On Multicast Capacity and Delay in Cognitive Radio Mobile Ad-hoc Networks

Yixuan Li, Xinbing Wang
Shanghai Jiao Tong University, China
Email: {lyx1990116, xwang8}@sjtu.edu.cn

Chee Wei Tan
City University of Hong Kong, Hong Kong
Email: cheewtan@cityu.edu.hk

Abstract—In this paper, we focus on the capacity and delay tradeoff for multicast traffic pattern in Cognitive Radio (CR) Mobile Ad-hoc Networks (MANET). In our system model, the primary network consisting of \( n \) primary nodes, overlaps with the secondary network consisting of \( m \) secondary nodes in a unit square. Assume that all nodes move according to an i.i.d. mobility model and each primary node serves as a source that multicasts its packets to \( k_p \) primary destination nodes whereas each secondary source multicasts its packets to \( k_s \) secondary destination nodes. Under the cell partitioned network model, we study the capacity and delay for the primary networks under two communication schemes: Non-cooperative Scheme and Cooperative Scheme. The communication pattern considered for the secondary network is Cooperative Scheme. Given that \( m = n^\beta \) (\( \beta > 1 \)), we show that per-node capacity \( O(1/k_p) \) and \( O(1/k_s) \) are achievable for the primary network and the secondary network, respectively, with average delay \( \Theta(n \log k_p) \) and \( \Theta(m \log k_s) \). Moreover, to reduce the average delay in the secondary network, we introduce the Redundancy Scheme and prove that a per-node capacity \( O(1/k_s \sqrt{m \log k_s}) \) is achievable with average delay \( \Theta(\sqrt{m \log k_s}) \). We find that the fundamental delay-capacity tradeoff in the secondary network is delay/capacity \( \geq O(mk_s \log k_s) \) under both the cooperative scheme and redundancy scheme.

I. INTRODUCTION

Since the seminal work by Gupta and Kumar [1], the study of the capacity of wireless networks has received great attention. Gupta and Kumar showed that the per-node capacity is \( \Theta(1/\sqrt{n \log n}) \) for a unicast traffic pattern, where \( n \) is the number of nodes in the network. Compared with unicast traffic pattern where packets are sent from one source node to one destination node, multicast traffic pattern has several benefits. Using this pattern, packets from source nodes are delivered to multiple destination nodes and certain links may be shared by different destinations. Therefore, the multicast traffic pattern is supposed to reduce the frequency resources that are required to establish communication with all destination nodes.

Compared to the capacity scaling in the static network model, mobility has been leveraged to improve the capacity bounds. In [2], M. Grossglauser et al. demonstrated that per-node capacity \( \Theta(1) \) is achievable under the mobility model. However, they neglected another significant issue of wireless networks besides throughput (capacity): the communication delay. This motivated the later focus on the relationship between delay and throughput. In [3], M. Neely et al. studied the capacity-delay tradeoff in cell partitioned MANET by introducing the redundancy scheme which reduced delay by sacrificing capacity. They developed communication schemes to achieve per-node capacity \( \Theta(1), \Theta(\frac{1}{\sqrt{n}}) \) and \( \Theta(\frac{1}{n \log n}) \) with average delay \( \Theta(n) \), \( \Theta(\sqrt{n}) \) and \( \Theta(\log n) \), respectively, which implies that the capacity-delay tradeoff is \( \frac{\text{delay}}{\text{capacity}} > O(n) \). Later on, the capacity-delay tradeoff for unicast networks was thoroughly studied using various mobility models such as the random walk model [4], the random way point mobility model [5] and the constrained mobility model [6], [7].

For a multicast traffic pattern, X. Li et al. [8] studied the capacity in static networks where each node sends messages to \( k - 1 \) destinations. Based on their interference model, they demonstrated that the per-node capacity is \( \Theta(\frac{1}{k \sqrt{n \log n}}) \) if \( k = O(\frac{n}{\log n}) \) whereas the per-node capacity \( \Theta(\frac{n}{\log n}) \) is achievable when \( k = \Omega(\frac{n}{\log n}) \). This result is actually the generalization of the other two traffic patterns: unicast [1] and broadcast [9]. Some relevant works concerning the capacity-delay tradeoff may also be found in [10], [11] and [12]. In [13], X. Wang studied the capacity and delay tradeoff in cell partitioned MANET under a multicast traffic pattern. They proved that the per-node capacity and delay are \( O(1/k) \) and \( \Theta(n \log k) \) respectively if no redundant relay nodes are introduced, and \( O(1/k \sqrt{n \log k}) \) and \( \Theta(\sqrt{n \log k}) \) respectively if there are redundant relay nodes.

All the aforementioned results concentrate on the throughput scaling and delay analysis for a single network. Recently, the ever-growing demand for frequency resources has motivated study of cognitive radio (CR) networks to efficiently utilize the idle spectrum in the time and space domain. CR networks can be regarded as a superposition of two independent networks: primary network and secondary network. In cognitive networks, primary users have a higher priority when accessing the spectrum, while secondary users could opportunistically access the licensed spectrum without harming the performance of primary users. Specifically, the secondary network has “vacuum” space in which no transmissions can take place. In particular, secondary nodes within the interference range of any primary nodes that are broadcasting or receiving messages can not have spectrum opportunities. Unlike a single network,

\footnote{Given two functions \( f(n) > 0, g(n) > 0 \); \( f(n) = o(g(n)) \) means \( \lim_{n \to \infty} \frac{f(n)}{g(n)} = 0 \); \( f(n) = O(g(n)) \) means \( \lim_{n \to \infty} \sup \{f(n)/g(n)\} < \infty \); \( f(n) = \omega(g(n)) \) is equivalent to \( g(n) = o(f(n)) \); \( f(n) = \Omega(g(n)) \) is equivalent to \( g(n) = O(f(n)) \); \( f(n) = \Theta(g(n)) \) means \( f(n) = O(g(n)) \) and \( g(n) = O(f(n)) \).}
the potential interaction between primary and secondary users should be considered when scaling the throughput and delay in CR networks. Therefore, the previous results of single networks may not be effectively applied to CR networks.

The study of capacity and delay scaling laws for both the primary and secondary network is a relatively new and challenging field. In [14] and [15], the authors investigated the capacity scaling of a homogeneous cognitive networks for unicast traffic by introducing a preservation region around primary receivers. Wang et. al [16] extended the CR network capacity scaling to multicasting traffic patterns under the Gaussian channel model. Interestingly, all these works showed that both the primary and secondary networks can achieve similar or the same performance bounds as they are stand-alone networks when the secondary network has a higher density than the primary one. Note that existing works mainly focus on static CR networks while the performance of CR networks under a mobility model has not been investigated to the best of our knowledge.

Motivated by the fact that mobility can dramatically enhance the throughput in stand-alone networks, we are interested in finding out:

- What throughputs and delay scaling can be achieved under a multicast traffic pattern in Cognitive Radio Mobile Ad-hoc Networks (CR MANET)?
- What is the delay and capacity tradeoff in CR MANET?

In our approach, we first adopt a non-cooperative scheme to investigate the achievable per-node capacity and delay for the primary network where one source is supposed to send messages to \( k_p \) destinations. However, the per-node capacity reduces from \( O(1/k_p) \) to \( O(1/k_p^2) \) as \( k_p \) increases from \( o(n) \) to \( \Theta(n) \), which motivates us to study the cooperative scheme among primary destination nodes to improve the per-node capacity when \( k_p = \Theta(n) \). Next, we focus on the capacity and delay analysis for the secondary network under a cooperative scheme and a redundancy scheme, respectively. Finally, we introduce a destination oriented redundancy scheme to efficiently utilize the network resources for the secondary network. The major contributions of this paper are as follows:

- We propose and investigate the mobility model for both the primary network and the secondary network. Compared with the mobility model in [17] and [18], which requires the primary nodes to be static and only the secondary nodes are allowed to move, our model is more practical and general as both the primary and secondary networks can be mobile. A novelty is to use the transmission queues to analyze the capacity and delay for CR MANET. Our results show that mobility can significantly enhance the capacity performance for both the primary and secondary networks.

- This paper is a nontrivial generalization of previous work since we directly study the capacity and delay under a multicast traffic pattern. Our results can be specialized to the unicast and broadcast traffic by setting the number of destinations to be 1 and \( n \) for the primary network (1 and \( m \) for the secondary network), respectively.

- We show that both the primary network and the secondary network can achieve the same throughput as the optimal one established for a stand-alone MANET in [13], if the number of secondary users \( m \) is larger than that of primary users \( n \) scaled by \( m = n^\beta (\beta > 1) \). Specifically, under the cooperative scheme, per-node capacity \( O(1/k_p) \) and \( O(1/k_s) \) are achievable for the primary network and the secondary network, respectively, with average delay \( \Theta(n \log k_p) \) and \( \Theta(m \log k_s) \).

- We introduce redundancy into the secondary network and prove that the tight bound of transmission delay can be reduced to \( \Theta(\sqrt{m \log k_s}) \) with achievable per-node capacity \( O(1/k_s \sqrt{m \log k_s}) \) if there are redundant relay nodes. To achieve the bound, a redundancy scheme is proposed. Moreover, we also proposed destination oriented redundancy scheme to help save wireless resources in practical operation. We find that the fundamental capacity-delay tradeoff in the secondary network is delay/capacity \( \geq O(m k_s \log k_s) \) under both the cooperative scheme and the redundancy scheme.

The rest of the paper is organized as follows. In section II, we introduce our network model and give some necessary definitions. In section III, we present the communication schedule for both the primary and the secondary network. In sections IV and V, we put forward our main results of capacity and delay in the primary network and secondary network, respectively. Finally, we conclude in section VI.

II. Network Model

**Cell Partitioned Network Model:** Our network model is based on the model utilized in [3]. Consider a unit square with a co-existing primary network and secondary network, where the two networks share the same space, time and frequency domain but with different priorities in accessing the spectrum. The primary and secondary nodes are distributed according to a Homogeneous Poisson Process (HPP) of density \( n \) and \( m \), respectively. And we assume \( n \) and \( m \) satisfy \( m = n^\beta (\beta > 1) \), which indicates that the secondary network is more dense than the primary network. It can be proved that the total numbers of primary nodes and secondary nodes within the unit area are of order \( \Theta(n) \) and \( \Theta(m) \), respectively (See Lemma 9). For simplicity, we will assume that there are \( n \) mobile primary nodes and \( m \) mobile secondary nodes distributed over the square throughout the paper, which does not influence our main results. We further divide the primary network into \( w = \Theta(n) \) non-overlapping cells with equal area \( \Theta(1/k_p) \) and divide the secondary network into \( c = \Theta(m) \) non-overlapping cells with equal area \( \Theta(1/k_s) \) as illustrated in Figure 1. The node density \( \tau_n \) is \( \frac{2}{\pi} \) for the primary network and \( \tau_s \) is \( \frac{2}{\pi}m \) for the secondary network. Clearly, both \( \tau_n \) and \( \tau_s \) are of constant order \( \Theta(1) \).

**Traffic Pattern:** In a multicast traffic pattern, we assume that there are a set \( V_p = \{ v_1^p, v_2^p, \ldots, v_n^p \} \) mobile primary nodes and another set \( V_s = \{ v_1^s, v_2^s, \ldots, v_m^s \} \) mobile secondary nodes in the unit square. For each multicast group in the
In CR MANET, there are three kinds of interference: inter-cell interference, intra-cell interference and inter-network interference. To avoid intra-cell interference, we assume that each cell (primary or secondary) allows at most one transmission during a time slot. In order to mitigate inter-primary-cell interference, neighboring primary cells have to transmit over orthogonal frequency bands and four bands are enough for the whole primary network to ensure this condition[3]. The inter-secondary-cell can be avoided by adopting a 25-TDMA schedule for the secondary network. Since the primary nodes have higher priority in using the spectrum, the secondary network has to operate without causing interference to the primary network. To limit the inter-network interference, no spectrum opportunities will be available for a secondary node if it resides in the region of an active primary cell.  

A primary cell becomes active when a transmission between two primary nodes takes place in it; otherwise, it is called an inactive primary cell. Accordingly, we define a secondary cell as active only when it is located in the region of inactive primary cells.

### Definitions

**Notations**

<table>
<thead>
<tr>
<th>Notations</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_p$</td>
<td>Input rate of primary source node (per-node capacity for the primary network).</td>
</tr>
<tr>
<td>$\lambda_s$</td>
<td>Input rate of secondary source node (per-node capacity for the secondary network).</td>
</tr>
<tr>
<td>$\lambda_{ps}$</td>
<td>Total service rate of primary source queue.</td>
</tr>
<tr>
<td>$\lambda_{ps}'$</td>
<td>Total service rate of secondary source queue.</td>
</tr>
<tr>
<td>$\lambda_{sub}$</td>
<td>Input rate of a primary source-to-destination sub-queue.</td>
</tr>
<tr>
<td>$\lambda_{sub}'$</td>
<td>Output rate of a primary source-to-destination sub-queue.</td>
</tr>
<tr>
<td>$\mu_{ps}$</td>
<td>Output rate of a secondary source-to-relay queue.</td>
</tr>
<tr>
<td>$D_{s\rightarrow d}$</td>
<td>Delay of primary source-to-destination communications.</td>
</tr>
<tr>
<td>$D_{s\rightarrow r}$</td>
<td>Delay of primary source-to-relay communications.</td>
</tr>
<tr>
<td>$D_{s\rightarrow sub}$</td>
<td>Delay of secondary source-to-relay communications.</td>
</tr>
<tr>
<td>$D_{r\rightarrow d}$</td>
<td>Delay of primary relay-to-destination communications.</td>
</tr>
<tr>
<td>$D_{r\rightarrow sub}$</td>
<td>Delay of secondary relay-to-destination communications.</td>
</tr>
<tr>
<td>$D_p$</td>
<td>Total delay for the primary network.</td>
</tr>
<tr>
<td>$D_s$</td>
<td>Total delay for the secondary network.</td>
</tr>
</tbody>
</table>

**Useful Notations**

In this section, we will introduce the communication schemes for both the primary network and the secondary network.
network. Due to the priority of the primary network, it operates without being aware of the existence of the secondary network. The secondary network adaptively chooses its protocol according to the given primary transmission scheme.

A. Communication Scheme for the Primary Network

In the primary network, we assume that the packets are delivered using at most two hops. The source node either sends packets directly to all the destinations or to one of the relays. Then the relay will forward the packet to all the corresponding destinations. Each cell becomes active when at least one successful transmission can happen. In the following, we consider two schemes: non-cooperative scheme and cooperative scheme.

Non-cooperative Scheme:
For an active cell containing at least two primary nodes, the transmissions are conducted in the following way.

1) If at least one primary source-destination pair can be found in the cell, then randomly pick one pair to finish the communication.
2) If no source-destination pairs can be found in one cell, perform the following two schedules with equal probability:
   - Source-to-Relay Transmission: If one primary source node can be found in the cell and at least one normal relay node is available for the source node in the same cell, pick the source node as the sender and a relay node as receiver to finish the transmission.
   - Relay-to-Destination Transmission: If one relay node, carrying a packet destined for a primary destination node, can be found in one cell and at least one corresponding “pristine” primary destination node stays within the same cell, pick the relay node as the sender and one of these “pristine” destination nodes as receiver to finish the transmission.
3) If neither of the two items mentioned above are satisfied, no transmissions will take place in this cell. A packet will be discarded whenever all its $k_p$ primary destinations have received it.

Cooperative Scheme:
According to the previous scheme, one primary destination node can not serve as a relay for other destinations within the same multicast group. That is to say, a certain destination node either receives packets from corresponding source node or acts as a relay for other multicast groups. However, under the cooperative scheme, we will no longer discriminate between

3We use the term “non-cooperative” to describe the situation when a destination node can not serve as a relay for other destinations within the same multicast group. Otherwise, the destination node can act as a relay for any multicast group, which is called “cooperative”.
4For a certain multicast group, all nodes in the area except for those destination nodes within this multicast group can be treated as normal relay nodes for this multicast group.
5A “pristine” primary destination node represents a destination node which has not received the desired packet, that the relay node is carrying. This is the same with a “pristine” secondary destination node.

the destination nodes and normal relay nodes. The first node that the source encounters in its cell will be treated as a relay node irrespective of whether it is a destination node or not. Once a destination node is chosen as a relay node, it can send packets to other destination nodes (possibly within the same multicast group or not). Under the cooperative scheme for the primary network, the following two transmission patterns are conducted with equal probability:

1) Source-to-Relay Transmission: If one primary source node can be found in the cell and one relay node is available in the same cell, pick the source node as the sender and one relay node as the receiver to finish the transmission. Note that the receiver is treated as a relay node irrespective of whether it is a destination node or not.
2) Relay-to-Destination Transmission: If one relay node, carrying a packet destined for primary destination nodes, can be found in this cell and meanwhile at least one corresponding “pristine” primary destination node resides within the same cell, pick the relay node as the sender and one destination node as receiver to finish the transmission.

If neither of the two items mentioned above are satisfied, no transmissions will take place in this cell.

B. Communication Scheme for the Secondary Network

Unlike the scheme in the primary network, a node in the secondary network can only opportunistically transmit whenever it is outside the region of active primary cells. Since a secondary cell is smaller than a primary cell, the region of a primary cell can actually cover $\Theta\left(\frac{m}{n}\right)$ secondary cells. All these secondary nodes within a certain primary cell region will transmit using the primary cell’s frequency. Hence, interference is likely to happen among the secondary cells within the same primary cell. To mitigate the interference between secondary cells, we adopt a 25-TDMA scheme to schedule the secondary cells within the same primary cell region. We assume the time slot of a secondary frame is equal to the time slot of a primary frame. Each secondary frame is further divided into 25 sub-frames. During each time slot, a secondary cell becomes active in one sub-frame.

From the capacity analysis for the primary network in section IV, we find that the cooperative scheme can achieve better performance than the non-cooperative scheme. Therefore, we only consider the cooperative scheme for the secondary network without considering the non-cooperative scheme anymore. The cooperative scheme in the secondary network is similar to that in the primary network.

Cooperative Scheme:
Under the cooperative scheme for the secondary network, the following two transmission patterns are conducted with equal probability for an active secondary cell:

1) Source-to-Relay Transmission: If one secondary source node can be found in the cell and one relay node is available in the same cell, pick the source node as the
sender and one relay node as the receiver to finish the transmission. Note that the receiver is treated as a relay node irrespective of whether it is a destination node or not.

2) Relay-to-Destination Transmission: If one relay node, carrying a packet destined for secondary destination nodes, can be found in this cell and at the same time at least one corresponding “pristine” primary destination node resides within the same cell, pick the relay node as the sender and one destination node as the receiver to finish the transmission.

If neither of the two items mentioned above are satisfied, no transmissions will take place in this cell. A packet will be discarded when all its \( k_s \) destinations have received it.

Note that in the cooperative scheme, only one relay node is used for sending a single packet. For the secondary network, we further propose a redundancy scheme in order to reduce the delay, which allows more than one relay to be used for delivering a single packet. Then, we improve the redundancy scheme by avoiding introducing extra nodes other than the destination nodes which have received packets to serve as relays. This is referred to as destination oriented redundancy scheme. In this way, we are able to better utilize network resources under this scheme. This communication schedule will be discussed in more details later on in section V.

IV. CAPACITY AND DELAY ANALYSIS FOR THE PRIMARY NETWORK

In this section, we will study the capacity and delay tradeoff in the primary network. Denote the probability that there are at least two primary nodes in one cell by \( p_1 \). Denote the probability of finding a source-destination pair in the primary network by \( p_2 \). We first calculate \( p_1 \) by

\[
p_1 = 1 - (1 - \frac{1}{w})^{kn} - \binom{k}{1} \frac{1}{w}(1 - \frac{1}{w})^{kn-1}
= 1 - (1 - \frac{1}{w})^{kn} - \frac{n}{w}(1 - \frac{1}{w})^{kn-1}
\sim 1 - (1 + \tau_p)e^{-\tau_p}.
\]

To get \( p_2 \), we need to model the primary nodes as mutually exclusive groups according to the following definition.

**Definition 1:** \( \{k_p + 1\} \)-grouped Primary Network: For each source node \( v^s_p \in V_p \) in the primary network, we put it and its \( k_p \) destination nodes into a group, denoted by \( G^p_i \). Then we will have \( \frac{n}{k_p+1} \) groups over the whole primary network. We also assume that \( G^p_i \cap G^p_j = \emptyset \), for \( i \neq j \), and \( i, j \in [1, \frac{n}{k_p+1}] \).

Note that in the \( \{k_p + 1\} \)-grouped primary network, each node within \( G^p_i \) can be a source node or destination node. Thus, any two nodes from the same group are a source-destination pair. However, for any randomly chosen two nodes from different groups, they cannot be viewed as a source-destination pair. Then, we get

\[
p_2 = 1 - \left( (1 - \frac{1}{w})^{k_p + 1} + \binom{k_p + 1}{1} \frac{1}{w} \right)^n e^{-\tau_p}
= 1 - \left( 1 - \frac{1}{w} \right)^{n k_p + 1} \left( 1 + \frac{k_p}{n} \right) e^{-\tau_p}
\sim \begin{cases} 1 - e^{-\tau_p} e^{-\tau_p} = 0, & \text{if } k_p = o(n); \\
1 - (1 + \tau_p) e^{-\tau_p}, & \text{if } k_p = \Theta(n). \end{cases}
\]

Next we move on to a capacity and delay analysis for the non-cooperative scheme.

A. Capacity and Delay Analysis of the Non-cooperative Scheme

In this subsection, we will study the achievable capacity and communication delay in the primary network when destination nodes can not relay packets to other destination nodes within the same multicast group. Our main results are presented in the following.

**Theorem 1:** In a cell-partitioned network with overlapping \( n \) primary nodes and \( m \) secondary nodes, the achievable per-node capacity for the primary network under the non-cooperative scheme is \( \lambda_p = O(1/k_p) \) with average delay \( E[D_p] = \Theta(n \log k_p) \) when \( k_p = o(n) \), and \( \lambda_p = O(1/k^2_p) \) with average delay \( E[D_p] = \Theta(n \log k_p) \) when \( k_p = \Theta(n) \).

**Proof:** In each time slot, a new packet arrives at primary source node \( v^s_p \) in the group \( G^p_i \) with rate \( \lambda_p \). Denote the rate that the packet is transmitted to a primary destination node or handed over to a relay node by \( R_1 \) and \( R_2 \), respectively. Then we have

\( \lambda_p = R_1 + R_2. \)

Since the transmission of source-to-relay and relay-to-destination are of equal probability, \( R_2 \) is also equal to the rate at which the relay nodes are sending packets to primary destinations. Thus, in every time slot, the total rate of transmission opportunities over the primary network is \( n(R_1 + 2R_2) \). Meanwhile, a transmission occurs in any given cell with probability \( p_1 \). Hence we obtain

\[
w p_1 = n(R_1 + 2R_2).
\]

Similarly, since \( p_2 \) denotes the probability that a primary source-destination pair can communicate with each other in one cell, we find

\[
w p_2 = n R_1.
\]

Combining Equations (10) and (2), we have

\[
R_1 = \frac{p_2}{\tau_p}; R_2 = \frac{(p_1 - p_2)}{2\tau_p}.
\]

Then we obtain that the total service rate for one primary source queue is \( \lambda^o_p = \frac{p_2}{\tau_p} \).

Now we will deal with the traffic delay for the primary network using queueing theory techniques. There are two possible routing strategies for a primary packet to reach its destination: the 1-hop source-to-destination path and the 2-hop source-to-relay-to-destination path. For the first strategy,
it is clear that transmission delay consists of only the queueing delay at the source since we neglect the propagation delay. On the other hand, in the second strategy, delay is composed of the queueing delay at both the source and the relay nodes.

If one packet is directly sent from the source node to destination nodes, it will wait at the source for a time period \( D_{s \rightarrow d} \) before the source can find its corresponding destinations to distribute this packet. Since a source node transmits packets to \( k_p \) destinations for multicast traffic, we assume that there are \( k_p \) identical copies of the packet in the buffer of the source node. Thus, we can model a source queue as a set of \( k_p \) independent sub-queues, in which each sub-queue is intended for one destination. Then the process of directly sending a packet to \( k_p \) destinations can be viewed as the packet being transmitted through \( k_p \) independent source-destination routings, as illustrated in Figure 2.

Denote the delay in each routing by \( D_{s \rightarrow d}^j \). Then we have

\[
D_{s \rightarrow d} = \max_{1 \leq j \leq k_p} \{D_{s \rightarrow d}^j\}.
\]

To calculate the delay in this scenario, we first obtain the input and output rate for each source-to-destination sub-queue in the following lemma.

**Lemma 1:** Each source-to-destination sub-queue is a \( M/M/1 \) queue corresponding to one destination node with input rate \( \lambda_{\text{sub}}^i = \frac{\lambda_p R_p}{k_p} \) and service rate \( \lambda_{\text{sub}}^o = \frac{R_p}{k_p} \).

**Proof:** It is clear that the probability that the packet will be sent directly from the source to the destination is \( \frac{R_p}{k_p} \). As illustrated in Figure 2, for each sub-queue, we pick one node from these \( k_p \) nodes in the source pool as the sub-queue’s source. Therefore, in each time slot, each sub-queue has a probability of receiving a packet. Hence, the input rate of each sub-queue is

\[
\lambda_{\text{sub}}^i = \lambda_p R_p \to d = \frac{\lambda_p R_p R_{1}}{\lambda_p}.
\]

The output rate of each sub queue is equal to the communication rate of a source-to-destination pair \( \lambda_{\text{sub}}^o = \frac{R_p}{k_p} \) because the \( k_p \) destinations are equally likely to get the packet directly from the source.

Thus, the expected queueing time for each sub-queue is \( \frac{1}{\lambda_{\text{sub}}^o - \lambda_{\text{sub}}^i} \). According to Lemma 9, we can conclude that

\[
D_{s \rightarrow d} = \begin{cases} \Theta \left( \frac{1}{\lambda_{\text{sub}}^o - \lambda_{\text{sub}}^i} \right), & k_p = 1 \\ \Theta \left( \frac{\log k_p}{\lambda_{\text{sub}}^o} \right), & k_p > 1 \end{cases}. \tag{3}
\]

On the other hand, if the packet is transmitted through a relay node, the delay is composed of both the queueing time at source node and relay node. We denote the delay in the first phase (from source to relay) by \( D_{s \rightarrow r}^p \) and delay in the second phase (from relay to destination) by \( D_{r \rightarrow d}^p \). Then we have the following lemma.

**Lemma 2:** In the first phase of 2-hop routing strategy under the non-cooperative scheme, the expected delay before a primary source node can find a normal relay node to send a packet is equal to \( 1/(R_{2p} \lambda_{s \rightarrow r} - \lambda_p) \), where \( \lambda_{s \rightarrow r} = 1 - \frac{1}{n} \) is the average response time (including both the waiting time and service time).

**Proof:** The probability that a packet is handed over by a relay in the primary network is \( \frac{R_{2p}}{\lambda_p} \). Thus, for a source-to-relay queue with input rate \( \lambda_p \) and service rate \( \lambda_{s \rightarrow r} \), the average response time is \( 1/(\lambda_{s \rightarrow r} - \lambda_p) \).

Next we need to find \( \lambda_{s \rightarrow r} \). Consider a source node \( v_i \) that is scheduled to send a packet to a relay node. According to our communication scheme, the source node can find a normal relay node only when there exists at least one node \( v_i \) in the same cell as the source. The probability for this condition is

\[
p_{s \rightarrow r} = 1 - (1 - \frac{1}{n})^{n-k_p-1}.
\]

Then we have \( \lambda_{s \rightarrow r} = \frac{R_{2p} \lambda_{s \rightarrow r}}{\lambda_p} \). Therefore, we obtain the delay in the first phase as

\[
D_{s \rightarrow r}^p = \frac{1}{\lambda_{s \rightarrow r} - \lambda_p} = \frac{1}{\frac{R_{2p} \lambda_{s \rightarrow r}}{\lambda_p} - \lambda_p}. \tag{4}
\]

Thus we conclude this lemma.

Next we will calculate the delay in the second phase \( D_{r \rightarrow d}^p \).

Note that the probability that the packet is sent to a relay node is \( p_{s \rightarrow r}^k = \frac{\lambda_{s \rightarrow r}}{n-k_p-1} = \frac{R_{2p}}{\lambda_p} (n-k_p-1) \), because all these \( n-k_p-1 \) primary nodes outside the group \( G_i \) are equally likely to be chosen as a relay under the non-cooperative scheme. Then

\[
\lambda_{s \rightarrow r} = \frac{1}{\lambda_{s \rightarrow r} - \lambda_p} = \frac{1}{\frac{R_{2p} \lambda_{s \rightarrow r}}{\lambda_p} - \lambda_p}.
\]

7Actually, according to Taylor’s formula, we have \( 1 - (1 - \frac{R_{2p}}{\lambda_p})^{n-k_p-1} \). Since the smaller part \( o(k_p \lambda_p) \) makes no difference to our results, we will omit it in paper and assume that \( p_{s \rightarrow r}^k = 1 - (1 - \frac{R_{2p}}{\lambda_p}) \).
the input rate for a relay-to-destination queue in the primary network $\lambda^i_r$ is

$$\lambda^i_r = \lambda_p P^i_{s \to r} = \lambda_p \mathcal{R}_2 p^p_{s \to r} (n - k_p - 1).$$

A relay node carrying packets from one source node will forward the packet to all the $k_p$ destinations. Hence, similar to the previous discussion, we can model this process as a packet transmitted through $k_p$ identical sub-queues and each destination is associated with one certain sub-queue, as shown in Figure 3.

![Figure 3. Illustration of the relay-to-destination transmission process in the primary network under the non-cooperative scheme. Since one packet is intended for $k_p$ destinations, we assume the designated relay will duplicate the packet into $k_p$ identical copies and then distribute them to the corresponding destinations. Then a relay queue can be modeled as a set of $n - k_p - 1$ independent sub-queues, among which are $k_p$ sub-queues intended for destinations.](image)

Thus, using the same technique as in the previous analysis, the input rate of one sub-queue is $k_p \lambda^i_r$, and the input rate is $\frac{\mathcal{R}_2}{n - k_p - 1}$ because one primary node can relay packets for all the other $n - k_p - 1$ nodes except the $k_p + 1$ nodes in its group. Therefore, the expected queuing time for each sub-queue in the relay node is $\frac{1}{k_p \lambda^i_r}$. By Lemma 9, we can obtain

$$D^p = \begin{cases} \left( \frac{1}{k_p \lambda^i_r} \right) \log n_k, & k_p = 1 \\ \left( \frac{1}{k_p \lambda^i_r} \right) - \frac{1}{k_p \lambda^i_r}, & k_p > 1 \end{cases}. \tag{5}$$

Combining Equations (3), (4) and (5), we can obtain that the total expected delay for the primary network under multicast traffic ($k_p > 1$) is

$$E[D_p] = \frac{\mathcal{R}_1}{\lambda_p^o} D^p_{s \to d} + \frac{\mathcal{R}_2}{\lambda_p^o} (D^p_{s \to r} + D^p_{r \to d})$$

$$= \frac{\mathcal{R}_1}{\lambda_p^o} \cdot \Theta \left( \frac{\log k_p}{k_p \lambda^i_r}, \frac{\mathcal{R}_2}{\lambda_p^o} \frac{1}{k_p - k_p \lambda^i_r} \log k_p \right)$$

$$+ \frac{\mathcal{R}_2}{\lambda_p^o} \left( \frac{\mathcal{R}_2 p^p_{s \to r}}{\lambda_p^o} \frac{1}{k_p - k_p \lambda^i_r} \log k_p \right). \tag{6}$$

In the mathematical expression of $E[D_p]$ in Equation (6), the first item represents the delay of the 1-hop source-to-destination transmission and the second item represents the delay of the 2-hop source-to-relay-to-destination transmission. From the previous results, we find that the value of $k_p$ (whether $k_p$ equals $o(n)$ or $\Theta(n)$) makes a major difference on the transmission pattern of the primary network since both $\mathcal{R}_1$ and $\mathcal{R}_2$ are determined by $k_p$. Therefore, we need to study the capacity and delay tradeoff based on the value of $k_p$.

1) If $k_p = o(n)$, then $p_2 \sim 0$ and $p^p_{s \to r} \sim 1 - e^{-p_1}$. Thus, $\mathcal{R}_1 \sim 0$ and $\mathcal{R}_2 = \lambda^p_{s \to r} \sim \frac{1 - (1 + p_2) e^{-p_1}}{2p_1}$, which implies that the dominant communication pattern for the primary network is the 2-hop source-to-relay-to-destination transmission strategy. Therefore, $E[D_p]$ is determined by the second item of Equation (6). Since the input rate of a stable queue must be smaller than its output rate, we have

$$\lambda_p < \frac{\mathcal{R}_2 p^p_{s \to r}}{\lambda_p^o} \sim \frac{(1 - e^{-p_1}) \mathcal{R}_2}{\lambda_p^o},$$

$$k_p \lambda^i_r \sim \frac{1 - e^{-p_1} \lambda p \mathcal{R}_2}{\lambda^i_r (n - k_p - 1)} < \frac{\mathcal{R}_2}{(n - k_p - 1)}.$$ 

The constraint of the above two inequalities allows the per-node capacity $\lambda_p < \frac{1 - e^{-p_1} \lambda p \mathcal{R}_2}{\lambda^i_r (n - k_p - 1)}$, which means a per-node capacity $O(1/k_p)$ for the primary network is achievable when $k_p = o(n)$. A delay $E[D_p] = \Theta(n \log k_p)$ is achievable in this case.

2) If $k_p = \Theta(n)$, then we have $p_1 \sim 1 - (1 + p_2) e^{-p_1}$. Thus, $\mathcal{R}_2 \sim 0$ and $\mathcal{R}_1 = \lambda^p_{s \to r} \sim \frac{1 - (1 + p_2) e^{-p_1}}{2p_1}$, which implies that the 1-hop source-to-destination transmission strategy is the dominant communication pattern for the primary network. Therefore, $E[D_p]$ is determined by the first item of Equation (6). Similarly, we have the following inequality

$$\lambda_p k_p \mathcal{R}_1 \lambda^i_r \sim \frac{\mathcal{R}_1}{\lambda_p}. \tag{7}$$

Thus, the per-node capacity $\lambda_p = O(1/k_p^2)$ and delay $E[D_p] = \Theta(k_p \log k_p) = \Theta(n \log k_p)$ are achievable for the primary network when $k_p = \Theta(n)$.

This concludes the proof of Theorem 1. \[
\square
\]

Note that when $k_p = \Theta(n)$, the per-node capacity decreases to $O(1/k_p^2)$ but the delay is still on the order of $\Theta(n \log k_p)$, which implies that the capacity performance of our communication scheme is relatively poor as $k_p$ increases to the same order as $n$. This motivates us to improve the scheme by allowing packets to be transmitted from one primary destination node to another, which will be discussed in the following subsection.

B. Capacity and Delay Analysis of the Cooperative Scheme

In this subsection, we bring forward the cooperative transmission scheme for the primary network. We will show that, by utilizing the cooperation among destinations within the same multicast group, the capacity of the primary network can be improved when $k_p = \Theta(n)$.

It is clear that under the cooperative scheme, packets can be transmitted only through the 2-hop source-to-relay-to-destination pattern. The difference from the previous subsection is that the relay node for each packet is selected from $n - 1$
nodes rather than from \( n - k_p - 1 \) nodes. Our main results under this scenario are presented in the following theorem.

**Theorem 2**: In a cell-partitioned network with overlapping \( n \) primary nodes and \( m \) secondary nodes, the achievable per-node capacity for the primary network is \( \lambda_p = O(1/k_p) \) with average delay \( \mathbb{E}[D_p] = \Theta(n \log k_p) \) under the cooperative scheme.

**Proof**: The proof is similar as that of Theorem 1. To avoid confusion, we will adopt the same mathematical symbols as the previous subsection. Since only the 2-hop source-to-destination pattern is available under the cooperative scheme, we have \( R_2 = \lambda_p^2 = \frac{\lambda_p}{R_1} \). For the source-to-destination delay \( D_{s \rightarrow r}^p \), we have the following lemma.

**Lemma 3**: In the first phase of 2-hop routing strategy under the cooperative scheme, the expected delay before a primary source node can find a relay node to send a packet is

\[
D_{s \rightarrow r}^p = \frac{1}{(R_2p_{s \rightarrow r}^p - \lambda_p)} = \frac{1}{p_{s \rightarrow r}^p - \lambda_p},
\]

(7)

where \( p_{s \rightarrow r}^p = 1 - (1 - \frac{1}{n})^{n-1} \).

**Proof**: The proof is quite similar to that of Lemma 2 and we omit it to avoid triviality. \( \blacksquare \)

Now we will discuss the delay in the second phase \( D_{r \rightarrow d}^p \). Under the cooperative scheme, when destination nodes can serve as relay nodes for any multicast group, the input rate for a relay-to-destination queue \( \lambda_r^i \) is

\[
\lambda_r^i = \frac{\lambda_p R_2 p_{s \rightarrow r}^o}{\lambda_p^o(n-1)} = \frac{\lambda_p p_{s \rightarrow r}^o}{n-1}.
\]

If a destination node is chosen as a relay node, then it only needs to transmit packets to the remaining \( k_p - 1 \) destination nodes. Otherwise, a normal relay node must send packets to all the \( k_p \) destination nodes. It is clear that whether \( k_p - 1 \) or \( k_p \) destination nodes that are supposed to receive the packet makes no difference to the delay in an order sense. Therefore, we assume that a relay node has to relay packets for \( k_p \) destination nodes regardless of whether it is a destination node or not. Hence, using the same technique as the previous subsection, we have that the input rate for a sub-queue is \( k_p \lambda_r^i \) and the output rate is \( \frac{\lambda_p}{n-1} \). Therefore, the expected queuing time for each sub-queue in the relay node is \( \frac{1}{\lambda_p - k_p \lambda_r^i} \).

According to Lemma 9, we can obtain

\[
D_{r \rightarrow d}^p = \begin{cases} 
\Theta\left(\frac{1}{\lambda_p - k_p \lambda_r^i}\right), & k_p = 1 \\
\Theta\left(\frac{\log k_p}{\lambda_p - k_p \lambda_r^i}\right), & k_p > 1
\end{cases}.
\]

(8)

Combining Equations (7) and (8), we have that the total expected delay under multicast traffic \( (k_p > 1) \) is

\[
\mathbb{E}[D_p] = D_{s \rightarrow r}^p + D_{r \rightarrow d}^p
\]

**From Equation (9), it is clear that the following two inequalities hold:**

\[
\begin{aligned}
\lambda_p < \frac{R_2}{\lambda_p^o} \sim \frac{(1-e^{-\tau_p})R_2}{\lambda_p^o} \\
k_p \lambda_r^i \sim \frac{(1-e^{-\tau_p})k_p \lambda_r^o R_2}{\lambda_p^o(n-1)} < \frac{R_2}{(n-1)}.
\end{aligned}
\]

By solving the above two inequalities, we have that the per-node capacity \( \lambda_p < \frac{R_2}{(1-e^{-\tau_p})k_p} \), which indicates that per-node capacity \( O(1/k_p) \) is achievable under the cooperative scheme regardless of the order of \( k_p \). Note that the expected delay \( \mathbb{E}[D_p] \) is still in the order of \( \Theta(n \log k_p) \). Therefore we have showed that the cooperative scheme is able to improve the per-node capacity for the primary network when \( k_p = \Theta(n) \). This concludes the proof for Theorem 2.

\( \blacksquare \)

From the above results, we find that the per-node capacity of the primary network is always scaled by \( O(1/k_p) \) irrespective of the value of \( k_p \) under the cooperative scheme, which implies that the cooperative scheme performs better than the non-cooperative scheme.

**V. Capacity and Delay Analysis for the Secondary Network**

In this section, we will discuss the capacity and delay tradeoff for the secondary network. Unlike the scheme in the primary network, a node in the secondary network can only opportunistically transmit whenever it is outside the region of active primary cells. Thus, when scaling the capacity and delay for the secondary network, we should consider the impact of the primary network and the transmission interruption of secondary nodes. In the following, we first analyze the average capacity and delay under the cooperative scheme. We then introduce the redundancy scheme and destination oriented redundancy scheme to help reduce the delay and efficiently utilize the network resources for the secondary network.

**A. Capacity and Delay Analysis of the Cooperative Scheme**

The following lemma indicates that, with the communication schemes defined previously, the secondary nodes (whether source nodes or relay nodes) have finite opportunities to deliver their packets. And later we can prove that the whole secondary network appears to have the same capacity and delay performance as a stand-alone MANET in order sense.

**Lemma 4**: With the proposed communication scheme, these two results hold:

1) In each time slot, a constant fraction of the secondary cells is outside the region of active primary cells, which can be scheduled successfully for transmission.

2) Each individual secondary cell has a finite spectrum opportunity to transmit.

**Proof**: Let \( c_q(n) \) be the fraction of primary cells with \( q \) nodes \( (q \geq 0) \). According to Lemma 1 in [19], \( c_q(n) = e^{-1}/q! \) w.h.p.. Then \( 1 - 2e^{-1} \approx 0.26 \) fraction of the primary cells contains at least two primary nodes w.h.p.. This implies that the rest fraction of primary cells may not be active and allow
the secondary nodes access to the spectrum for transmitting. Thus we conclude the first part of the lemma.

The above part implies a finite transmission opportunity for the overall secondary cells. We have to further consider the transmission opportunity of each individual secondary cell. Recall that the secondary network adopts a 25-TDMA scheme with adjacent-neighbor communication. In each primary time slot, we have a complete secondary TDMA frame in our communication scheme. Each active secondary cell will be assigned with at least one active secondary TDMA slot within each secondary frame. This completes the proof for the lemma.

Lemma 5: The total service rate for a secondary source queue is $\lambda_s^p = \frac{p_3}{25}$, where $p_3 \sim 1 - (1 + \tau_s)e^{-\tau_s}$.

Proof: Denote the probability that there are at least two secondary nodes in one cell by $p_3$, which can be calculated by

$$p_3 = \left(1 - \left(1 - \frac{1}{c}\right)\right)^m - \frac{m}{c}\left(1 - \frac{1}{c}\right)^{m-1} \sim 1 - (1 + \tau_s)e^{-\tau_s}.$$  

Then we obtain that the total service rate for a secondary source node is $\lambda_s^p = \frac{p_3}{25}$. Thus we conclude this lemma. \hfill \Box

Our results on achievable per-node capacity and transmission delay are given in the following.

Theorem 3: In a cell-partitioned network with overlapping $n$ primary nodes and $m$ secondary nodes, the achievable per-node capacity for the secondary network under the cooperative scheme is $\lambda_s = O(1/k_s)$ with average delay $\mathbb{E}[D_s] = \Theta(m \log k_s)$.

Proof: For the 2-hop source-to-relay-to-destination communication strategy, both the queueing time at the secondary source node and relay node are accounted for in the delay. We denote the delay in first phase (from a secondary source node to a relay node) by $D_s^{*} \rightarrow r$ and delay in second phase (from a relay node to all the $k_s$ secondary destination nodes) by $D_s^{*} \rightarrow d$. Then, similar to Lemma 2 in previous section, we have the following lemma.

Lemma 6: In the first phase of the cooperative scheme in the secondary network, the expected delay for a secondary source to send one packet to a relay node equals $25/c_1(1 - e^{-\tau_s} - \lambda_s)$, where $c_1(0 < c_1 < 1)$ is a constant.

Proof: Since a secondary frame is divided into 25 sub-frames, the transmission rate for a secondary cell decreases by a factor of 25. Moreover, according to Lemma 4, there are a constant fraction of secondary cells in the unit area that can access the spectrum. Here we use constant $c_1$ to represent the overall likelihood that the secondary cells will be successfully scheduled during one second. Therefore, the input rate and output rate for a secondary source-to-relay queue are $rac{c_1}{25}\lambda_s$ and $\frac{c_1}{25}\mu_s^{d}$, with average delay $\frac{1}{25}(\mu_s^{d}-\lambda_s)$. Then we need to calculate the service rate $\mu_s^{r}$. The probability that a secondary source node is able to send its packets to a relay node is $1 - (1 - \frac{1}{c})^{m-1} \sim 1 - e^{-\tau_s}$ since any of the secondary nodes in the same cell with the secondary source node can serve as relays. Therefore $\mu_s^{r} = 1 - e^{-\tau_s}$ and then we can conclude that

$$D_s^{*} \rightarrow r = \frac{25}{c_1(1 - e^{-\tau_s} - \lambda_s)}. \quad (11)$$

Now we need to get the delay in the second phase $D_s^{*} \rightarrow d$. Note that a secondary source node relays a packet from a secondary source node with probability $p^{s} \rightarrow r = \frac{1 - e^{-\tau_s}}{m-1}$ because all the secondary nodes are equally likely to be chosen as a relay. Then the input rate for a relay-to-destination queue is $\mu_s^{r} = \frac{c_1}{25}\lambda_s p^{s} \rightarrow r = \frac{1 - e^{-\tau_s}}{25}(m-1)$.

Similar to the previous section, we model the process that a relay node distributes one packet from a secondary source node to all the $k_s$ secondary destinations as the packet being transmitted through $k_s$ independent sub-queues and each destination node is associated with one certain sub-queue.

Therefore, we can conclude that the input rate of one sub-queue is $k_s \mu_s^{r}$ and the output rate is $c_1(\lambda_s^{p}/(m-1)-k_s \mu_s^{r})$ because one secondary node can relay packets destined for all the other $m-1$ nodes with equal probability. Therefore, the expected queueing time for each sub-queue in the relay node is $\frac{25}{c_1(\lambda_s^{p}/(m-1)-k_s \mu_s^{r})}$. By Lemma 9, we can obtain

$$D_s^{*} \rightarrow d = \left\{ \begin{array}{ll}
\Theta\left(\frac{25}{c_1(\lambda_s^{p}/(m-1)-k_s \mu_s^{r})}\right), & k_s = 1 \\
\Theta\left(\frac{25 \log k_s}{c_1(\lambda_s^{p}/(m-1)-k_s \mu_s^{r})}\right), & k_s > 1
\end{array} \right. \quad (12)$$

Combining the Equations (11) and (12), we get the total expected delay for the secondary network under multicast traffic ($k_s > 1$) is

$$\mathbb{E}[D_s] = D_s^{*} \rightarrow r + D_s^{*} \rightarrow d = \frac{25}{c_1(1 - e^{-\tau_s} - \lambda_s)} + \Theta\left(\frac{25 \log k_s}{c_1(\lambda_s^{p}/(m-1)-k_s \mu_s^{r})}\right). \quad (13)$$

To make sure that the number of packets waiting to be transmitted in each queue does not increase to infinity with time, the following two constraints must be satisfied.
\[
\begin{align*}
\lambda_s &< 1 - e^{-\tau_s k_s}, \\
k_s \mu^i_s &< \frac{c_1(1-e^{-\tau_s})\lambda_s k_s}{20(m-1)} < \frac{c_1 \lambda_s}{25(m-1)}.
\end{align*}
\]

From the above two inequalities, we can obtain \( \lambda_s < \frac{\lambda_s}{(1-e^{-\tau_s} k_s)} \). Thus per-node capacity \( \lambda_s = O(1/k_s) \) and average delay \( \mathbf{E}[D_s] = \Theta(m \log k_s) \) are achievable for the secondary network. Thus we conclude the Theorem 3. ■

Note that in our system model and communication scheme, the primary network and secondary network achieve per-node capacity \( O(1/k_p) \) and \( O(1/k_s) \), respectively. Therefore, their co-existence and mutual interference make no difference for their throughput scaling and the two overlapping networks can achieve similar capacity and delay scaling as if they were two independently stand-alone networks.

B. Capacity and Delay Analysis of the Redundancy Scheme

In this part, we will discuss the capacity and delay tradeoff when more than one secondary node can serve as a relay, i.e., the redundancy scheme, for the secondary network.

Intuitively, the time needed for a packet to reach its destination can be reduced by repeatedly sending this packet to many other secondary nodes, i.e., using more than one node as a relay. In this way, the chances that some node carrying an original or duplicate version of the packet finds a destination will increase. Thus, it is supposed that adopting the redundancy scheme can reduce communication delay although this may not help to improve the network throughput. Previous works [3] and [13] have also introduced the redundancy scheme for a stand-alone ad hoc network in which the source node sends packets to more than one relay node. In CR networks, the redundancy scheme is also applicable for both the primary network and the secondary network. However, considering that secondary nodes suffer from a larger average-delay than the primary nodes, we namely introduce this redundancy scheme to the secondary network to help effectively reduce its end-to-end delay.

Here, we first show the lower bound of average delay, which is given in the following.

Theorem 4: In a cell-partitioned network with overlapping \( n \) primary nodes and \( m \) secondary nodes, no communication scheme with redundancy can achieve an average delay lower than \( O(\sqrt{m \log k_s}) \) for the secondary network.

Proof: We prove this result by considering an ideal situation where the secondary network is empty except that one secondary source node sends a packet to \( k_s \) destinations. The optimal scheme for the source is to send duplicate versions of the packet whenever it meets new nodes that have never received the packet before. These duplicate-carrying nodes will then act as relays to help transmit the packet to the destinations as soon as they enter the same cell as one of the destinations.

During the time slots \( \{1, 2, \ldots, i\} \), let \( \vartheta_j \) be the total number of intermediate relay nodes at the beginning of time slot \( j(1 \leq j \leq i) \). Clearly, \( \vartheta_1 \leq \vartheta_2 \leq \cdots \leq \vartheta_i \) since the number of relays is non-decreasing across time. Note that a relay can only be generated by the secondary source node, hence at most one node can be a new relay in every time slot. Thus \( \vartheta_i \leq i \) for all \( i \geq 1 \). Furthermore, denote the probability that a destination node can meet at least one relay node during time slots \( \{1, 2, \ldots, i\} \) by \( p^{(i)} \), we have

\[
p^{(i)} = 1 - \prod_{j=1}^{i} \left( 1 - \frac{1}{c} \vartheta_j \right) \leq 1 - \left( 1 - \frac{1}{c} \right)^i \leq \left( 1 - \frac{1}{c} \right)^i.
\]

However, a destination meeting a relay does not necessarily ensure the transmission can be scheduled since the secondary cell in which they meet may not be active during that slot. Similar to previous discussion in Lemma 6, a constant factor \( \frac{1}{p^{(i)}} \) should be multiplied. Hence, the probability that a destination node can successfully receive a packet from a relay during time slots \( \{1, 2, \ldots, i\} \) is \( c_1 p^{(i)} / 25 \). Then we have

\[
\begin{align*}
\Pr(D_s \geq i) &\geq 1 - \frac{c_1}{25} p^{(i)} k_s = 1 - \frac{c_1}{25} k_s \left( 1 - \left( 1 - \frac{1}{c} \right)^2 \right) k_s \geq 1 - \frac{c_1}{25} k_s e^{-1},
\end{align*}
\]

We choose \( i = \sqrt{\frac{\log k_s}{\tau_s}} \) and let \( k_s \to \infty \), we obtain that

\[
\Pr(D_s \geq i) \geq 1 - \frac{c_1}{25} k_s \left( 1 - \frac{1}{c} \right) k_s \geq 1 - \frac{c_1}{25} e^{-1}.
\]

Therefore, the average delay in the secondary network with redundancy satisfies

\[
\mathbf{E}[D_s] \geq \mathbf{E}[D_s | D_s \geq i] \Pr(D_s \geq i) \geq (1 - \frac{c_1}{25} e^{-1}) \mathbf{E}[\frac{m \log k_s}{\tau_s}]
\]

as \( m \) and \( k_s \) approach infinity, which concludes the theorem. ■

To achieve the above proved lower bound of delay, we propose the following redundancy scheme for the secondary network.

Redundancy Scheme: For an active secondary cell containing at least two nodes, the following two transmission patterns are performed with equal probability:

1) Source-to-Relay Transmission: Randomly choose a node that the secondary source node meets in the same secondary cell as a relay. Pick the source node as the sender and the relay node as receiver to finish the transmission. If the source node has forwarded the duplicate of the packet to at least \( \sqrt{\log k_s} \) distinct relays (possibly some of the destinations), the packet is deleted from the buffer of source node.

2) Relay-to-Destination Transmission: If one relay node, carrying a packet destined for secondary destination...
nodes, can be found in this cell and meanwhile at least one corresponding “pristine” secondary destination node resides within the same cell, pick the relay node as the sender and the destination node as receiver to finish the transmission. A packet will be discarded when all its $k_s$ destinations have received it.

Then we have the following theorem.

**Theorem 5:** In a cell-partitioned network with overlapping $n$ primary nodes and $m$ secondary nodes, the achievable per-node capacity for the secondary network is $O(1/k_s \sqrt{m \log k_s})$ with average delay $E[D_s] = O(\sqrt{m \log k_s})$ under the redundancy scheme.

**Proof:** Under the redundancy scheme, the expected time required for a certain packet to reach its corresponding $k_s$ destinations $E[D_s]$ is less than $E[D^*_1] + E[D^*_2]$, where $E[D^*_1]$ represents the expected time required to distribute the duplicates to $\sqrt{m \log k_s}$ different nodes, and $E[D^*_2]$ is the expected time required to reach all the $k_s$ destinations given that $\sqrt{m \log k_s}$ relays are holding the packet. We then upper bound $E[D^*_1]$ and $E[D^*_2]$, respectively.

$E[D^*_1]$ bound: For the duration of $D^*_1$, there are at least $m - \sqrt{m \log k_s}$ secondary nodes that do not have the packet. Hence, in every time slot, the probability that at least one of these nodes is located in the same cell as the source is at least $1 - (1 - \frac{1}{\sqrt{m \log k_s}}) = m^{1/2}/\sqrt{m \log k_s}$. Every time slot in the duration $D^*_1$, a source node can successfully deliver a duplicate packet to a new node with probability of at least $p^*$ given by

$$p^* \geq \frac{c_1}{25} \alpha_1 \alpha_2 \log k_s \left( 1 - \frac{1}{e} \right)^{m - \sqrt{m \log k_s}},$$

where $\alpha_1$ is the probability that the source is selected from all the other nodes in the same cell to be the transmitting node, and $\alpha_2$ is the probability that this source is chosen to operate the “source-to-relay” transmission which is equal to $1/2$. According to Lemma 6 in [3], $\alpha_1 \geq 1/(2 + \tau_s)$. Thus,

$$p^* \rightarrow \frac{c_1}{25} \frac{1 - e^{-\tau_s}}{4 + 2\tau_s}$$

as $m$ approaches infinity.

The average time for a duplicate packet to reach a new node is upper bounded by $1/p^*$. Since there are $\sqrt{m \log k_s}$ duplicates to be transmitted, in the worst case, $\sqrt{m \log k_s}$ of these times are required. Therefore, $E[D^*_1]$ is upper bounded by $\sqrt{m \log k_s}/p^*$. Hence, we have $E[D^*_1] \leq \frac{c_1}{25} \frac{1 - e^{-\tau_s}}{4 + 2\tau_s} \sqrt{m \log k_s}$.

$E[D^*_2]$ bound: Every time slot in the duration of $D^*_2$, there are at least $\sqrt{m \log k_s}$ nodes holding the duplicates of the packet. The probability that at least one other node is in the same cell as the destination is $1 - \left(1 - \frac{1}{\sqrt{m \log k_s}} \right)^{m-1}$. Each time slot, a destination node can successfully receive a duplicate packet from one of these relay nodes with probability of at least $p^*$ given by

$$p^* = \frac{c_1}{25} \beta_1 \beta_2 \beta_3 \left( 1 - \frac{1}{e} \right)^{m-1},$$

where $\beta_1$ is the probability that the destination is selected from all the other nodes in the same cell to be the receiver, $\beta_2$ is the probability that the sender is chosen to operate the “relay-to-destination” transmission which is equal to $1/2$, and $\beta_3$ represents the possibility that the sender is one of these $\sqrt{m \log k_s}$ nodes possessing a duplicate packet intended for the destination. Similarly, we have $\beta_1 \geq 1/(2 + \tau_s)$ and $\beta_2 = \sqrt{m \log k_s}/(m-1) \geq \sqrt{\log k_s}/m$. Thus,

$$p^* \rightarrow \frac{c_1}{25} \frac{1 - e^{-\tau_s}}{4 + 2\tau_s} \sqrt{m \log k_s}/m$$

as $m$ approaches infinity. The average time that a single destination node receives a duplicate packet destined for it is upper bounded by $1/p^*$. The time needed for all the $k_s$ destination nodes to receive the packet is the maximum of them. Again according to Lemma 9, $E[D^*_2]$ is upper bounded by $\log k_s/p^*$. Hence, we have $E[D^*_2] \leq \frac{25}{4 \tau_s} \frac{1 + e^{-\tau_s}}{\sqrt{m \log k_s}}$.

Finally, according to Lemma 2 in [3], the total delay can be upper bounded by $E[D_s] = O(\sqrt{m \log k_s})$ and we obtain an achievable per-node capacity $O(1/k_s \sqrt{m \log k_s})$. Thus we conclude the theorem.

**Theorem 6:** In a cell-partitioned network with overlapping $n$ primary nodes and $m$ secondary nodes, the achievable per-node capacity for the secondary network is $O(1/k_s \sqrt{m \log k_s})$ with average delay $E[D_s] = \Theta(\sqrt{m \log k_s})$ under the redundancy scheme.

**Proof:** This can be directly obtained by combining Theorem 4 and Theorem 5.

## C. Capacity and Delay Analysis of Destination Oriented Redundancy Scheme

The above mentioned redundancy scheme can effectively reduce the end-to-end delay, however, the shortcoming is that repeatedly sending one packet to more than one relay node will consume more network resources and increase interference. To avoid this, we improve the original redundancy scheme and propose a novel redundancy scheme, named the **Destination Oriented Redundancy Scheme**, which not only reduces delay but also better utilizes network resources.

### Destination Oriented Redundancy Scheme

For an active secondary cell containing at least two secondary nodes, the following two transmission patterns are performed with equal probability.

1. **Source-to-Relay Transmission:** Choose the first node that the secondary source node meets as a relay irrespective of whether it is a destination node or not. Pick the source node as the sender and the relay node as receiver to finish the transmission.

2. **Relay-to-Destination Transmission:** If one relay node, carrying a packet destined for secondary destination nodes, can be found in this cell and meanwhile at least one corresponding “pristine” secondary destination node resides within the same cell, pick the relay node as the sender and the destination node as receiver to finish the transmission. For the destination nodes which have received the packet, they can also serve as relay nodes to conduct the relay-to-destination transmission.
In the above mentioned scheme, to reduce the delay by bringing in redundancy, we allow the destination nodes who have received the message to perform as relay nodes. Note that at most one node besides the source and $k_s$ destinations is chosen as a relay, hence we do not introduce extra relay nodes to relay the packet. Thus we save the network resources with this improved scheme.

With this new redundancy scheme, we show that the lower bound of communication delay for secondary network is $O(\sqrt{m \log k_s})$ if we allow only one transmission in one time slot. However, if we assume that all the available transmissions among different active cells can be conducted in one time slot, we prove that the lower bound of communication delay is $O(\frac{m \log k_s}{k_s})$. Formally, we define these two communication patterns as Destination Oriented Solo-redundancy Scheme (use “Solo-redundancy Scheme” for short) and Destination Oriented Ensemble-redundancy Scheme (use “Ensemble-redundancy Scheme” for short), respectively.

**Definition 2:** Solo-redundancy Scheme refers to the scheme when at most one destination node within the same multicast group is allowed to receive packets in one time slot, even though there may be more than one active cell in which a packet from a certain source secondary node can be sent to a “pristine” secondary destination node. Whereas the Ensemble-redundancy Scheme allows all the available transmissions within the same multicast group among different active cells to be conducted in one time slot.

We are now ready to present our main results in the following.

**Theorem 7:** The lower bound of communication delay in the secondary network under the destination oriented redundancy scheme is

1) $B_1 \triangleq \Omega(\sqrt{m \log k_s})$ if we adopt the solo-redundancy scheme.

2) $B_2 \triangleq \Omega(\frac{m \log k_s}{k_s})$ if we adopt the ensemble-redundancy scheme.

**Proof:** We start with the proof of the first item in Theorem 7. Suppose during the time slots $\{1, 2, \ldots, i\}$, there are $\psi_i$ ($\psi_i \leq k_s$) destination nodes that have received the packet. Denote the number of destination nodes who have received the message in time slot $j$ ($1 \leq j \leq i$) by $\psi_j$. It is clear that $\psi_1 \leq \psi_2 \leq \cdots \leq \psi_j$. Furthermore, denote the probability that one destination node has not received the packet during the time slots $\{1, 2, \ldots, i\}$ by $p^i$. Then $p^i$ satisfies

$$p^i = \prod_{j=1}^{i} (1 - \frac{1}{e})^{\psi_j}$$

$$= (1 - \frac{1}{e})^{\sum_{j=1}^{i} \psi_j} \geq (1 - \frac{1}{e})^{\psi_i}.$$  

$$(\text{since } \psi_i \leq i \text{ under solo redundancy scheme})$$

Then we have

$$Pr(D_s \geq i) \geq 1 - \frac{c_1(k_s - \psi_i)}{25}(1 - \frac{1}{e})^{\psi_i}$$

$$\geq 1 - \frac{c_1(k_s - \psi_i)}{25} \left[1 - (1 - \frac{1}{e})^2\right](k_s - \psi_i) \quad (16)$$

We choose $i = \sqrt{\frac{m \log k_s}{\tau_k}}$. Substituting the value of $i$ into Equation (16), we obtain that

$$Pr(D_s \geq i) \geq 1 - \frac{c_1(k_s - \psi_i)}{25} \left(1 - e^{-\log k_s}\right)(k_s - \psi_i)$$

$$= 1 - \frac{c_1(k_s - \psi_i)}{25} \left(1 - \frac{1}{k_s}\right)(k_s - \psi_i)$$

$$\geq 1 - e^{-1}.$$  

Therefore, the expected communication delay in the secondary network under solo-redundancy scheme satisfies

$$E[D_s] = E\{D_s | D_s < i\} Pr(D_s < i)$$

$$+ E\{D_s | D_s \geq i\} Pr(D_s \geq i)$$

$$\geq E\{D_s | D_s \geq i\} Pr(D_s \geq i)$$

(since we omit the first item)

$$\geq (1 - e^{-1}) \sqrt{\frac{m \log k_s}{\tau_s}}$$

$$\sim \Omega(\sqrt{m \log k_s}).$$

Thus, we prove the first claim in Theorem 7. Next we prove the second claim. In this scenario, we have that $p^i$ satisfies

$$p^i = \prod_{j=1}^{i} (1 - \frac{1}{e})^{\psi_j}$$

$$= (1 - \frac{1}{e})^{\sum_{j=1}^{i} \psi_j} \geq (1 - \frac{1}{e})^{\psi_i}.$$  

$$(\text{since } \psi_i \leq k_s \text{ is always satisfied})$$

Then we have

$$Pr(D_s \geq i) \geq 1 - \frac{c_1(k_s - \psi_i)}{25}(1 - p^i)(k_s - \psi_i)$$

$$\geq 1 - \frac{c_1(k_s - \psi_i)}{25} \left[1 - (1 - \frac{1}{e})^2\right](k_s - \psi_i) \quad (18)$$

Here, we choose $i = \frac{m \log k_s}{\tau_k}$. Substituting that value into Equation (18), we obtain

$$Pr(D_s \geq i) \geq 1 - \frac{c_1(k_s - \psi_i)}{25} \left(1 - e^{-\log k_s}\right)(k_s - \psi_i)$$

$$\geq 1 - e^{-1}.$$  

Hence, the expected communication delay in the secondary network under ensemble-redundancy scheme satisfies
\[ E[D_s] \geq E[D_s | D_s \geq i] Pr(D_s \geq i) \geq (1-e^{-1}) \frac{m \log k_s}{\tau_s k_s} \sim \Omega\left(\frac{m \log k_s}{k_s}\right). \]

Note that under ensemble-redundancy scheme, more transmissions are allowed in one time slot compared to solo-redundancy scheme, thus the delay in ensemble-redundancy scheme should be smaller than that in the solo-redundancy scheme correspondingly. Therefore, the two delay lower bounds \( B_1 \) and \( B_2 \) should satisfy \( B_1 > B_2 \). It is clear that

1. If \( k_s = \Theta(m) \), then \( B_2 = \Omega\left(\frac{m \log k_s}{k_s}\right) \sim \Omega(\log k_s) \).
   Thus we conclude \( B_2 < B_1 \). This is acceptable.
2. If \( k_s = o(m) \), then \( B_2 > B_1 \), which implies that \( B_2 \) is not a very tight lower bound in this scenario.

For this case of \( B_2 \) when \( k_s = o(m) \), we can find an intuitive explanation according to our network model. To make the explanation more straightforward, here we assume that the first relay node met by the source node is one destination node. Note that the “pristine” destination nodes can get packets only when they meet other destination nodes carrying messages in one active cell. Denote the probability that one “pristine” destination node receives packets in one time slot by \( p_4 \). Then we have

\[
p_4 \leq 1 - \left( \frac{k_s}{1} \right)^1 (1 - \frac{1}{c})^{k_s - 1} - (1 - \frac{1}{c})^{k_s} = 1 - \left( 1 + \frac{k_s - 1}{c} \right) (1 - \frac{1}{c})^{k_s - 1} = 1 - \left( 1 + \frac{k_s - 1}{c} \right) e^{-\tau_s k_s}. \]

Hence, if \( k_s = o(m) \), \( p_4 \) is close to zero, which implies that relay-destination pairs are less likely to be found in a single active cell. Therefore, even though we allow more transmissions to be finished in different active cells during one time slot under ensemble-redundancy scheme, the actual number of transmissions conducted in one time slot may be very small since it is difficult to satisfy the condition for successful communication. Thus \( \psi_i \) is smaller than \( i \) with high probability even under ensemble-redundancy scheme. Recall the process of finding \( B_2 \), we used the inequality \( \psi_i \leq k_s \), a very loose constraint, while deducing Equation (17). So if we adopt a tighter constraint \( \psi_i \leq i \), we find \( B_2 = \Omega(\sqrt{m \log k_s}) \) which is much better than \( \Omega\left(\frac{m \log k_s}{k_s}\right) \). Therefore, we may modify our results on the lower bound of communication delay under ensemble-redundancy scheme as follows:

1. \( B_2 = \Omega(\sqrt{m \log k_s}) \) if \( k_s = o(m) \);
2. \( B_2 = \Omega(\log k_s) \) if \( k_s = \Theta(m) \).

Thus we conclude Theorem 7.

VI. CONCLUSION AND DISCUSSION

A. Results Review

In this paper, we have studied the capacity and delay tradeoff in CR MANET. For the primary network, we first studied a non-cooperative scheme. Our results showed that this scheme achieves a per-node capacity \( O(1/k_p) \) and average delay \( \Theta(n \log k_p) \) when \( k_p = o(n) \). However, if \( k_p = \Theta(n) \), the per-node capacity reduces to \( O(1/k_p^2) \). To improve this, we introduced the cooperative scheme and proved that per-node capacity \( O(1/k_p) \) is also achievable even if \( k_p = \Theta(n) \), which implies that cooperation among destinations within the same multicast group improves the throughput scaling.

For the secondary network, we proved that per-node capacity \( O(1/k_s) \) and average delay \( \Theta(\log k_s) \) are achievable under the cooperative scheme. Note that our results indicated that both the primary network and the secondary network can achieve nearly the same capacity and delay scaling in CR networks as if they were two independent stand-alone networks. Furthermore, in order to improve the delay performance in the secondary network, a redundancy scheme was considered and we showed that the per-node capacity \( O(1/k_s \sqrt{\log k_s}) \) and average delay \( \Theta(\sqrt{\log k_s}) \) are achievable. Finally, we improved the original redundancy scheme and proposed the destination oriented redundancy scheme to efficiently utilize wireless resources and reduce the delay for the secondary network users.

Formally, we present our results in the following three Tables II, III and IV.

![Table II](attachment:image2.png)

**TABLE II**

<table>
<thead>
<tr>
<th>Communication scheme</th>
<th>Capacity</th>
<th>Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-cooperative ( k_p = o(n) )</td>
<td>( O(1/k_p) )</td>
<td>( \Theta(n \log k_p) )</td>
</tr>
<tr>
<td>Non-cooperative ( k_p = \Theta(n) )</td>
<td>( O(1/k_p^2) )</td>
<td>( \Theta(n \log k_p) )</td>
</tr>
<tr>
<td>Cooperative scheme</td>
<td>( O(1/k_p) )</td>
<td>( \Theta(n \log k_p) )</td>
</tr>
</tbody>
</table>

**TABLE III**

<table>
<thead>
<tr>
<th>Communication scheme</th>
<th>Capacity</th>
<th>Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperative scheme</td>
<td>( O(1/k_s) )</td>
<td>( \Theta(\log k_s) )</td>
</tr>
<tr>
<td>Redundancy scheme</td>
<td>( O(1/k_s \sqrt{\log k_s}) )</td>
<td>( \Theta(\sqrt{\log k_s}) )</td>
</tr>
</tbody>
</table>

**TABLE IV**

<table>
<thead>
<tr>
<th>Communication scheme</th>
<th>Lower Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solo-redundancy scheme</td>
<td>( \Omega(\sqrt{\log k_s}) )</td>
</tr>
<tr>
<td>Ensemble-redundancy scheme ( k_s = o(m) )</td>
<td>( \Omega(\sqrt{\log k_s}) )</td>
</tr>
<tr>
<td>Ensemble-redundancy scheme ( k_s = \Theta(m) )</td>
<td>( \Omega(\log k_s) )</td>
</tr>
</tbody>
</table>

B. Comparison with Previous Work

Compared with the multicast capacity of static CR networks developed in [16], we find that the capacity performance is
better when nodes are mobile. Also, compared with the partial mobility model developed in [17] and [18], which requires the primary nodes to be static and only the secondary nodes are allowed to move, our model is more practical and general as both the primary and secondary networks can be mobile. Moreover, compared with the results of CR network under unicast traffic in [18], we find that the capacity diminishes by a factor of $1/k_p$ and $1/k_s$ for the primary network and the secondary network respectively under the cooperative scheme. This is because we need to forward a packet to $k_p$ primary destinations (or $k_s$ secondary destinations). Particularly, if $k_p = \Theta(1)$ and $k_s = \Theta(1)$, our results can be specialized to the unicast traffic; if $k_p = \Theta(n)$ and $k_s = \Theta(m)$, our results can be specialized to the broadcast traffic.

Furthermore, we find that although redundancy can reduce the transmission delay in the secondary network, it will lead to a decrease in the capacity. This suggests that transmissions with redundant relays can reduce delay at the cost of the capacity. The tradeoff between the delay and capacity always satisfies $\text{delay/capacity} \geq O(mk_s \log k_s)$ under both the cooperative scheme and the redundancy scheme in the secondary network, as shown in Figure 4. However, if we schedule the transmission in the secondary network by multiple unicast from source to $k_s$ destinations, the capacity will diminish by a factor of $k_s$ and the delay will increase by a factor of $k_s$, which indicates that the delay-capacity tradeoff becomes $\text{delay/capacity} \geq O(mk_s^2)$ in CR MANET. This demonstrates that our tradeoff is better than that of naively extending the tradeoff for unicast to multicast.

![Diagram of capacity-delay tradeoff in the secondary network](image)

C. Rationality of System Model

In our system model, we consider an ideal i.i.d fast mobility model, which allows nodes to choose new locations every timeslot from overall cells in the network. Indeed, actual mobility is better characterized by Markovian Dynamics or Random Walk mobility model, where nodes can only roam to adjacent cells in every timeslot. However, analysis under the ideal i.i.d mobility model provides a significant bound on capacity and delay performance in the limit of infinite mobility [3]. Under the i.i.d mobility model, all the nodes’ locations can not be predicted from time to time, hence the communication schemes are required to be more robust and adaptable. Our schemes are accordingly designed to better fit such features of the mobility model. Compared with other communication schemes, our schemes need less current and future information about users’ locations. In addition, it has been proved in [20] that the network capacity using an i.i.d mobility model is equivalent to the capacity region of the networks using non-i.i.d mobility model. Hence the results obtained in our paper are not only theoretically rational but also practically useful when applying to large-scale networks.

D. Future Work

Note that in our network model, we achieve a “harmonious” co-existence of the primary network and the secondary network by assuming primary nodes confine their interference to one cell. For a more general and practical network, we can introduce a “guard zone” to limit the interference received by primary and secondary nodes. The analysis of capacity and delay under this network model will be considered in future work. Additionally, the per-node capacity under the destination oriented redundancy scheme in CR MANET is still an open question. Finally, if the number of primary and secondary nodes are of the same order, it would be interesting to study the impact of the interference between them.

REFERENCES

as $n \to \infty$, which completes the proof for primary nodes. The number of secondary nodes can also be bounded in the similar way.

The following two lemmas are needed when scaling the throughput under multiscat traffic. The first lemma can be found in [13], and we include it here for completeness.

**Lemma 8.** \( \sum_{i=1}^{k} \frac{(-1)^{i-1}}{i} \binom{k}{i} = \ln(k+1)+r, \) where $k \geq 1$ and $r$ is the Euler constant.

**Proof:** Denote the left-hand side of the equation by $S(k)$, then we have $S(k-1) = \sum_{i=1}^{k} \frac{(-1)^{i-1}}{i} \binom{k-1}{i}$. Notice that $\binom{k-1}{i} = \binom{k-1}{i-1} + \binom{k-1}{i}$, and it follows

$$S(k) - S(k-1) = \sum_{i=1}^{k} \frac{(-1)^{i-1}}{i} \binom{k-1}{i-1}$$

Recall that $(1 - 1)^k = \sum_{i=0}^{k} (-1)^i \binom{k}{i} = 0$, hence we obtain $\sum_{i=1}^{k} (-1)^i \binom{k}{i} = -\sum_{i=1}^{k} (-1)^i \binom{k}{i} = -\left[ \sum_{i=0}^{k} (-1)^i \binom{k}{i} \right] - 1 = 1$. Combining with Equation 19, we get $S(k) - S(k-1) = \frac{1}{k}$, then

$$S(k) = S(1) + \sum_{i=2}^{k} \left[ S(k) - S(k-1) \right]$$

where the right-hand-side is the harmonic series. Hence this lemma holds.

**Lemma 9.** Suppose $X_1, X_2, \ldots, X_k$ are continuous i.i.d exponential random variables with expectation of $1/a$. Denote $X_{\max} = \max\{X_1, X_2, \ldots, X_k\}$, then

$$E\{X_{\max}\} = \begin{cases} \Theta(1/a), & k = 1 \\ \Theta(k \log k/a), & k > 1 \end{cases}$$

**Proof:** Consider the cdf of $X_{\max}$, which can be calculated by [22]

$$F_{X_{\max}}(t) = P\{X_{\max} \leq t\} = (1 - e^{-at})^k.$$

Thus, the pdf of $X_{\max}$ can be accordingly obtained by

$$f_{X_{\max}}(t) = \frac{dF_{X_{\max}}}{dt} = k(1 - e^{-at})^{k-1} \cdot ae^{-at}.$$
Then, the expectation of $X_{\text{max}}$ is

$$E\{X_{\text{max}}\} = \int_0^\infty k(1 - e^{-at})^{k-1} \cdot ae^{-at} \cdot t dt$$

$$= ka \int_0^\infty \sum_{i=0}^{k-1} \binom{k-1}{i} (-1)^i e^{-a(i+1)t} \cdot t dt$$

$$= \sum_{i=0}^{k-1} ka \binom{k-1}{i} (-1)^i \frac{1}{a(i+1)^2}$$

$$= \sum_{i=1}^{k} ka \binom{k-1}{i-1} (-1)^{i-1} \frac{1}{a^2i^2}$$

$$= \frac{k}{a} \sum_{i=1}^{k} \frac{(-1)^{i-1}}{i^2} \binom{k-1}{i-1}$$

$$= \frac{1}{a} \sum_{i=1}^{k} \frac{(-1)^{i-1}}{i} \binom{k}{i}.$$  

(21)

Then according to Lemma 8, we conclude this lemma.  
