding them using equation (11), we get,
\[ (Y)K_{m+1}(\Phi'(r, 1)) + (1-Y)K_{m+1}(\Phi'(r, 0)) = \min \left \{ \eta_1 + \beta_1 \frac{(qr + s(1-r))k}{Y}, \eta_2 + \beta_2 \frac{(qr + s(1-r))k}{Y}, \cdots, \eta_n + \beta_n \frac{(qr + s(1-r))k}{Y} \right \} + (1-Y) \min \left \{ \eta_1 + \beta_1 \frac{(qr + s(1-r))(1-k)}{1-Y}, \eta_2 + \beta_2 \frac{(qr + s(1-r))(1-k)}{1-Y}, \cdots, \eta_n + \beta_n \frac{(qr + s(1-r))(1-k)}{1-Y} \right \} \quad (16) \]
We can reduce above to:
\[ (Y)K_{m+1}(\Phi'(r, 1)) + (1-Y)K_{m+1}(\Phi'(r, 0)) = \min \left \{ \eta_1 + \beta_1 \frac{(qr + s(1-r))k}{Y}, \eta_2 + \beta_2 \frac{(qr + s(1-r))k}{Y}, \cdots, \eta_n + \beta_n \frac{(qr + s(1-r))k}{Y} \right \} + \min \left \{ \eta_1 + \beta_1 \frac{(qr + s(1-r))(1-k)}{1-Y}, \eta_2 + \beta_2 \frac{(qr + s(1-r))(1-k)}{1-Y}, \cdots, \eta_n + \beta_n \frac{(qr + s(1-r))(1-k)}{1-Y} \right \} \quad (17) \]
Since minimum of finite number of concave function is concave, $E_m(r)$ is piece-wise linear and concave in $r$.

APPENDIX B
PROOF OF PROPOSITION[2]
We will first prove $K_m(r) \geq K_{m+1}(r)$, then we will use this to prove $E_m(r) \geq E_{m+1}(r)$. First we start for base case $m = M-1$ and the first term in recursion $m = M-2$:
\[ K_{M-1}(r) = \min \left \{ rD_1, (1-r)D_2 \right \} \]
and
\[ K_{M-2}(r) = \min \left \{ rD_1, (1-r)D_2, E_{M-2}(r) \right \} \]
respectively. We can easily see that $K_{M-1}(r) \geq K_{M-2}(r)$. We now prove it for first two terms of the recursion $K_{M-2}(r)$
and $K_{M-3}(r)$. Let’s denote $Y = (rq + (1 - r)s)k + (r(1 - q) + (1 - r)(1 - s))g$, we can write $K_{M-3}(r)$ as:

$$K_{M-3}(r) = \min\{r(D_1, (1 - r)D_2, c_r + (Y)K_{M-2}(\Phi'(r, 1)) + (1 - Y)K_{M-2}(\Phi'(r, 0)))\}$$

$$\leq \min\{r(D_1, (1 - r)D_2, c_r + (Y)K_{M-1}(\Phi'(r, 1)) + (1 - Y)K_{M-1}(\Phi'(r, 0)))\}$$

$$= K_{M-2}(r)$$

Hence $K_{M-2}(r) \geq K_{M-3}(r)$. Similarly it proceeds for other $m$ and hence $K_{m+1}(r) \geq K_m(r)$. Now let us see this for $E_m(r)$ using previous proof for $K_m(r)$:

$$E_m(r) = e_r + (Y)K_{m+1}(\Phi'(r, 0)) + (1 - Y)K_{m+1}(\Phi'(r, 0))$$

$$\leq e_r + (Y)K_{m+2}(\Phi'(r, 0)) + (1 - Y)K_{m+2}(\Phi'(r, 0))$$

$$= E_{m+1}(r)$$

Hence $E_{m+1}(r) \geq E_m(r)$.

**APPENDIX C**

**PROOF OF THEOREM**

For the last time period $M-1$, there exists $\rho = \frac{D_2}{D_1 + D_2}$ such that stop exploration on current link and start exploring other links from $U^i$ if $r < \rho$, otherwise stop exploration on current link and choose it for data transmission if $r \geq \rho$. At $\rho$, $K_{M-1}(\rho)$ attains its maximum value of $\frac{D_2}{D_1 + D_2}$.

For general time instants $m$, we can say that $E_m(0) > e_r$ and $E_m(1) > e_r$. Using this fact, proposition [1] and proposition [2] we can say that if for some $r'$, $E_{M-2}(r') < \frac{D_2D_1}{(r' + 1)D_2}$, then $E_{M-2}(r)$ and subsequent $E_m(r)$ will intersect $\min\{rD_1, (1 - r)D_2\}$ at two points. Hence we can say that for general time instant $m$ we will get an optimal policy as: stop exploration on current link and start exploring other links from $U^i$ if $r_m \leq \alpha_m$, or stop exploration on current link and choose it for data transmission if $r_m \geq \beta_m$, otherwise continue sending probe packet once more on the current link which is being explored to check the link quality if $\alpha_m < r_m < \beta_m$. Here $\alpha_m$ and $\beta_m$ are found by satisfying the corresponding equations: $\alpha_m D_1 = E_m(\alpha_m)$ and $(1 - \beta_m)D_2 = E_m(\beta_m)$.

If for all $r'$, $E_{M-2}(r') > \frac{D_2D_1}{D_1 + D_2}$, then terms $rD_1$ and $(1 - r)D_2$ will contribute to the single threshold $\rho$ which is trivial for time instant $M-2$ due to proposition [1]. In this case, using proposition [1] and proposition [2] one can have either single or two thresholds for other time instants $m$ depending upon whether $E_m(r)$ intersects $\min\{rD_1, (1 - r)D_2\}$ at two or no points as shown in figure below.

For the second part, using proposition [2] and if $E_{M-2}(\rho) < \frac{D_2D_1}{D_1 + D_2}$, we can say that $\alpha_{M-2} < \rho < \beta_{M-2}$. This is easy to see because $K_{M-1}(r)$ first increases till $\rho$ and then decreases. Also,

$$K_{M-1}(r) \geq K_{M-2}(r) = \min\{K_{M-1}(r), E_{M-2}(r)\}$$

and $\alpha_{M-1} = \beta_{M-1} = \rho$. Hence, $\alpha_{M-2} \leq \alpha_{M-1}$ and $\beta_{M-1} \leq \beta_{M-2}$. Similarly, using proposition [2] we can say this for other instants $m$ that $\alpha_m \leq \alpha_{m+1}$ and $\beta_m \geq \beta_{m+1}$.

Hence we can say that with respect to $m$:

$$\cdots \geq \alpha_m \geq \alpha_{m+1} \geq \cdots \geq \alpha_1$$

and similarly

$$\cdots \leq \beta_m \leq \beta_{m+1} \leq \cdots \leq \beta_1.$$