ABSTRACT

LI, YUJIN. Mobility and Traffic Correlations in Device-to-Device (D2D) Communication Networks. (Under the direction of Wenye Wang.)

The development of wireless networking technologies has brought tremendous changes to people's everyday life, such as cellular and WiFi Internet access. It empowers users to communicate through different wireless access technologies, and also enables new applications in new paradigm, such as vehicular ad hoc networks (VANETs) and mobile cloud computing (MCC). Wireless communication can be achieved through not only infrastructure wireless networks (i.e., cellular and WiFi networks) but also opportunistic device-to-device (D2D) communications. One of the major features of D2D communication networks is user mobility, which affects network connectivity and the design of network protocols. Another important feature of D2D communication networks is complex traffic flow due to dynamic network topology and the double communication opportunities (i.e., infrastructure or infrastructureless wireless networks).

In this dissertation, we aim at understanding the mobility and traffic correlation in D2D communication networks, especially the emerging wireless applications such as VANETs and MCC. We first characterize node mobility and mobility correlation among users so that we can identify the autonomous ad hoc networks. Based on observations from real mobility traces, we define a metric, namely Dual-Locality Ratio, to quantify mobility correlation and evaluate group structures. As correlated mobility leads to presence of groups in which nodes have unequal abilities to relay data to other parts of the network, we further study how the information propagates in VANETs, which have highly dynamic and correlated vehicle mobility due to road layout and speed limit. We derive the farthest distance that message dissemination reaches at time t and the first time that message reaches distance d from the original source location under different dissemination strategies. Our analytical bounds provide not only spatial and temporal limits of message dissemination but also guidelines for design of message dissemination algorithms. Recently, D2D communication network is also used to accommodate mobile cloud computing. Besides access remote cloud through cellular or WiFi networks, users can employ nearby mobile devices for mobile cloudlet computing. In order to find out whether the traffic of computation offloading goes to the remote cloud or mobile cloudlet, we address the issue of whether/when mobile cloudlet can provide mobile application services by investigating its properties and computing performance. Finally, we investigate the content delivery in the D2D communication networks such as to accommodate the explosive mobile traffic. We find out how likely D2D communications can deliver contents to mobile users through discovering content distribution in network. The work in this dissertation advances our understanding of mobility and traffic correlation and offers guidance into the design of D2D communication networks.

Mobility and Traffic Correlations in Device-to-Device (D2D) Communication Networks

by
Yujin Li

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Electrical Engineering

Raleigh, North Carolina

2014

APPROVED BY:

Michael Devetsikiotis	Do Young Eun
Min Kang	Wenye Wang Chair of Advisory Committee

DEDICATION

To my parents and family.

BIOGRAPHY

Yujin Li earned her B.S. and M.S. degrees both in Control Science and Engineering from Beijing Institute of Technology in 2007 and 2009, respectively. In August 2009, she started her Ph.D. study at North Carolina State University in the Department of Electrical and Computer Engineering. Her research focuses on modeling and analyzing the performance of emergency message dissemination in VANETs and mobile application offloading in mobile cloud computing.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor, Dr. Wenye Wang. I feel so fortunate to have the opportunity to work closely with Dr. Wenye Wang and learn from her research passion, vision, wisdom, and much more, the attitude toward life. I sincerely appreciate Dr. Wang's suggestions, encouragement and guidance during my Ph.D. study. I believe that the knowledge and skills that she has imparted to me are far beyond this dissertation.

I would like to thank my committee members, Dr. Michael Devetsikiotis, Dr. Eun, and Dr. Kang, for their valuable feedback and comments, which significantly improved the quality of my research. I am also grateful to all the professors that have taught me in and out of classes at NC State for imparting their knowledge, which allows me to be prepared for my future career.

I would also like to give my thanks to my fellow labmates for their help during my Ph.D. study: Ming Zhao, Yi Xu, Shawqi Kharbash, Lei Sun, Zhuo Lu, Chi Yi, Mohit Khanna, Xiang Lu, Haiyang Zheng, Mohit Shah, Huan Luo, Mingkui Wei, and Sigit Pambudi.

Finally, I would like to give my special thanks to my families and friends for their encouragement and support during my Ph.D. study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

The development of wireless networking technologies has brought tremendous changes to people's everyday life. For example, cellular network, one of the most large-scale wireless networks, can provide not only phone-call services, but also various data services, such as video chat, mobile web and gaming. At the same time, the WiFi networks have become more flexible and configurable to provide ubiquitous and high-speed wireless access to the Internet. Moreover, mobile ad hoc networks (MANETs) that do not rely on a pre-existing fixed infrastructure, such as a wired line backbone network or base stations, have received lots of attentions in the past decade. For examples, nearby mobile users can use Bluetooth for direct file transfer services with low power consumption; handheld devices can also form a autonomous ad hoc network to allow communication in disaster areas (i.e., pocket switched networking) when networking infrastructure is destroyed. Opportunistic communication between mobile devices underlaying infrastructure wireless networks depict the blueprint of ubiquitous device-to-device (D2D) communication networks.

In D2D communication networks, devices can communicate with each other through both infrastructure wireless networks (e.g., cellular and WiFi) and mobile ad hoc networks. D2D communications underlaying a cellular infrastructure [1] have been proposed to take advantage of the physical proximity of communicating devices, offloading the cellular system, increasing bit-rate, and improving cellular coverage and robustness to infrastructure failures. For instance, an iPhone can be set up as a personal hotspot so that nearby devices can access the Internet through the iPhone's cellular data connection. Vehicular Ad Hoc Networks (VANETs) can use vehicle-to-vehicle communications as well as vehicle-to-roadside unit communications to achieve safe and efficient driving environment.

One of the major features of D2D communication networks, in contrast to a traditional

public switched telephone network, is user mobility. In cellular network, when a user moves from one cell to another, the call in progress has to be handed off from one base station to another to ensure continuity of service [2]. Similarly, WiFi networks perform handovers between adjacent WiFi access points in order to follow users' movements and provide seamless continuing service. Moreover, intelligent mobility management techniques [3] are expected to achieve global roaming among various access technologies, whenever the mobile host is moving across different access network domains. Last but not least, the impact of mobility on the link and route lifetimes in MANETs is of major importance for the design of efficient MAC and network layer protocols [4]. When two nodes move outside of each other's transmission zone, link and path breakage requires routing protocols to discover and update route information, resulting in long transmission delay, high overhead and energy consumption, which can dramatically degrade network performance and deteriorate the utilization of valuable network resources.

Another important feature of D2D communication networks, in contrast to pure infrastructure or infrastructureless wireless networks, is complex traffic flow. Recent developments in communication technologies, implemented in the latest smart mobile devices (including Google Nexus, iPhone, and iPad), make bulk data transfers between users in proximity a reality. Peer-to-peer assisted forwarding through Bluetooth is demonstrated to be feasible to deliver information originally scheduled for transmission over the cellular networks so as to offload cellular traffic [5]. An individual user can make a probabilistic decision whether to download contents through cellular networks or opportunistic communications with other users [6]. Exploring the double opportunities of peer-to-peer communication and mobile device to infrastructure communication can improve user throughput [7] and offload cellular traffic [8], but may also lead to redundant traffic and waste of network resources. In order to overcome the challenges induced by node mobility and complex traffic flow in D2D communication networks, we are dedicated to study mobility and traffic characteristics of mobile users and how they affect network performance in this dissertation.

As user mobility greatly affects opportunistic communication between nodes, there have been many studies on mobility, including mobility modeling [9–11], mobility impact analysis [12], and mobility-assisted schemes design [13, 14]. Most of existing studies have assumed random uncorrelated node mobility or simple group mobility because of its simplicity for analysis and simulation. However, in real life, mobility of wireless devices, which are associated with mobile human (e.g., pedestrians or drivers), exhibits significant degree of correlation [15, 16] due to geographic constraints or social correlations of human beings. Actually, such correlated human mobility leads to node groups providing communication opportunities for information delivery. The correlated mobility may invalidate or undermine existing results and insights on mobility in wireless communication networks. Therefore, we examine mobility correlation between users so that we can identify node groups (i.e., autonomous ad hoc networks). We characterize node

mobility using real traces and quantify mobility correlation in both spatial and temporal locality domains. Moreover, we use quantified mobility correlation to identify on-the-fly group structures and group evolutions, assist network topology management and data forwarding.

Presence of groups due to correlated mobility means that nodes have unequal abilities to relay data to other parts of the network, thus affecting network traffic flows and information propagation. Message dissemination for D2D communication networks is one of the most important applications to distribute and share information among a group of mobile users. In particular, VANETs as part of the intelligent transportation system have received lots of attention not only from the academia but also from industry [17, 18]. Many VANET applications are heavily dependent on message dissemination under highly dynamic and correlated vehicle mobility. In order to improve road safety, traffic efficiency, and driving convenience, emergency or entertainment information, such as collision warning and advertising, is disseminated to users that are affected by the incident or interested in the information. Different from multicasting in other wireless communication networks that targets specific destinations, message dissemination in VANETs mainly targets nodes in specific geographic regions, which is referred to as *geocast*. For example, two cars involved in an accident initiate a broadcast message about this collision to inform all vehicles in a region that geographically surrounds the original location of the accident. Because of vehicle mobility, geocast in VANET faces challenges of *dynamic* destinations as well as intermittent connectivity. Our current understanding of dissemination latency and information propagation speed for pre-defined destination nodes [19–22] may no longer be valid for geocast in VANET. To bridge the gap of our understanding on correlated mobility and message dissemination, we will study the performance of geocast in 2-Dimensional (2D) VANET with realistic vehicle mobility.

Besides the goal of information distribution, D2D communication network is also used to accommodate a new computing paradigm—mobile cloud computing (MCC) [23, 24]. In MCC, *remote cloud* provides data storage and computing service while mobile devices are clients to access the services through wireless communication networks, mainly cellular network and WLAN (Wireless Local Area Network). In light of the increasing memory and computational power of mobile devices [25], a *peer-to-peer* MCC (referred to as *mobile cloudlet*) [26–29] is also proposed in order to speedup computing and conserve energy. When cellular or WiFi connections are unavailable or costly, clustered mobile devices can share their resources in order to compute a common task. Mobile cloudlet is appealing to users with correlated mobility that pursue a common goal in group activities, such as multimedia sharing for audience at an event and language translation for a group of tourists in a foreign country. In mobile cloud computing, users have the double opportunities of utilizing remote cloud and local mobile cloudlet, introducing both traffic flows between mobile devices and cloud servers and among peer devices. Consequently, we are motivated to investigate whether or when the traffic of

transferring computational tasks should go to cloud or mobile devices, i.e., whether and under what conditions mobile cloudlet is feasible for providing mobile application services.

At last, with the proliferation of mobile handheld devices and the explosion of mobile applications, mobile traffic volume, especially multimedia traffic, is increasing dramatically, which leads wireless communication networks to be likely overloaded [30]. Cisco forecasts that global mobile data traffic grew 70 percent and mobile video traffic exceeded 50 percent in 2012 [31]. How users can efficiently fetch and share contents emerges as an important issue such as to satisfy users' enormous demand with limited network resources. Users can directly fetch data from cloud servers through cellular/WiFi networks, and users can also cache content for content sharing among one another through peer-to-peer communications. Such multi-choice content fetching together with node mobility produce complicated contents distribution and mobile data traffic in D2D communication networks, which challenge the design of efficient content delivery mechanisms. Motivated by such emerging mobile traffic volume and its threats to overload current wireless communication networks, we are interested in finding the distribution of contents in a network to facilitate content delivery networks (CDN) based on D2D communications.

In summary, our study of the mobility and traffic correlations deepens our understanding on network topology, information propagation, mobile cloud computing, and content sharing. Our study greatly expands our knowledge on the challenges and benefits of mobility and traffic patterns in D2D communication networks, and meanwhile provides instrumental guidelines for the efficient design of D2D communication networks (e.g., dissemination strategy and content delivery mechanism).

1.2 Objectives

In line with the four problems identied above, we aim to achieve the following objectives in this Ph.D. study that contribute to understandings of mobility and traffic correlations in D2D communication networks for emerging applications (e.g., VANETs, MCC, and CDN).

1.2.1 Identifying Network Structures of MANETs

In order to understand node grouping for opportunistic communications in D2D communication networks, we examine how node mobility is correlated in both temporal and spatial domains using three real traces. Based on our observations, we define a metric, namely dual-locality ratio (DLR), to quantify mobility correlation among nodes. We use both simulations and real traces to validate that DLR characterizes mobility correlation and can be used to identify mobile group structures in dynamic network environments. We also investigate stability and evolutions of groups that are identified by DLR. We derive conditions for group stability and

network evolution. In addition, we incorporate DLR in assisting network topology management and data forwarding, which increase cluster stability and packet delivery ratio, respectively. Our objectives are summarized as follows.

- We will highlight the correlation of node mobility in both *spatial* locality and *temporal* locality by observing on-the-fly groups in real traces.
- Similarities of spatial and temporal localities of nodes are measured and a metric, DLR, will be defined to quantify mobility correlation.
- We will show that comparing with existing community detection metrics, DLR can effectively identify on-the-fly groups using simulations and real traces.
- DLR will also be used to evaluate group stability and evolutions, assist clustering and data forwarding.

1.2.2 Analyzing Performance of Message Dissemination in VANETs

Correlated user mobility leads to node groups, providing opportunities for information propagation by opportunistic communications. To study the performance of message dissemination in VANETs with highly dynamic vehicle mobility, we consider the constrained vehicle mobility model as movements of vehicles are limited by both geometric and speed limits. Because of the crucial impact of dissemination mechanisms on geocast performance, we define a general *L-copy* ($L \geq 1$) direction-invariant and geographic-assisted message dissemination in which one or multiple nodes are actively spreading the message at the same time and disseminators are chosen direction-invariantly or based on node's geographic information, respectively. In order to shroud dynamic destination nodes of geocast, we introduce message mobility that includes movements of active message carriers (i.e., disseminators) and jumps incurred by transmissions from proceeding disseminators to succeeding disseminators. Message mobility enables us to study message dissemination without specifying relay nodes on information propagation path. Based on message mobility, we derive lower and upper bounds for the farthest distance that active messages reach at time t (denoted as *dissemination distance* $|D(t)|$) and the first time that active messages reach distance d from the original source location (denoted as *hitting time* $\tau(d)$) under different message dissemination strategies. We validate our analytical bounds using simulation results of several well known dissemination algorithms. Guidelines are provided for message dissemination strategy design in two realistic scenarios of VANETs. Our objectives are summarized as follows.

- Message mobility is formulated in order to study message dissemination in VANETs with dynamic node mobility and intermittent connectivity. Message mobility empowers us to

focus on where the dissemination (i.e., information propagation traffic) is rather than who carries the message.

- We will derive lower and upper bounds for dissemination distance and hitting time using message mobility, providing temporal and spatial limits of message dissemination performance.

- We will use simulations to validate our analytical bounds, which combining with two real VANET applications provide guidelines on dissemination mechanism design.

1.2.3 Evaluating Feasibility of Mobile Cloudlet in MCC

In mobile cloud computing, mobile users can offload computational tasks to either remote cloud through infrastructure networks, or local mobile cloudlet that includes groups of nearby mobile devices. We start with studying the performance of mobile cloud computing when users access cloud servers through cloudlet infrastructure located at different community sites (e.g., computers at coffee shop and office). When mobile devices form mobile cloudlet and share their computing resources for a common task, we address the issue of when mobile cloudlet can provide mobile application services through investigating the properties of a mobile cloudlet with respect to *cloudlet size*, cloudlet node's *lifetime* and *reachable time*. Cloudlet size is defined as the number of mobile devices that a mobile initiator encounters within time τ. A cloudlet node's lifetime is from its first contact to its last contact with an initiator before the task expires, reflecting a node's maximum time to perform computation for an initiator. Reachable time is the total contact duration between a cloudlet node and the initiator within time τ, indicating the cloudlet node's connection likelihood for task dissemination and retrieval. These properties not only determine how much computing resource a mobile cloudlet can provide but also implicate how reliable a mobile cloudlet is. We use both traces and analysis to derive distribution of the mobile cloudlet size, expected lifetime, and expected reachable time. Based on the properties of mobile cloudlet, we further derive upper and lower bounds of computing capacity and long-term computing speed of mobile cloudlet, which can be used by an initiator to decide whether to execute a task in mobile cloudlet. Our objectives are as follows.

- When mobile users access the cloud through cloudlet, we will study the performance of cloudlet computing, such as cloudlet access probability and task success rate.

- We will examine the mobile cloudlet properties, which are determined by the contact and inter-contact between mobile devices, using contact-based mobility traces.

- We will derive the distribution of cloudlet size, the expectations of node's lifetime and reachable time based on mathematical analysis of the alternating contact and inter-contact

process between two mobile users.

- The upper and lower bounds of mobile cloudlet computing capacity and long-term computing speed are derived based on cloudlet size, the lifetime and reachable time of cloudlet nodes. These bounds are used by mobile users to decide whether to offload mobile applications to mobile cloudlet.

1.2.4 Assessing Content Sharing Opportunities through D2D Communications

As emerging mobile traffic volume threats to overload current wireless communication networks, it is important to efficiently delivery contents to users. In order to do so, we investigate where copies of a content are in a network, i.e., the distribution of contents in network hosts including cloud servers and mobile devices. Because content distribution is affected by characteristics of contents and user request patterns, we examine content distribution in 3 key dimensions: i) popularity and diversity of contents, ii) users' behaviors of content fetching, and iii) the storage of contents on servers using real YouTube video traces. Mathematical models are further developed to characterize distributions of contents among mobile devices as time varies. In addition, we analyze the performance of content sharing through D2D communications. Our objectives are as follows.

- Using real YouTube traces, we examine content popularity, user's request pattern, and content transmission time.
- Mathematical modeling and analysis are developed to capture content distribution, i.e., who carries a content during what time.
- Based on characteristics of content distribution, we analyze the performance of content sharing through D2D communication, such as hit rate.

1.3 Outline and Organization

The rest of this report is organized as follows. Chapter 2 presents our observations of correlated node mobility and a metric to measure mobility correlation, which is used for identifying node groups and network structure as well as evaluating link and group stability and assisting data forwarding. Chapter 3 presents our analysis on performance of message dissemination in intermittently connected VANETs, including upper and lower bounds on dissemination distance and hitting time. Chapter 4 presents how we extract and analyze mobile cloudlet properties to derive computing capacity and speed of mobile cloudlet. Chapter 5 presents our discovery of

practical opportunities that exist for content sharing through D2D communication. Finally, we conclude our research results and discuss the possible extension directions in Chapter 6.

Chapter 2

Measuring Mobility Correlation and Identifying Network Structures

Recent studies on mobility-assisted schemes for routing and topology control, as well as on mobility-induced link dynamics have presented significant finding on the properties of a *pair* of nodes (e.g., the inter-meeting time and link life time), or a *group* of nodes (e.g., network connectivity and partitions). In contrast to the study on the properties of a set of nodes rather than individuals, many works share a common ground with respect to node mobility, that is, *independent* mobility in multi-hop wireless networks. Nonetheless, in D2D communication networks, mobile devices installed on vehicles or held by human are not isolated, yet *dependent* to each other. For example, speed of a vehicle is influenced by its close-by vehicles and vehicles on the same road move at similar speeds. Therefore, the gap between our understanding of the impact of *independent mobility* and our interest in the properties of *correlated mobility*, along with the real systems altogether declare an interesting question: how we can *measure the inter-node mobility correlation* such as to uncover the node groups and network components, and explore their impact on link dynamics and network connectivity. Bear this question in mind, we first examine several traces and find that node mobility exhibits *spatial locality* and *temporal locality* correlations, which are closely related to node grouping. In order to study the properties of such *groups on-the-fly*, we introduce a new metric, *dual-locality ratio* (DLR), which quantifies mobility correlation of nodes. In light of taking spatial and temporal locality dimensions into account, the DLR can be used to effectively identify stable user groups and evaluate group stability and evolutions, which in turn can be used for network performance enhancement.

2.1 Motivation and Related Work

In contrast with the random mobility modeled by popular mobility models (such as random walk), the moving behaviors of mobile users usually follow some mobility patterns and exhibit significant degree of correlations, which leads to overlapping movement trajectories of nodes [32]. As observed in corporate/campus WLAN traces [15, 16], mobile users spend most of their time at their home locations, where nodes gathering yields connected components. Traces in [33] show that nodes belonging to one community have frequent contacts and long contact durations. The correlated user mobility (also called group mobility) clearly leads to node grouping on the road or at community sites where human perform tasks or social activities.

As a result, many research works have been elaborated upon the impact of group mobility. Particularly, simulations in [12] show that routing protocols (i.e., DSR, DSDV and AODV) achieve the highest throughput and the least overhead with RPGM comparing to Freeway and Manhattan models. Authors in paper [32] observe a significantly reduced packet delivery ratio when employing the realistic trace simulator to control mobility of nodes. Ciullo et al. [34] reveal that correlated node movements have huge impact on asymptotic throughput and delay, and can sometimes lead to better performance than the one achievable under independent node movements. Huang and Chen [35] exploit the group mobility (RPGM) in replica allocation scheme to improve the data accessibility.

Existing studies on correlated mobility rely heavily on group mobility models with simplified nodes grouping behaviors, such as RPGM [36] and Virtual Track model [37]. It is commonly assumed that nodes are partitioned into several groups beforehand and group memberships either never change (e.g., RPGM) or evolve according to certain stochastic process. For instance, the virtual track model [37] for vehicular network scenario binds nodes' group movements on edges in a graph and group split and merge only happen at vertices. On the other hand, paper [38] assumes that groups of nodes merge or split according to a Markov chain process.

Nevertheless, in a spontaneously deployed ad hoc network with no pre-configurations, mobile nodes have no prior knowledge about the mobility groups and their memberships. Moreover, rather than the simplified nodes grouping behaviors in existing group mobility models, node groups in reality evolve not only in various ways (such as growth, contraction, combination and split) but also according to complex mechanisms due to the autonomous human mobility and complicated social behaviors of mobile users. The dynamic movement behaviors of nodes mean that group mobility is not *a prior* knowledge in wireless networks, nor does group structures, which makes the insights and benefits of group mobility claimed by applications based on pre-defined group mobility and structure questionable.

In a spontaneously deployed adhoc network with no pre-configurations, mobile nodes have no prior knowledge about the mobility groups and their memberships. In other words, there is a

need to identify where and how nodes form ad hoc networks to enable D2D communication. Few works have been done on group identification in mobile network environments. B. Li et al [39] firstly pointed this problem that mobile users are unable to acquire information about group structure beforehand, and proposed a sequential clustering (SC) group identification algorithm based on speed similarity among nodes for predicting network partition. In paper [40], B. Li et al proposed two entropy metrics for mobility identification, namely speed entropy and relation entropy, which cluster nodes with same movement properties, such as speed level, into mobile groups. However, metrics in papers [39] and [40] fail to consider the underlying social dimension of mobile hosts, which may undermine their effectiveness in identifying meaningful and strongly connected node groups.

Recently, there has been a growing body of work on the detection of *dynamic* communities and their evolution, such as [41, 42] and references inside. Most of these studies are two-stage approaches that extract clusters at consecutive timestamps of the network and then identify evolution of communities by comparing the group structures at different times. all centralized algorithms which are only useful for offline data analysis on mobility traces. For self-organized mobile ad hoc networks, it's desirable for the mobile devices sensing and detecting their local community structures in real time instead of relying on a centralized server. Community detection mechanisms usually are centralized and assume a prior knowledge of connections among all pairs of nodes. However, in mobile networks, correlations between pairs of nodes are not prior-knowledge, and are dynamically changing due to nodes movements. It's desirable for the mobile devices sensing and detecting their local community structures in real time instead of relying on a centralized server. The first study of distributed community detection without assuming prior-knowledge of inter-node correlation was carried out by Hui et al. [43], who proposed three distributed algorithms, that categorize nodes into familiar sets and local communities based on their contact durations and number of contacts. These algorithms can satisfactorily approximate the centralized mechanisms, but may be unable to capture the dynamically evolving network group structures due to the essential hysteresis in contact information.

In summary, current group identification methods and community detection algorithms are insufficient to capture the dynamic node grouping due to correlated user mobility, which is under-explored in existing studies. Therefore, in this chapter, we try to identify dynamic node groups that are induced by correlated mobility of wireless users, which are dynamically changing and evolving due to the autonomous human mobility and social behaviors. We achieve this by describing mobility correlation between any two nodes through a novel metric, dual-locality ratio, which incorporates mobility correlations of nodes in both spatial and temporal domains. The effectiveness of DLR is evaluated using traces as well as simulations. In addition, we demonstrate how DLR can be leveraged to assist data forwarding and clustering. Our contributions are three-fold.

1. Using three real traces, we identify mobility correlation in both *spatial* locality (i.e., locations and speed) and *temporal* locality (i.e., mobility patterns over time). A new metric, DLR, is defined to measure mobility correlation by incorporating similarity measure of two nodes' movements in both spatial and temporal locality domains.

2. We validate the effectiveness of DLR via simulations and traces. We find that DLR can not only properly measure the spatial locality similarities among nodes, such as distance, relative speed and direction, but also effectively identify groups in which nodes have same mobility pattern (i.e., traveling route). Compared with clustering geographically nearby nodes into groups, groups identified by DLR are more stable and meaningful because DLR can capture the grouping movements, destination of mobile users, as well as their current spatial locality similarities.

3. Besides measuring mobility correlation and identifying group structures, we further show that DLR can be used to evaluate link and group stability, manage network topology and assist data forwarding. DLR is shown to be a good indicator of link lifetime and link stability. We apply DLR to rigorously analyze the group coherence degree and find out conditions for stable groups and their evolutions (e.g., node switching and group merging). Simulation results show that choosing relays based on node mobility correlation can increase packet delivery ratio of data forwarding, and compared with lowest-ID clustering algorithm [44], choosing a node that has high DLR with its neighbors as cluster-head can increase cluster stability (i.e., lifetime).

The rest of this chapter is organized as follows. In Section 2.2, we use three real traces to identify how node mobility are correlated in spatial and temporal domains. In Section 2.3, we present our main results of the DLR for measuring mobility correlation by quantifying mobility similarities in spatial and temporal domains. In Section 2.4, we show that DLR can identify stable and meaningful groups via simulations and traces. In Section 2.5, we utilize DLR to evaluate group stability and evolutions, assist data forwarding and clustering. Finally, we conclude in Section 2.6.

2.2 Observations of Correlated Mobility

In wireless networks, mobile devices are mostly carried by pedestrians, vehicles, actuators, and even animals. The mobility of these carriers, in contrast to random i.i.d. mobility model, is restricted by geographic surroundings and dependent to each other due to social interactions. In this section, we highlight the *spatial and temporal* locality in node mobility by observing on-the-fly groups in real traces.

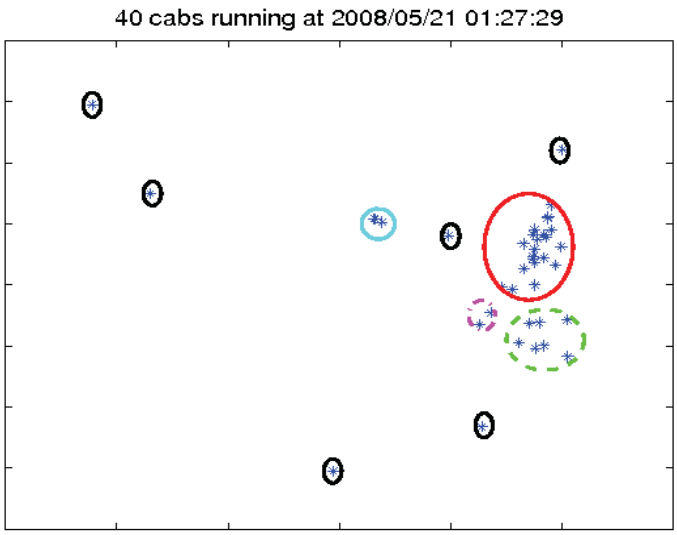

Figure 2.1: Geographical grouping in SFCAB.

2.2.1 Spatial Locality

To find out *whether there exist correlations among mobile users in real world*, we start with a taxicab trace in San Francisco (SF) from Cabspotting project [45]. This trace is chosen for our study because the customers may have quite different destinations and independent movements, which do not show obvious mobility patterns. The SF cab trace (SFCAB) contains GPS logs of 536 yellow cabs for over 30 days. Cab location is updated almost every minute if the cab stays online. We extract locations of 40 cabs running at 01:27 on 2008/05/21, shown in Figure 2.1.

Although cabs are likely moving independently since their destinations span a broad area, we still observe several cab clusters with different sizes, i.e., there exist mobility correlations among cabs. For instance, in Figure 2.1, there are two larger groups (in red and green) and two smaller groups (in blue and pink), along with six individual nodes in the space domain at a specific time. In the largest group (in red circle), 22 cabs locate in downtown San Francisco, which means these 22 users have similar *geographic* properties (i.e., adjacent locations and possible similar speeds). Such grouping phenomenon is probably caused by road layout, attraction of hot spots (such as airport, shopping center), and neighborhood distribution, thus demonstrating the *spatial locality* of user groups.

Remark 1 *The spatial locality of user groups observed in the cab trace above tells that mobile users are correlated in geographic dimension (e.g., locations and speeds). The more similar the spatial locality properties of nodes, such as adjacent locations and similar speeds, the higher their mobility correlations are and the more likely they form a group.*

2.2.2 Temporal Locality

In addition to the observation of spatial locality in node mobility, we turn to another trace, a bus trace in Seattle from Ad-Hoc City project [46], to explore the temporal locality incurred by daily moving schedules of mobile users. This trace is considered because the movements of buses are not independent as the cab-based trace, serving people who may share similar stops and timing patterns. The Seattle bus trace (STBUS) contains GPS logs of more than 1200 buses running 239 routes for around 20 days. Bus location is updated about every 2 minutes while the bus is running. Since there are frequent holes in STBUS trace, we observe 17 buses running 6 routes at 11:30 am on 2001/10/31, shown in Figure 2.2. The curves are trajectories of bus routes, and the points are current locations of buses.

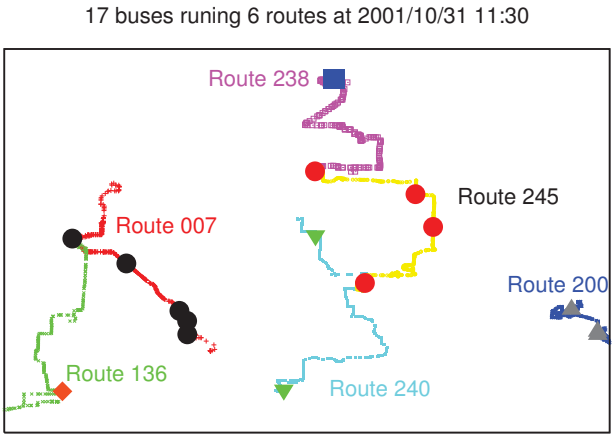

Figure 2.2: Social grouping in STBUS.

Figure 2.2 shows that over time buses running same or overlapping routes tend to meet more frequently (e.g., buses running red route 007). Clearly, buses on the same route have identical movement trajectory, thus likely meet frequently (e.g., the three black dots on the lower part of Route 007). In other words, buses can be grouped with respect to stops for different purposes, like shopping center, industry park, which are related to social behavior. Notice that buses running the same route do not always locate near to each other at specific times, e.g., the four buses (in red spots) on Route 245 are almost evenly distributed enroute. That reveals that the *temporal locality* in user groups is different from spatial locality of groups as it is induced by social behavior over time rather than geographic limitations. The correlations of the temporal locality properties of nodes are reflected in similar mobility patterns and overlapping moving routes, which are essentially due to the social correlations among mobile users.

Remark 2 *Besides the correlation of mobile users with geographic constraints, social behaviors or duties also have impact on mobility patterns and grouping of nodes, i.e., the temporal locality of node mobility. This adds another modality to explore the inter-node correlation.*

In order to properly interpret the temporal locality properties of vehicles, we further use a data set of students daily trace collected during a three-month period. The recorded trips include car trips and bus trips. Note that as the GPS receiver carried in a car requires a line of sight (LoS) from satellites, it cannot log short trips in a building complex area, which is a shortcoming for almost every trace file collected by GPS receiver. Since our trace has lots of detailed information of individually visited locations and driving paths, it is preferred over other large scale data sets, such as city wide transport system traces, which do not have such detailed information of each vehicle.

Figure 2.3 illustrates an example of one day car moving trace of a student. From the figure, there are total 4 trips during that day. Upon the time sequence, the student visits four places: home $\rightarrow$ lab $\rightarrow$ class $\rightarrow$ church $\rightarrow$ home. This moving trace shows that a vehicle changes its movement path over time because the driver has different destinations at different time in order to execute various social activities. For instance, in the morning, the vehicular node moves on the roads from home to lab (trip 1); during the daytime, it moves around campus (trip 2); and in the evening, it travels from church back to home (trip 4). As its driver targets different places to execute different activities, a vehicle accordingly changes movement paths, i.e., vehicle's location preferences are time-varying.

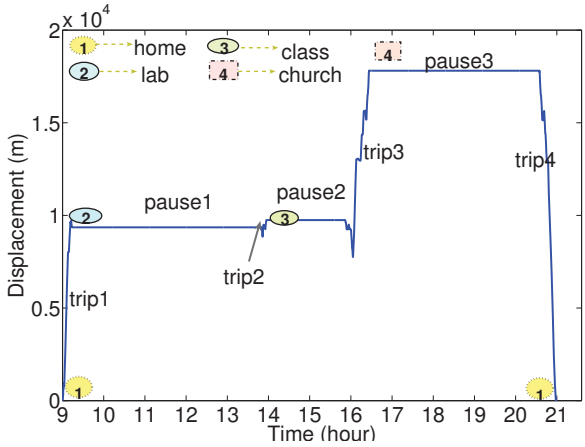

Figure 2.3: An example of one day car moving traces.

In addition, the aggregated locations visited by students within one week, shown in Figure

2.4, demonstrate that the majority community sites students daily visited are within 2-km-wide campus area. Vehicular nodes in this trace frequent several communities sites around campus and mostly move on roads among these places. In other words, vehicles frequently travel to preferred locations, i.e., vehicle mobility shows *temporal locality*.

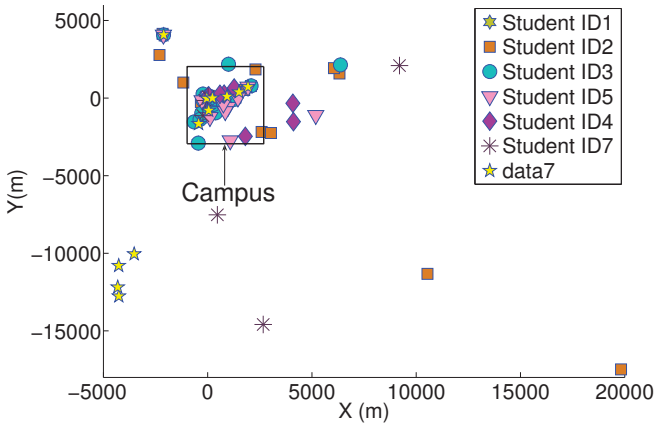

Figure 2.4: Aggregated visiting locations of students.

More interestingly, given the collected GPS traces, we have found that there are many overlapping trajectories among car trips of students. That's probably because similar social duties and life schedules of students induce similar temporal locality. This observation reveals that vehicles with similar temporal locality likely meet and have potential to move as groups.

Remark 3 *The vehicular mobility exhibits temporal locality, i.e., vehicles show different preferences to different community sites over time. Similarity in temporal locality of vehicles, i.e., preference of same/adjacent locations, leads to overlapping moving paths and grouping phenomena.*

2.2.3 Groups on-the-fly

Being aware of mobility correlations among nodes, group mobility models, such as RPGM [36] and its variances [47, 48], have been proposed to characterize such correlation by *preconfiguring* group membership of nodes and group movement behaviors. For instance, a group leader determines movement paths and moving speeds for its group members. To find out whether existing group mobility models can sufficiently capture inter-node mobility correlation, we compare two snapshots of STBUS trace to observe how network structures change over time.

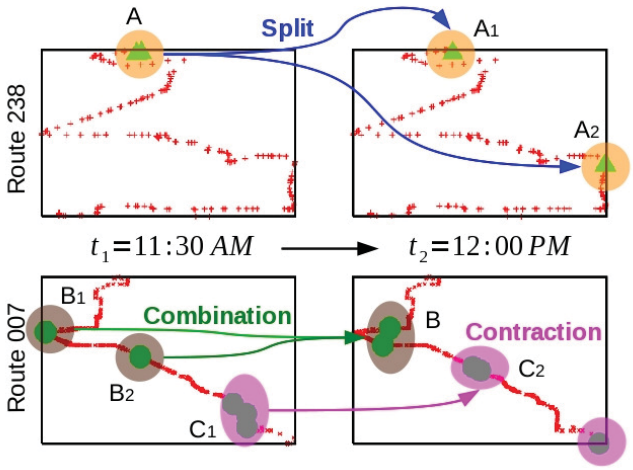

Figure 2.5: Group evolutions in STBUS: split, combination, and contraction.

Figure 2.5 shows the locations of 7 buses (2 buses on Route 238 and 5 buses on Route 007) at time t_1 and t_2, respectively. Group structures change dramatically as group A splits into A_1 and A_2 on the top, while groups B_1 and B_2 merge into a larger group B, and C_1 with 3 members contracts to a smaller group C_2 with 2 members on the bottom over the time period of 11:30am to 12:00pm. *On-the-fly* groups in Figure 2.5 implies that

- mobility correlations among nodes are *dynamically changing* due to autonomous mobility and social behaviors of mobile users [41];
- pre-configured group mobility (e.g., RPGM) *fails* to capture evolving correlations among nodes in real world (or non-mission oriented services).

Hence, a new metric is needed for measuring mobility correlation and exploring its impact on network properties, such as link dynamics and network connectivity.

In addition, it is necessary to point out that mobility correlations in *both* spatial and temporal dimensions have their implications on properties of a pair or a group of nodes. For instance, a group of nodes that locate near to each other and move closely (high spatial locality similarity), would form a stable group. In another example, a pair of nodes with high temporal locality similarity, having similar mobility patterns, are likely to meet frequently and stay together for a long period of time resulting in short inter-meeting time and long contact duration.

Remark 4 *In order to fully understand the properties of a set of nodes, i.e., group, it is desirable to capture* both *spatial and temporal locality similarities of node mobility.*

2.3 Mobility Correlation Measurement

Based on our observations of node mobility in real traces, we highlight the correlation of node mobility in both *spatial* locality and *temporal* locality, which are reflected by movement properties in spatial domain (location and speed) and in temporal domain (mobility pattern over time), respectively. Hence, we will proceed to measure the similarities of spatial and temporal localities of nodes in order to fully evaluate nodes' mobility correlation.

2.3.1 Spatial Locality Similarity

Intuitively, mobile users in a local group not only locate near to each other but also move more closely with similar speeds. Hence, the distance and relative speed between a pair of nodes reveal their similarity in spatial locality.

To proceed, let $d_{i,j}(t)$ and $v_{i,j}(t)$ be the Euclidean distance and relative velocity vector between node pair (n_i, n_j) at time t, As location and speed are two factors of the spatial locality of node mobility, the *distance* $d_{i,j}(t)$ and *relative speed* $v_{i,j}(t)$ provide information for the similarity of spatial locality between two nodes at time t. The spatial locality similarity measure $SLS_{i,j}(t)$ is time varying such as to capture dynamic mobility correlations. $SLS_{i,j}(t)$ should increase when either the distance $d_{i,j}(t)$ or relative speed $v_{i,j}(t)$ decreases. To make sure the spatial and temporal locality similarity measures have the same order of magnitude, we also require $SLS_{i,j}(t)$ to be normalized within $[0, 1]$.

With the statistical distance measure defined in statistical analysis [49], we are able to quantify the spatial locality similarity of two nodes with their relative movement information. Regarding location and speed as the two attributes of the 2-dimensional spatial locality observation for a node, the statistical distance between the two 2-dimensional observations of two nodes (i, j) at time t is $\sqrt{w_1 d_{i,j}^2(t) + w_2 v_{i,j}^2(t)}$, where $d_{i,j}(t)$ and $v_{i,j}(t)$ are their relative distance and speed, w_1 and w_2 are the weight coefficients for each attribute. Note that $v_{i,j}(t) = \sqrt{|v_i(t)|^2 + |v_j(t)|^2 - 2|v_i(t)||v_j(t)|cos\theta}$, where $|v_i|$ and $|v_j|$ are magnitudes of nodes v_i and v_j's speeds and θ is the angle between their moving directions. Let $r_{\max}$ denote the maximum transmission range and $v_{\max}$ denote the maximum speed. Because two neighboring nodes satisfy $0 \leq d_{i,j}(t) \leq r_{\max}$ and $0 \leq v_{i,j}(t) \leq 2v_{\max}$, we can set $w_1 = 1/(r_{\max})^2$ and $w_2 = 1/(2v_{\max})^2$ so that $d_{i,j}(t)$ and $v_{i,j}(t)$ can be scaled to have the same order of magnitude (i.e., be within range $[0, 1]$ for a neighboring node pair). Therefore,

Definition 1 *The spatial locality similarity (SLS) between a neighboring node pair (i, j) is defined as*

$$SLS_{i,j}(t) = \frac{1}{1 + \sqrt{\left(\frac{d_{i,j}(t)}{r_{max}}\right)^2 + \left(\frac{v_{i,j}(t)}{2v_{max}}\right)^2}}; \tag{2.1}$$

$SLS_{i,j}(t) = 0$ *for two non-neighboring nodes* i *and* j *as no communication link exists between them.*

Clearly, this SLS measure satisfies all our requirements. $SLS_{i,j}(t)$ is symmetric and $0 \leq SLS_{i,j}(t) \leq 1$. This definition also reflects the argument that nodes in a group move more closely than a random node pair [36]. When $d_{i,j}(t)$ and $v_{i,j}(t)$ decrease to 0, which means the two nodes move together at the same speed, like two people sitting in the same car, we have $SLS_{i,j}(t)$ approaches to 1 implying the highest spatial locality similarity. On the opposite, when the distance and the relative speed between these two nodes, $d_{i,j}(t)$ and $v_{i,j}(t)$ are large, then the value of $SLS_{i,j}(t)$ becomes very small, even to 0, which indicates very little or no spatial locality similarity between this pair of nodes. In other words, the more adjacent locations and similar speeds of nodes, the higher SLS they have, which indicates stable communication links and local groups.

2.3.2 Temporal Locality Similarity

Besides spatial locality similarity in the spatial domain, temporal locality similarity is also critical for understanding mobility correlation of nodes. To characterize degree of temporal locality similarity, we take an *entropy*-based approach to quantify similarity of mobility patterns. Specifically, we present a mathematical model, *cave profile,* to establish the mobility pattern of individual node with time-varying location preference. Finally, we use the idea of *relative entropy* to measure the similarity of different mobility patterns.

Cave profile modeling

Existing studies on mobility patterns of mobile users [15,16,50,51] have shown that, in contrast to the random trajectories predicted by the prevailing Levy flight and random walk models [52], movements of mobile users show a high degree of *temporal* and *spatial* dependence. More specifically, a mobile user spends *most of its time* at a *few locations*, while occasionally visits other places. Different mobile users not only have different probabilities of visiting the same community, and stay there for unequal amounts of time, but also may visit these sites at different times. In other words, a mobile user's mobility pattern is characterized by the *frequency* of its visits to each location, the *time* spent there, as well as the *order* in which the locations are visited [53]. Therefore, we model mobility pattern of a mobile user by characterizing the *tempo-spatial* dependence in its trajectory.

Assume that N nodes move in a network with M communities, which are referred to as *"caves"* because these nodes are tied to the communities. Let $\Omega = \{c_1, \ldots, c_M\}$ be the set of all possible community sites (e.g., home, office), where a node may visit frequently. Further we assume that mobile nodes can record its location, either by GPS or other localization methods

[54], like the traces we have used. Let $X_i(t)$ be the community site at which node n_i presents at time t, and $T_i(t) = \{X_i(1), X_i(2), \ldots, X_i(t)\}$ denote the historic sites visited, that is, n_i being present in these sites at each time interval Δt, which can be slotted time, e.g., $\Delta t = 10$ minutes over time. In fact, $T_i(t)$ generalizes the data in many traces, including the three traces we used in the earlier section. Therefore, $T_i(t)$ includes information of n_i's visiting *frequency*, *sojourn time*, and visiting *order* of community sites, thus captures mobility pattern in temporal and spatial domains.

We also notice that prior studies on human mobility [53] find that user mobility can be predicted with high accuracy ($\geq 93\%$) based on its movement trajectory, because human tend to repeat their daily schedules, such as working at office during daytime and staying at home in the evening. Therefore, based on mobility observation $T_i(t)$, we can estimate the probability of n_i appearing at cave c_m at next time slot $t+1$, denoted as $P_i^t(c_m)$.

Definition 2 *The* cave profile *of node n_i at time t is defined as $P_i(t) = \{P_i^t(c_1), P_i^t(c_2), \ldots, P_i^t(c_M)\}$, and*

$$P_i^t(c_m) = P(X_i^{t+1} = c_m | T_i^t) = \frac{N(T_i(t-k,t)c_m, T_i(t))}{N(T_i(t-k,t), T_i(t))}, \tag{2.2}$$

where $T_i(t-k,t) = \{X_i(t-k), \ldots, X_i(t)\}$ is the substring of $T_i(t)$, and $N(s', s)$ denotes the number of times that the substring s' occurs in s.

As $N(T_i(t-k,t), T_i(t))$ is proportional to the probability of finding a sequence of visited locations $T_i(t-k,t)$ over time, the probability $P_i(t)$ is able to capture both temporal and spatial dependence of mobility patterns. Since $\sum_{m=1}^{M} P_i^t(c_m)$ may not equal to 1, we rewrite the normalized cave profile as

$$\hat{P}_i^t(c_m) = \frac{P_i^t(c_m)}{\sum_{j=1}^{M} P_i^t(c_j)}. \tag{2.3}$$

Note that in Eq. (2.2), $P_i(t)$ depends only on the k most recent locations, thus we refer the above definition as $O(k)$ *cave profile*. If $N(T_i(t-k,t), T_i(t)) = 1$, to avoid $P_i^t(c_m) = 0, \forall c_m \in \Omega$, $O(k)$ cave profile degrades to $O(k-1)$. If $X_i(t)$ never occurs before, $P_i(t)$ degrades to be *temporal-uncorrelated* $O(0)$ cave profile.

$$P_i^t(c_m) = N(c_m, T_i(t)) / |T_i(t)|, \tag{2.4}$$

where $|T_i(t)|$ is the length of $T_i(t)$.

For the simplicity of analysis and practical applications, we need to find the value of k. To this end, we take the suggestion from an earlier work [55] that human mobility can be well predicted depending on the 2 most recent locations. Therefore, we use $O(2)$ cave profile throughout this chapter. That means,

$$P_i^{t+1}(c_m) = \frac{N(X_i(t-1)X_i(t)c_m, T_i(t))}{N(X_i(t-1)X_i(t), T_i(t))}. \tag{2.5}$$

With the above definition, we are able to model the temporal locality due to *individual* movement pattern. Next, we move on to the measurement of similarity between *different* mobility patterns.

Measuring Temporal Locality Similarity

Among various similarity/distance measures that compare two probability distributions (see survey paper [56]), *Kullback-Leibler Divergence* (KLD) (also relative *entropy*) is a well-known method of measuring the difference between two probability distributions in information theory. The KLD is well-defined for both discrete and continuous distributions, and is always non-negative. However, KLD is a non-symmetric measure, and is sensitive to quantization effects in the histogram computation. *Jenson-Shannon Divergence* (JSD), the symmetrized and smoothed version of KLD, is an empirically derived divergence that is numerically stable, and also robust in the presence of noise. Thus, we choose Jenson-Shannon Divergence to measure the similarity between mobility patterns:

Definition 3 *The JSD between $\hat{P}_i(t)$ and $\hat{P}_j(t)$ is defined as follows:*

$$JSD(\hat{P}_i(t)||\hat{P}_j(t)) = \frac{1}{2}\sum_{m=1}^{M} \hat{P}_i^t(c_m) log_2 \frac{2\hat{P}_i^t(c_m)}{\hat{P}_i^t(c_m) + \hat{P}_j^t(c_m)} + \frac{1}{2}\sum_{m=1}^{M} \hat{P}_j^t(c_m) log_2 \frac{2\hat{P}_j^t(c_m)}{\hat{P}_i^t(c_m) + \hat{P}_j^t(c_m)}, \tag{2.6}$$

The JSD measure of temporal locality similarity is defined as

$$TLS_{i,j}(t) = 1 - JSD(\hat{P}_i(t)||\hat{P}_j(t)). \tag{2.7}$$

$0 \leq TLS_{i,j}(t) \leq 1$ *and* $TLS_{i,j}(t) = 1$ *if and only if* $\hat{P}_i(t) = \hat{P}_j(t)$*, i.e., two users that have the same mobility pattern have the strongest temporal locality similarity* 1.

An illustrative example

To better understand the cave profile model and temporal locality similarity measure, let us take an example to observe the temporal locality similarities among four mobile users with five-site options, that is, $\mathcal{M} = \{1, 2, 3, 4, 5\}$. The location history for each user is shown in Table 2.1. The historic observations although are short, mimic the temporal and spatial dependence in mobility pattern. For instance, n_1 visits communities c_1 and c_2 much more often than the other three communities, while n_3 mostly stays at c_2 and c_5. In contrast to the sequence of site visits of n_2, n_1 goes to c_2 after staying at c_1.

By applying Eq. (2.5) for $O(2)$ cave profile, the probability of n_i visiting cave m after visiting caves c_3 and c_4 can be calculated, e.g., $P_1(1) = N(\{341\}, T_1)/N(\{34\}, T_1) = 1/2$ while $P_1(i) = N(\{34i\}, T_1)/N(\{34\}, T_1) = 0$ for $i = 2, 3, 4$. The resulting cave profile for each user is shown in Table 2.1.

Table 2.1: Example of User Cave Profiles

User	Location history T_i	Cave Profile
n_1	111122234111122234	{1/2, 0, 0, 0, 0}
n_2	222111134222111134	{0, 1/2, 0, 0, 0}
n_3	555522234555522234	{0, 0, 0, 0, 1/2}
n_4	111133334111133334	{1/2, 0, 0, 0, 0}

Accordingly, we can easily obtain the normalized $\hat{P}_i$ by Eq. (2.3). Using Eqs. (2.6) and (2.7), the temporal locality similarity between each user pair can be calculated, which is $TSL_{1,4} = TSL_{4,1} = 1$, and $TSL_{i,j} = 0$ otherwise. The results indicate that nodes n_1 and n_4 have high temporal locality similarity, thus likely move together as both of them probably will go to *Cave 1*.

It is interesting to see that even n_1 and n_2 have the same historical probability of visiting each cave according to Eq. (2.4), they are not likely moving together because they seldom appear at one location at the same time. In other words, without considering the temporal dependency in human mobility, $O(0)$ cave profile is inadequate for modeling mobility pattern.

Remark 5 *Based on mobility history, cave profile can be used to estimate the probability that each community is chosen as a user's next destination. By measuring the similarity of cave profiles, temporal locality similarity shows the likelihood of two users visiting the same cave during next time slot, thus telling the tendency of two users moving as a group.*

2.3.3 Dual-Locality Ratio

By far, we have investigated spatial and temporal locality similarities, both of which are essential in characterizing inter-node mobility correlation. From the perspective of a specific time, nodes are mainly correlated in spatial locality, e.g., cabs in SFCAB project, which affects link stability. From the perspective of a time period, nodes are mainly correlated in temporal locality, which influences contact-based properties. In order to characterize inter-node mobility correlation such as to adapt to complex node mobility as well as various network applications, we propose a new metric, namely *Dual-Locality Ratio* (DLR), by introducing a tune-up parameter α to jointly

consider above observations.

Definition 4 *The* Dual-Locality Ratio $DLR_{i,j}(t)$ *between two nodes* n_i, n_j *at time* t *is given by*

$$DLR_{i,j}(t) = (1-\alpha)SLS_{i,j}(t) + \alpha TLS_{i,j}(t), \tag{2.8}$$

where $0 \leq \alpha \leq 1$ *and* $0 \leq DLR_{i,j}(t) \leq 1$.

Through adjusting the value of α, the weights of spatial and temporal locality similarities can be adapted for different network scenarios. For networks that most nodes move independently, DLR with small α can represent the mobility correlation in spatial locality; while for nodes with clear mobility patterns, DLR with large α can manifest the similarity of nodes' temporal locality. DLR can also accommodate different applications by choosing different α. Large α (even $\alpha = 1$) is suitable for mobility pattern recognition or communication detection. On the other hand, small α (even $\alpha = 0$) is fit for link or path duration estimation. As we discuss groups on-the-fly in this chapter, we omit t for simplicity, e.g., $SLS_{i,j}(t)$, $TLS_{i,j}(t)$, and $DLR_{i,j}(t)$ are simplified as $SLS_{i,j}$, $TLS_{i,j}$, and $DLR_{i,j}$, respectively.

Remark 6 *Although the above definition of DLR measures the mobility correlation between a* pair *of nodes, it is shown in the following sections that DLR can be used to study properties of a* group *of nodes, such as group structure, stability, and evolution, as well as properties of a pair of nodes (e.g., link lifetime).*

2.4 Group Identification

A *group* means a number of nodes bounded together as being related in some way.

Definition 5 *Let* DLR_{th} *be the required* grouping threshold *for two users to belong to a group. Two neighboring nodes* n_i, n_j *are in the same group, if* $DLR_{i,j} \geq DLR_{th}$.

According to this definition, nodes can identify whether encountered nodes belong to a group. Each node v_i first obtain information of cave files, locations, and speeds from its neighbors in order to calculate DLR. When $DLR_{i,j}$ exceeds DLR_{th}, v_i will consider node v_j as its group member. Through further exchanging each other's group member information among encountered nodes, group structures can be uncovered.

The main overhead of computing DLR is due to exchange of mobility information between neighboring mobile nodes. Two types of messages are generated and periodically broadcasted by nodes to update 1) spatial locality information and 2) temporal locality information. The spatial locality message includes node's location, speed, and moving direction. The temporal locality message includes node's cave profile (i.e., caves the nodes visited over the past t time).

Assume that the sizes of spatial locality message and temporal locality message are S_s and S_t, and the corresponding broadcast frequencies are f_s and f_t, respectively. The overhead at each node is $S_s f_s + S_t f_t$. Clearly, spatial locality message has small S_s. The spatial locality update frequency is related to relative velocity and transmission range r of nodes [57]. f_s can be set as $r/2v_{\max}$ so that the link establish or break can be captured. For example, f_s approximately equals to 1 message per 4 second when $r = 250$ m and $v_{\max} = 30m/s$. On the other hand, f_t can be set as 1 message per 10 minutes, which is sufficient to capture the user's visit to different communities. Temporal locality message only needs to record visited caves over 24 hours as human tend to repeat their daily movement schedules. The total overhead in the network is $(S_s f_s + S_t f_t)N$, where N is the number of nodes in the network.

2.4.1 Trace Evaluation

In order to use DLR to identify groups, spatial and temporal locality properties need to be extracted from traces. Since locations of nodes are logged in both datasets SFCAB and STBUS, location and speed can be easily obtained. Hence, we focus on how we extract cave profiles for calculating temporal locality similarity, and the results of group identification.

Group identification in SFCAB

First of all, we investigate whether there is a traceable mobility pattern in cab mobility. Since SFCAB records whether a cab is carrying customers or not, we can extract locations where cabs pick up or drop off customers, i.e., locations of *stops*. It is worthy of noting that the customers are autonomous and take cabs without coordination, which means that there is no correlations among their destinations. Surprisingly, the stops of 3 cabs shown in Figure 2.6 reveal spatial dependency, i.e., locations of stops are not uniformly distributed in the city area. For instance, *Cab 1* visits the western area more often than the eastern part of the city. *Cab 2*, in contrast, prefers the eastern area. This is because cab drivers prefer working in different areas or hot spots, such as airport or downtown. In other words, cabs exhibit different mobility patterns.

To measure the similarity of mobility patterns between two cabs, we then identify hot spots, or "caves", where people frequently get on or off. The stops of 100 cabs are clustered to 5 caves (hot spots) through *k-means clustering* algorithm in MATLAB, shown in Figure 2.7. Stop locations of one cab in Figure 2.7 show that this cab visits caves 2, 4, 5 more frequently while occasionally visiting caves 1 and 3. For simplicity, we use $O(0)$ cave profile, i.e., the probability that a cab stays in each cave. The probability of car n_i at cave c_m can be obtained by

$$P_i(c_m) = \frac{\text{number of stops in cave } c_m}{\text{total number of stops of car } n_i}. \tag{2.9}$$

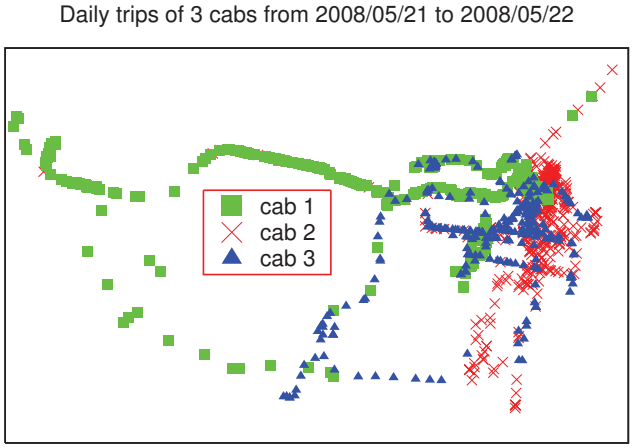

Figure 2.6: SFCAB: temporal and spatial dependency.

By parsing the data in SFCAB trace file, we have identified three groups (1, 4, 6) with multiple members, and other 9 groups with single node only as shown in Figure 2.8. The length and direction of an arrow represent the speed and moving direction of a node, respectively; the dotted line between n_i and n_j means that $DLR_{i,j} \geq DLR_{th}$, where $DLR_{th} = 0.2$ is the average DLRs of all node pairs. By taking a close look, we further observe that node 5 does not join *Group* 1 because node 5 is not within the transmission range of any node in this group. Also nodes 2 and 3 cannot be clustered to a group because they move at different speeds, e.g., the arrow length for node 3 is much longer (faster) than that of node 2. Nodes 9 and 10 are classified as two groups because of their opposite moving directions.

When we change $0 < \alpha < 1$ with different values, that means changing the weights of temporal and spatial locality similarities, we have not observed much differences in identified groups as shown in Figure 2.8. This implies that α has little impact on group structures of cabs. One possible reason is that nodes in SFCAB mostly move independently. Thus, temporal locality similarities among nodes are at the same level and spatial locality similarity dominates mobility correlation and group identification in SFCAB. In other words, temporal locality has little impact on group structures of nodes with homogeneous mobility.

Remark 7 *DLR can properly measure the spatial locality similarities among nodes, such as distance, relative speed and direction. Figure 2.8 shows that DLR can effectively identify groups in which nodes have similar mobility features (such as location, velocity), have more connections among them than connections with rest of the network, and form connected components.*

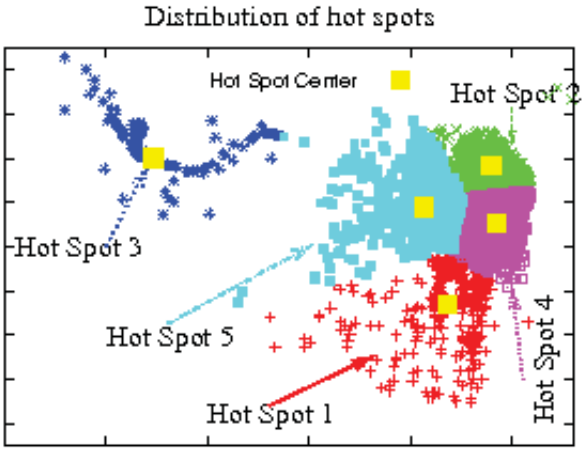

Figure 2.7: SFCAB: "Caves" are visible.

Group identification in STBUS

Assume that each bus stop is a "cave", and Ω is the set of all bus stops. For simplicity of analysis, we consider temporal-uncorrelated cave profile $P_i = \{P_i(c_1), \ldots, P_i(c_M)\}$, where $P_i(c_m)$ is the probability that bus running route i appears at bus stop c_m. For c_m on route i, $P_i(c_m)$ can be approximated by l_m/L_i, where l_m is the length between c_m and its previous stop, and L_i is the length of route i; otherwise, $P_i(c_m) = 0$. Using Eqs. (2.6) and (2.7), the temporal locality similarity between two buses running routes i and j can be simplified as:

$$TLS_{i,j} = 1 - \frac{1}{2} log_2\left(\frac{L_i - L_{i,j}}{L_i} + \frac{L_j - L_{i,j}}{L_j}\right), \tag{2.10}$$

where $L_{i,j} = L_i \cap L_j$ is the overlapping length of two routes.

An example of two buses is illustrated in Figure 2.9. Note that for two buses running on the same route, $TLS_{i,j}$ is 1, i.e., they have same mobility pattern; for buses running different routes, $TLS_{i,j}$ depends on the proportion of overlapping trajectory over total route length. The more overlapping between two routes (i.e., similar mobility patterns), the higher their temporal locality similarity is.

Using DLR in Eq. (2.8) and setting DLR_{th} as average of DLR over all neighboring node pairs, group structure in STBUS is shown in Figure 2.10. Groups in dashed rectangles are obtained by using $\alpha = 1$, which are consistent with bus routes. Based only on spatial locality similarity ($\alpha = 0$), buses belonging to different routes may be clustered as a group (the small solid square in the middle). By considering both spatial locality and temporal locality similarities, e.g., $\alpha = 0.5$, identified groups (the dashed circles) only include buses running same route and moving closely.

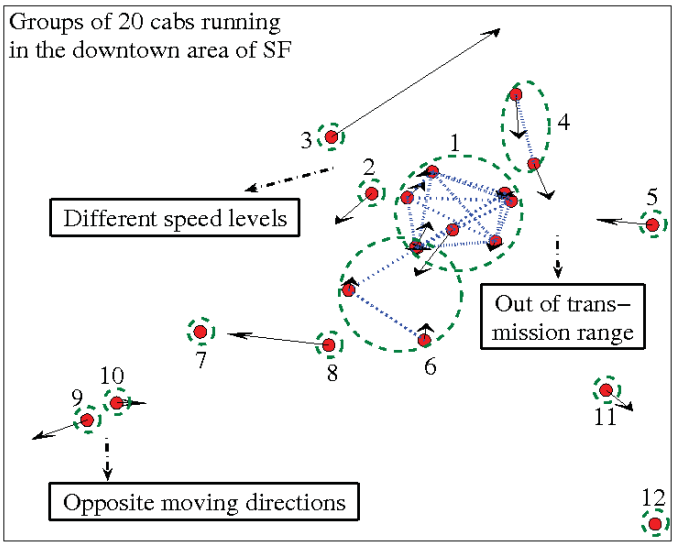

Figure 2.8: SFCAB: Identified groups with $\alpha = 0.5$ and $DLR_{th} = 0.2$.

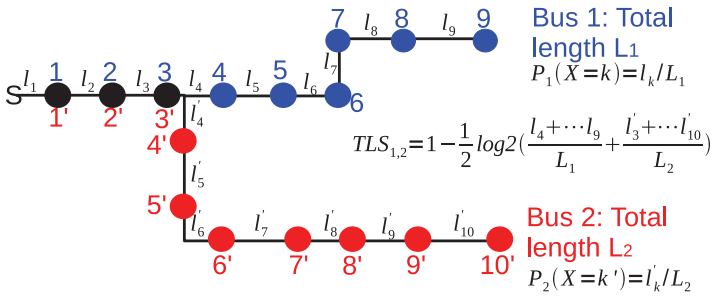

Figure 2.9: Temporal locality similarity between two buses.

Remark 8 *DLR can effectively detect temporal locality similarity in user groups in STBUS, which implies a good level of measurement of temporal locality similarities among mobile nodes. Figure 2.8 shows that DLR can effectively identify groups in which nodes have same mobility pattern (i.e., traveling route), and have more interactions among them than interactions running on other routes.*

2.4.2 Simulation Evaluation

By examining real traces, we have observed how DLR can be used in identifying groups with either spatial or temporal locality similarity. To the best of our knowledge, there is no trace with both spatial and temporal locality information available for public use, we resort to simulations

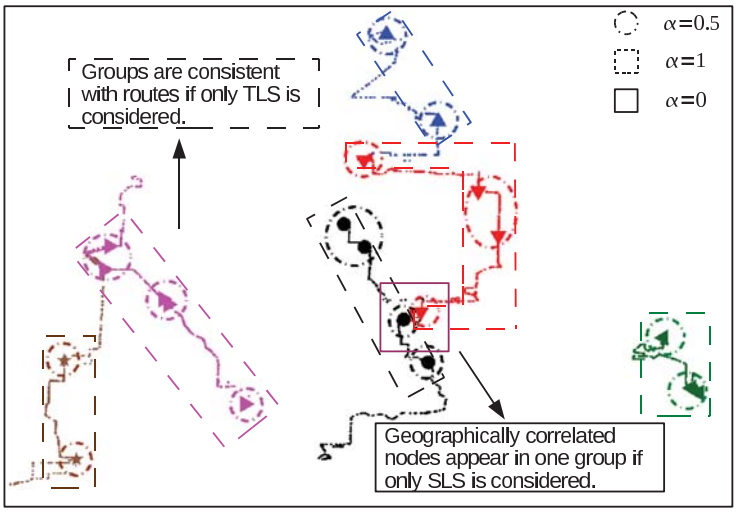

Figure 2.10: STBUS: Identified groups with $\alpha = 0, 0.5,$ and 1.

for a thorough evaluation under realistic human mobility as well as random mobility.

Time-Space varying Caveman Mobility

Besides the commonly used RWP mobility model, we implement a *Time-Space varying Caveman (TSC) mobility model* in OMNeT++ and INET-Framework [58]. We use TSC to reproduce the time-space dependency of human mobility.

Let $W_i = \{w_i(c_1), w_i(c_2) \dots w_i(c_M)\}$ denote the waiting times of node n_i at M caves, By generating pause times $w_i(c_1), \dots, w_i(c_M)$ according to truncated power-law (TPL) distribution, few of them can be much larger than others, i.e., people spend most of their time at few caves and stay shortly at other caves. Accordingly, let nodes stay in different caves with certain *weights/probabilities*. Node n_i's preference of cave c_m, or probability of staying at c_m, is given by:

$$P_i(c_m) = \frac{w_i(c_m)}{\sum_{j=1}^{M} w_i(c_j)}. \tag{2.11}$$

The above definition enables hierarchical location preferences, which can vary over time by generating pause times $W_i(t)$.

An example with 5 communities in the network is shown in Figure 2.11. Vehicle n_i at home location generates the sojourn time in community site c_m as $\omega_i^{c_m}(t)$. Using Eq. (2.11), location profile $\{P_i^{c_m}(t), 1 \le m \le 5\}$ at time t is obtained. Then, each node selects one of the M communities as its next target with probability $P_i^{c_m}(t), 1 \le i \le N, 1 \le m \le M$, and randomly chooses a destination point around c_m. The node moves to its destination using

smooth movement that first speeds up, then moves at stable speed, finally slows down before coming to a stop (see Figure 2.11). The speed of smooth movement is proportional to the distance between starting point and the destination [59]. When node n_i reaches its destination at c_m, it stays there for $\omega_i^{c_m}(t)$ period of time with small movements around c_m or short pauses. Then, node n_i repeats this process again. All N nodes in the network continue their movements until the end of simulation.

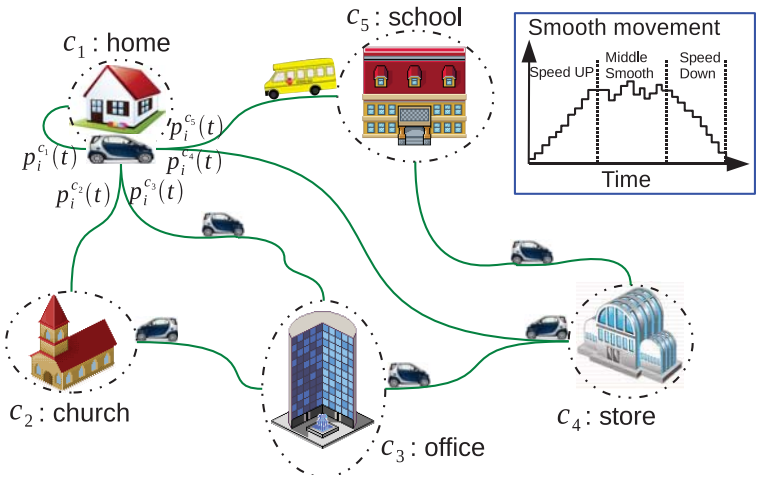

Figure 2.11: TSC mobility: temporal locality modeling and smooth movement.

Before using time-space varying caveman (TSC) mobility for VANETs detection, we run simulations to make sure that TSC mobility exhibits empirically observed truncated power-law decay of inter-contact time [60]. Simulation runs for 24 hours with 20 nodes moving in a 5-community network area as shown in Figure 2.11. The transmission range is set as 250 meters. Figure 2.12 shows the complementary CDF (CCDF) of the inter-contact time, i.e., $\mathcal{P}\{T_I > t\}$, on a log-log scale with simulation area ranging from $1000m \times 1000m$ to $5000m \times 5000m$. Consistent with the studies in [60], the inter-contact time follows a truncated power-law distribution and for the $5000m \times 5000m$ simulation area, the power-law behavior is dominant over up to $O(10^4)$ seconds, followed by a sharp decrease beyond that timescale.

Remark 9 *TSC mobility model not only mimics time-space varying human mobility but also yields power law and exponential decay dichotomy of inter-contact time. Therefore, we use TSC to evaluate DLR.*

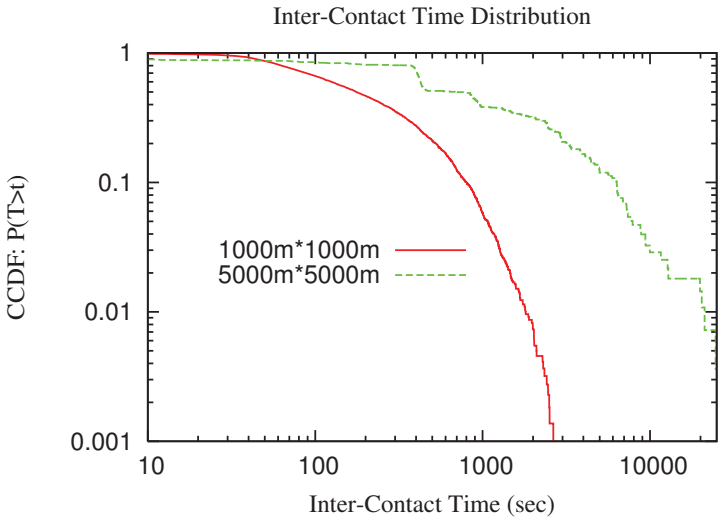

Figure 2.12: Inter-contact time under TSC mobility.

Group identification under TSC mobility

Here we first study group identification in a simple 4-cave network, then in a general network. As a simple scenario, the network is partitioned into 4 caves $\{C1, C2, C3, C4\}$, with 10 nodes moving for 48 hours according to TSC mobility. Nodes update their mobility observations every 10 minutes. We introduce two mobility patterns: $O(0)$ cave profile for $\{n_1, \ldots, n_5\}$ is $\{0.85, 0.05, 0.05, 0.05\}$, while for $\{n_6, \ldots, n_{10}\}$ is $\{0.05, 0.05, 0.85, 0.05\}$.

Figure 2.13 shows that DLR identifies two groups, in which $\alpha = 0.5$ and $DLR_{th} = 0.5$ are used. *Group* 1 in $C1$ includes nodes $\{n_1, n_2, n_3, n_4, n_7\}$ and *group* 2 in $C3$ includes nodes $\{n_5, n_6, n_8, n_9\}$. Notice that node n_7 is loosely connected to other nodes in *group* 1 because it has very different mobility pattern (i.e., cave file) from nodes in *group* 1, e.g., it has a higher preference of *cave* $C3$, while the nodes in *group* 1 spend most of their time in *cave* $C1$. The connection between node n_7 and other nodes in *group* 1 is due to spatial locality similarity (i.e., adjacent locations). Similarly, node n_5 is connected to n_8 because they locate closely, disconnected to nodes n_6 and n_9 in *group* 2 due to different mobility patterns.

Remark 10 *DLR can identify user groups in which nodes either have* similar *mobility patterns or are closely located.*

In spite of the effectiveness of DLR, there is still one concern: why not just borrow community detection algorithms from social network (see review papers [61–63]) for group identification in mobile networks? By an extensive study of community detection algorithms in the literature, we find that 1) most of these algorithms are *centralized*, thus are not applicable to *self-organized*

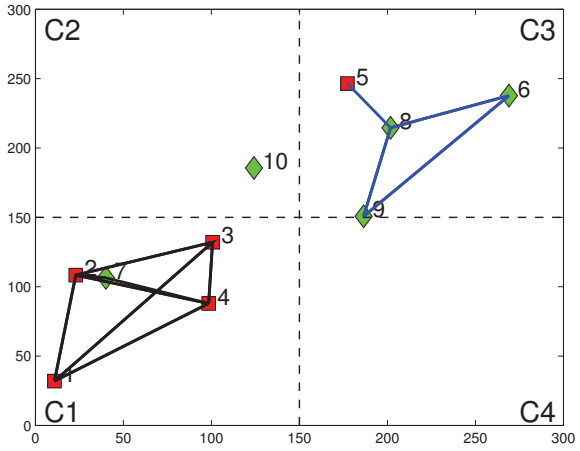

Figure 2.13: Nodes with similar mobility patterns are identified as groups.

multihop wireless networks; and 2) most of them assume a *prior* knowledge of correlations among nodes, which, however, is not true in mobile networks. Some works [43] present an effort in overcoming these problems by proposing distributed community detection algorithms in mobile networks, which is shown to be unsuitable for group identification in mobile networks (i.e., there is a difference between community detection and group identification.) in the following.

In a more general network with 25-cave and 20-node under TSC mobility, we compare DLR with an community detection algorithm, SIMPLE, in [43]. SIMPLE works as follows: node n_i inserts an encountered node to its familiar set if their contact duration exceeds a certain threshold, and n_i adds an encountered node n_j to its community set if number of common nodes in their familiar sets over total number of n_j's familiar members is larger than a merging threshold.

By carefully choosing familiar and merging thresholds for SIMPLE, which have significant impact on the results, detected communities by SIMPLE are shown in a snapshot of nodes' positions in Figure 2.14. In this illustration, an arrow from node n_i to n_j means that n_i considers n_j as a community member. Although SIMPLE may reveal long-term community structures, it tells little about current user groups. In other words, this algorithm cannot capture the dynamics of network structure, which is an intrinsic and unique feature in wireless networks, like MANETs, VANETs, and DTNs.

By setting $\alpha = 0.8$ and $DLR_{th} = 0.75$, user groups identified by DLR are shown in Figure 2.15. Comparing Figures 2.14 and 2.15, we observe: i) when nodes from same community meet, they are likely to move as a group. For example, nodes 10, 13, and 17 are connected in both Figures 2.14 and 2.15. ii) Nodes from different communities may occasionally move together,

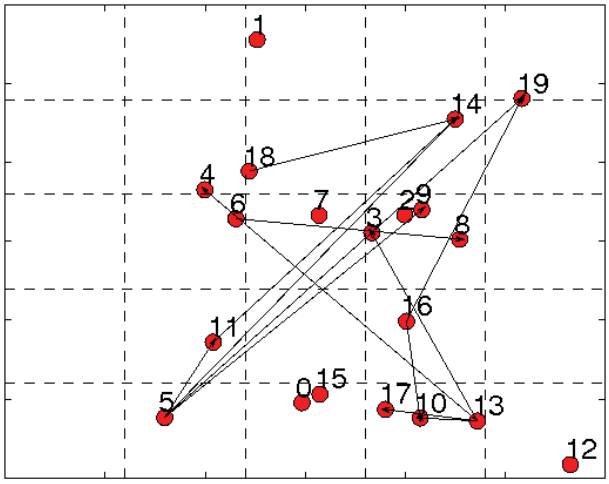

Figure 2.14: Communities detected by SIMPLE under TSC mobility.

such as nodes 0 and 15. iii) Nodes with long contact durations not necessarily form group all the time. For instance, nodes 5, 11, and 14 join a community in Figures 2.14, but they locate at different caves in Figure 2.15.

Remark 11 *DLR performs better than SIMPLE in identifying the* real-time *user groups in wireless networks since it can capture the grouping movements, destination of mobile users, as well as their current spatial locality similarities.*

Group identification under RWP mobility

In this part, we apply both DLR as well as SIMPLE to nodes moving according to RWP mobility model, as a case study, to find out whether they can identify groups or communities under random mobility.

Since nodes are homogeneous and move randomly in RWP model, contact duration between one pair of nodes is identical to another pair of nodes if the simulation runs for a long time. Through simulations, we find that choosing a large or small contact duration threshold for SIMPLE may lead to either completely partitioned or fully connected network, i.e., a node either has no community member or joins the giant community including almost all nodes in the network. As shown in Figure 2.16, SIMPLE is unable to identify communities for nodes under random mobility.

By measuring the mobility correlation, we find that temporal locality similarities among nodes are almost identical, which is consistent with the homogeneity of nodes mobility. Consequently, spatial locality similarities among nodes determine the group structures. In other

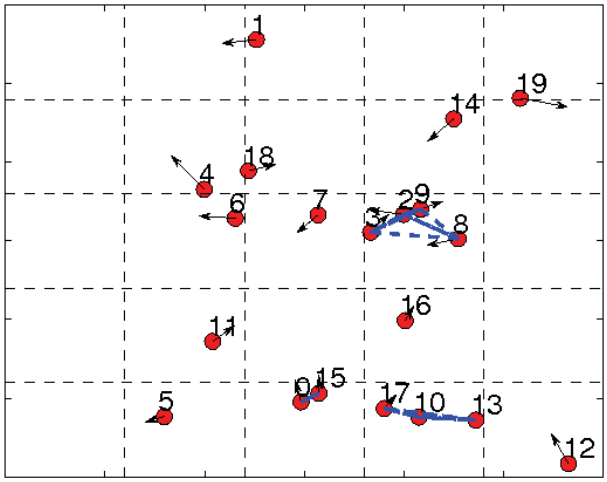

Figure 2.15: Groups identified by DLR under TSC mobility.

words, nodes are forming groups because they are located near to each other and move closely rather than similar mobility patterns. The user groups (clusters in green ovals) in Figure 2.17 show that DLR still catches the dynamic group structures of nodes under random mobility.

Remark 12 *When nodes in the network are homogeneous with approximately random mobility, community detection algorithms fail to identify meaningful group structures, while DLR can still identify groups in which nodes have similar movement features.*

Impact of α and DLR_{th}

This part provides guidelines on how to choose appropriate α and grouping threshold DLR_{th}.

One important parameter in DLR is α, which can adjust the weights of spatial and temporal locality similarities. On one hand, if mobile users have relatively stable movement habits, DLR with large α tends to detect long-term communities that mobile users have similar mobility patterns. Figure 2.18 shows that, by using larger α, perfect group identification is guaranteed for a wider range of DLR_{th}. Thus, *large* α is preferred for networks with well defined *communities*. On the other hand, DLR with high weight of spatial locality similarity tends to cluster nodes into closely moving groups. Without taking into account spatial locality similarity ($\alpha = 1$), the group structure may be unable to capture the dynamics of node movements. Therefore, for network with high or random node mobility, *smaller* α may be preferred in order to capture network *dynamics*.

Next we take a look at whether DLR is robust to changes of grouping threshold DLR_{th} in Definition 5. Let us define *false* identification as if nodes n_i and n_j have very *different* mobility patterns, $DLR_{i,j} > DLR_{th}$, or if n_i and n_j have *similar* mobility patterns and $X_i = X_j$,

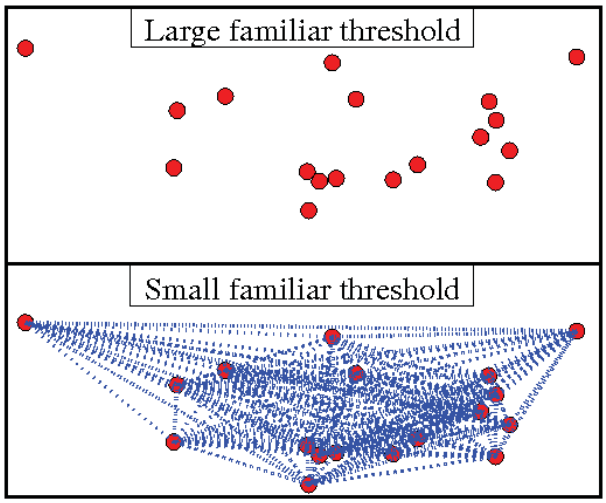

Figure 2.16: Communities identified by SIMPLE under RWP mobility.

$DLR_{i,j} < DLR_{th}$. The probability of correct identification in the 4-cave network is shown in Figure 2.18.

When $\alpha = 0.8$, we observe that 100% correct group identification holds for $0.58 \leq DLR_{th} < 0.81$. Therefore, for identifying groups in networks where nodes have different mobility patterns, DLR is *robust* to a wide range of DLR_{th}.

Although dual-locality ratio threshold DLR_{th} is an important parameter in identifying user groups, an universal DLR_{th} may be hard to determine due to the dependence of group structures on *node mobility* and *application requirements*. An easy way to set DLR_{th} is averaging DLRs over all node pairs, since nodes in groups are more closely related to each other. Another possible way to choose appropriate DLR_{th} is adaptively changing DLR_{th} through learning the stability of identified groups. If nodes move in or out the group frequently, increasing DLR_{th} can help to identify more stable groups. To find an suitable DLR_{th}, application requirements also need to be considered. For reliable applications, such as sending control messages, we suggest to use a large DLR_{th} to ensure the stability of links and group structures; otherwise, a small DLR_{th} is feasible.

2.5 Applications of Mobility Correlation

The previous section shows that correlations among nodes can be used to identify group structures. Following that, an interesting question is how stable the groups or the links among nodes are. In contrast to many prior work on link dynamics and network topology under *i.i.d.* mobility (e.g., [64]), we aim to find insights of *inter-node* mobility correlation on properties of a pair

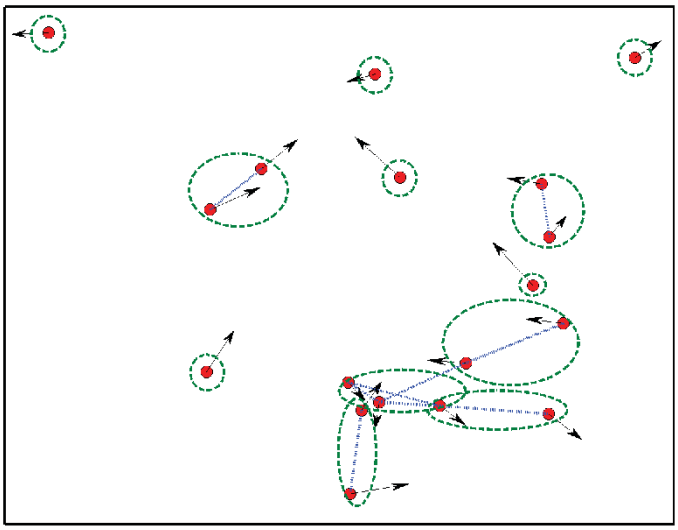

Figure 2.17: Groups identified by DLR under RWP mobility.

or group of nodes. We first investigate the relationship between inter-node mobility correlation and *link lifetime*. Then, we evaluate the *stability* of group structure, and find out the conditions for stable groups and their *evolutions*. Finally, DLR is applied to assist data forwarding and clustering algorithm.

2.5.1 Evaluating Link Lifetime

To begin with, we aim to find some insights on link lifetime, or alternatively contact time between a pair of nodes. Specifically, we examine the average contact durations of nodes that have different levels of DLRs. Five nodes move according to TSC mobility in a 6-cave network. The average speeds of nodes are 10m/s, 20m/s, and 30m/s, representing low, medium, and high mobility, respectively. The cave preferences of nodes are shown in Table 2.2, where n_0 is the reference node. Based on $O(0)$ cave profile, we can estimate temporal locality similarities between n_0 and n_i $(1 \leq i \leq 4)$ using Eqs. (2.6) and (2.7), which are $TLS_{0,1} = 1, TLS_{0,2} = 0.75, TLS_{0,3} = 0.5, TLS_{0,4} = 0.25$. Assume that two nodes are in the transmission range of each other or in contact if and only if they locate at the same cave. Simulation collects the contact durations and number of contacts between n_0 and n_i $(1 \leq i \leq 4)$, and calculates their average contact durations under low, medium, and high mobility scenarios, respectively.

The results in Figure 2.19 are in agreement with our intuition that two nodes with high temporal locality similarity have similar movement schedules, thus being able to maintain long contact duration or link lifetime. Figure 2.19 shows that, regardless of mobility intensity (low, medium, or high), average contact duration increases approximately linearly as temporal locality

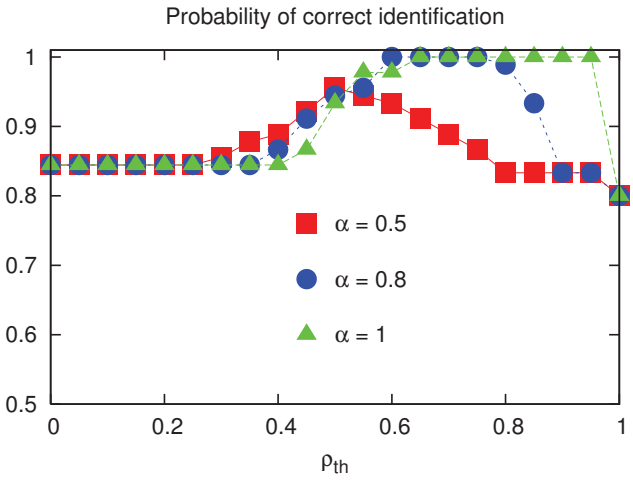

Figure 2.18: Probability of correct group identification.

Table 2.2: $O(0)$ cave profiles of 5 nodes in a 6-cave network

	c_0	c_1	c_2	c_3	c_4	c_5
n_0	0.25	0.25	0.25	0.25	0	0
n_1	0.25	0.25	0.25	0.25	0	0
n_2	0.25	0.25	0.25	0	0.25	0
n_3	0.25	0.25	0	0	0.5	0
n_4	0.25	0	0	0	0	0.75

similarity increases. Moreover, more closely two nodes move together, more stable the link between them is, i.e., spatial locality similarity implies link stability. Therefore, DLR could be a good indicator of link lifetime and link stability, which can help us to establish reliable routes in routing protocol.

2.5.2 Evaluating Stability and Evolution of Groups

In mobile networks, the groups may change remarkably over time due to the dynamic movements of mobile users. Thus, group stability is essential to characterize the correlation degree of group mobility and predict group evolution [41,42], which has an immediate impact on routing and system performance in multi-hop wireless networks. Since DLR has direct impact on link stability and link lifetime, as discussed earlier, we further explore how it affects network connectivity. In what follows, we first apply DLR to rigorously analyze the group coherence degree, and then find out conditions for stable groups and their evolutions.

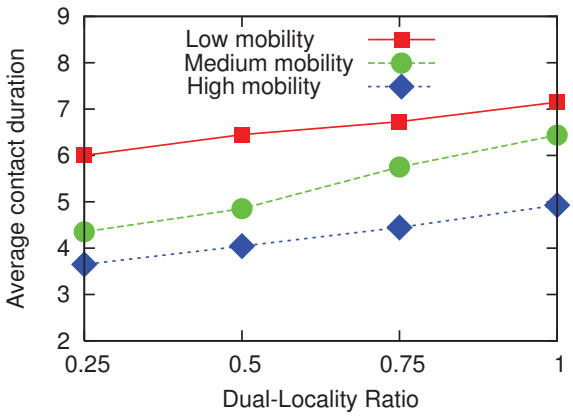

Figure 2.19: Link lifetime or contact duration versus DLR.

Group cohesiveness

Group structure in a network has many equivalent features to what a cluster structure has in a graph. To properly describe and specify group structure, we consider graph G consisting of group member set $V(G)$, which has the associated link connection set $E(G)$. Let $h(i,j)$ denote the i,j-path with the minimum hop length (number of hops). If node n_i is a neighbor of node n_j, $h(i,j)=1$, i.e., $h(i,j)=1$ is equivalent to $e_{i,j} \in E$. Let N_i be node n_i's neighbors set and Δ_i be the degree of n_i, then $N_i = \{n_j | e_{i,j} \in E\}$ and $\Delta_i = |N_i|$.

In graph theory, *local clustering coefficient* has been widely used to measure the local group cohesiveness. Basically, the local clustering coefficient C_i of a vertex n_i in a graph quantifies how close n_i and its neighbors are to form a complete graph [65]. Given a group $G=(V,E)$, the local clustering coefficient of n_i is

$$C_i = \frac{2|\{e_{j,k}\}|}{\Delta_i(\Delta_i - 1)}, n_j, n_k \in N_i, e_{j,k} \in E. \tag{2.12}$$

From Eq. (2.12), $0 \leq C_i \leq 1$. It is clear that high local clustering coefficient of a vertex implies a well interconnected neighborhood. But this claim may not be true for estimating group coherence in a weighted graph, which is equivalent to a network in which correlations among nodes are different.

To explain this discrepancy, we present two examples for calculating a unweighted and weighted node local clustering coefficient in Figure 2.20. There are six nodes in a group with levels of DLRs ($DLR = \{0.2, 0.5, 0.8\}$) among them. Weighted clustering coefficient for each vertex in both case 1 and case 2 in Figure 2.20.

In Case 1, Eq. (2.12) gives the unweighted local clustering coefficient of node A, $C_A = 0.6$,

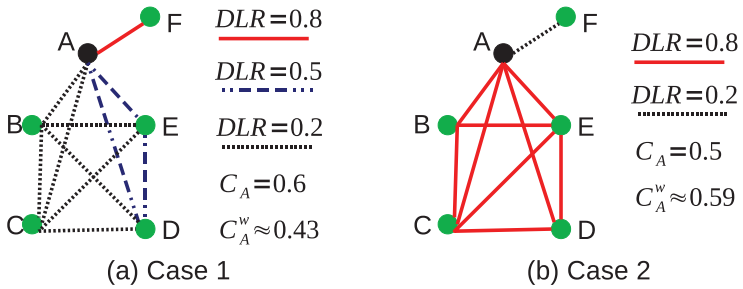

Figure 2.20: Unweighted and weighted local clustering coefficients.

which means a well interconnected neighborhood of node A. However, the weights of edges forming the interconnected triangle vertices (A,D,E), or correlations among (A,D,E), are comparably larger than those forming the interconnected triangle vertices (A,B,C). This indicates that the triangle (A,D,E) plays more important role in determining the local group cohesiveness than triangle (A,B,C). Furthermore, among all A's neighbors, though F has the strongest correlation with A, it has no direct connection with A's other neighbors. It further indicates that A's clustering properties may be *overestimated* by simply considering the physical topological information.

Based on above observations, DLRs among nodes must be taken into account in order to accurately estimate the local group coherence. In analogy with the weighted clustering coefficient defined in [66], which however only considers the weights of the edges *connected* to n_i, we explicitly take into account weights of *all* edges in triangle (n_i, n_j, n_k), and define the weighted local clustering coefficient of node n_i as

$$C_i^\omega = \frac{\sum_{j,k\in N_i} \frac{DLR_{i,j}+DLR_{i,k}+DLR_{j,k}}{3} \cdot 1_{\{h(j,k)=1\}}}{\sum_{j,k\in N_i} \frac{DLR_{i,j}+DLR_{i,k}}{2}}, \tag{2.13}$$

where $1_{\{.\}}$ is the indicator function.

Distinguished from Eq. (2.12), Eq. (2.13) takes into account the correlations between n_i and its neighbors. When correlations among nodes are the same, $C_i^\omega = C_i$. By applying Eq. (2.13) to recalculate C_A^ω for Case 1 in Figure 2.20, as we expected, $C_A^\omega \approx 0.43 < C_A$. Furthermore, although the physically interconnected neighbors in Case 2 is less than Case 1 ($C_A(2) = 0.5 < C_A(1) = 0.6$), $C_A^\omega(2) \approx 0.59 > C_A^\omega(1)$ due to the strong DLRs of triangle pairs (A,B,C) and (A,D,E) incident to node A in Case 2.

Remark 13 *For node n_i in a group, it either has $C_i^\omega > C_i$ or $C_i^\omega \leq C_i$. In the former case, n_i has strong correlations with its neighbors. In contrast, for the latter case, the group is generated by the nodes connected with weak correlations. Therefore, n_i can show even strong local group*

coherence when it has high clustering coefficient C_i and $C_i^\omega > C_i$, which is the reverse case when showing very weak group coherence.

Definition 6 *Given the weighted local clustering coefficient for each node in a group, we can easily find the coherence level of entire group. Define* group coherence *as*

$$C_G = \frac{1}{V(G)} \sum_{i \in V(G)} C_i^\omega. \tag{2.14}$$

With the same example in Figure 2.20, we observe that the group coherence in Case 2 is higher than Case 1, even though there are more direct connections among group members in Case 1. Therefore, similar to the property of C_i^ω, a high value of C_G implies an overall strongly connected group structure, which is due to stable link connections among mobile nodes.

Conditions for stable groups

Basically, stable group means that its composition remains unchanged over time. In other words, the internal connections among nodes within a group are stronger than the external ones. To proceed, let Δ_i denote the degree of node n_i. Define $\Delta_i = \Delta_i^{in}(G) + \Delta_i^{out}(G)$, where $\Delta_i^{in}(G)$ and $\Delta_i^{out}(G)$ represent the number of links between n_i and its neighbors *inside* and *outside* of the group G, respectively. Then, $\Delta_i^{in}(G)$, along with the sum of DLRs of node n_i with its neighbors *inside* group G, indicate the strength of n_i with group G.

Theorem 1 A group $G = (V(G), E(G))$ is **stable in a strong sense** if $\forall i \in V(G)$

$$\sum_{j \in N_i^{in}} DLR_{i,j} > \sum_{k \in N_i^{out}} DLR_{i,k}, \quad \text{and} \quad \Delta_i^{in}(G) > \Delta_i^{out}(G). \tag{2.15}$$

And it is **stable in a weak sense** if

$$\sum_{j \in N_i^{in}} DLR_{i,j} > \sum_{k \in N_i^{out}} DLR_{i,k}, \sum_{i \in V(G)} \Delta_i^{in}(G) > \sum_{i \in V(G)} \Delta_i^{out}(G), \tag{2.16}$$

where N_i^{in} and N_i^{out} are sets of n_i's neighbors inside and outside group G, respectively.

For a stable group in a *strong* sense, each node has more connections within the group than outside the group. For a stable group in a *weak* sense, the sum of all connections within the group is greater than the total connections toward the nodes outside the group. It is clear that the condition of Eq. (2.15) satisfies Eq. (2.16), but the reverse is not true. More importantly, in Theorem 1, a stable group in both strong and weak sense must satisfy the condition that for each node, sum of correlations with its neighbors inside the group is stronger than that with

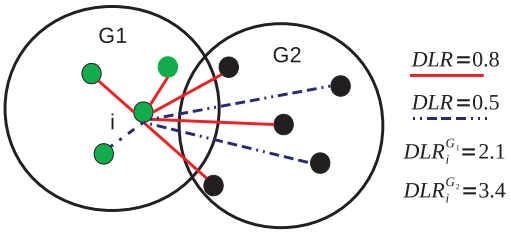

Figure 2.21: Stable Groups and Network Topology: Switching Groups.

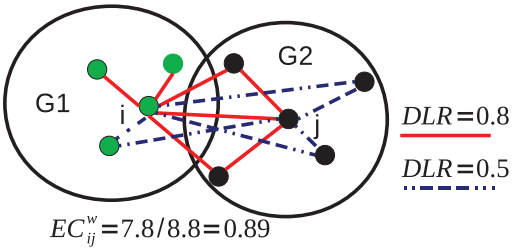

Figure 2.22: Stable Groups and Network Topology: Groups Merging.

neighbors outside the group. Otherwise, the group is unstable and the node may switch to other groups.

In wireless networks, stable groups mean less dynamics in network topology. In other words, unstable groups, such as node switching among groups and a large group splitting into several small groups, can lead to broken links, even to network partitions. Conditions for stable groups can assist evaluating network dynamics, according to which different routing strategies may be applied for better performance.

Conditions for group evolution

Group evolutions frequently happen in mobile networks, such as group contraction, split, and combination in Figure 2.5. Therefore, we investigate the conditions for group evolution, which are beneficial for predicting network connectivity and partitions. We focus on two common group evolution phenomena, that are node *switching* from one group to another and two groups *merging* into one group.

Based on Theorem 1, a group is reluctant to change its structure when the condition of Eq. (2.15) is satisfied. Otherwise, a node probably can switch from one group to another. Hence, n_i will *switch* from G_1 to G_2 when

$$\sum_{k \in N_i^{G_2}} DLR_{i,k} > \sum_{j \in N_i^{G_1}} DLR_{i,j}. \tag{2.17}$$

In an example shown in Figure 2.21, node n_i has 3 neighbors in G_1 and 5 neighbors in G_2, we can calculate values of both sides of Eq. (2.17). As $\sum_{j \in N_i^{G_1}} DLR_{i,j} = 2.1$, and $\sum_{k \in N_i^{G_2}} DLR_{i,k} = 3.4$, node n_i is likely to switch from G_1 to G_2.

Note that two groups can communicate when there exists at least one link between two users from different groups. Intuitively, the correlation strength of the inter-group edge $e_{i,j}$ is proportional to the number of their common neighbors. In graph theory, a common neighbor

n_k implies that there is a triangle connection among node tuple (n_i, n_j, n_k). Accordingly, the more triangles the edge $e_{i,j}$ is attached, the stronger tie groups G_i and G_j has. Therefore, the weighted *edge* clustering coefficient $EC^{\omega}_{i,j}$ can be used to represent *betweenness of groups* in social networks [66]. Specifically, $EC^{\omega}_{i,j}$ is defined as the ratio of the number of triangles that edge $e_{i,j}$ is currently attached to over the maximum number of triangles that can be potentially included:

$$EC^{\omega}_{i,j} = \frac{\sum_{k \in N_i \cup N_j, k \neq i,j} (DLR_{i,k} + DLR_{j,k}) \cdot 1_{\{h(i,k)=h(j,k)=1\}}}{\sum_{k \in N_i, k \neq j} \varepsilon_i + \sum_{k \in N_j, k \neq i} \varepsilon_j}, \tag{2.18}$$

where $\varepsilon_i = DLR_{i,k} \cdot 1_{\{h(j,k)=1\}} + 1_{\{h(j,k)\neq 1\}}$ for n_i, and $\varepsilon_j = DLR_{j,k} \cdot 1_{\{h(i,k)=1\}} + 1_{\{h(i,k)\neq 1\}}$ for n_j. They represent the total possible weighted edge contribution in forming the triangles in the neighborhood of n_i or n_j.

To observe the betweenness of two groups, we present an example in Figure 2.22. Many common neighbors and stronger correlations between n_i and n_j result in high weighted edge clustering coefficient, that is $EC^{\omega}_{i,j} = 0.89$. There exists a higher chance that two groups can merge to be a larger single group.

In multihop adhoc networks, node switching means that some links may break while new links may establish, i.e., changes of network topology; while groups merging implies that nodes in these groups are well connected, i.e., quality of network connectivity is good. Therefore, conditions for group evolutions can help us to predict network connectivity.

In addition, we need to point out that the inter-node correlation can be helpful in many applications in addition to our analysis of link dynamics and network topology. For example, in mobility-aware routing, two nodes with high DLR move closely and probably maintain a stable link, and thus are more suitable for establishing routes with high stability. In data forwarding in DTNs, if a node currently carrying a message sends a copy to a node with different mobility patterns, i.e., low temporal locality similarity between them, the chance of at least one of them meeting destination could be increased. In mobility-aware clustering, since the node with high weighted clustering coefficient in Eq. (2.13) is much less likely to be disconnected from its neighbors, therefore communication overhead for changing clusterhead can be reduced if it is selected as clusterhead.

2.5.3 Assisting Data Forwarding

As an application, we develop a TLS based data forwarding mechanism for information dissemination in delay tolerant networks (DTNs). The relays are selected among the neighbors of message carrier based on their DLR with destination, to ensure that the delay of delivery and number of relays can be reduced.

Assume neighboring nodes in DTNs can exchange their cave profiles, and the data source

has knowledge about the destination D's $O(0)$ cave profile $P_D = \{P_D(c_1), \ldots, P_D(c_M)\}$ (refer to Eq. (2.4)), which is the long-term cave preference probability set. Suppose the neighbors set of data source S is $N_S = \{n_1, \ldots, n_k\}$, which can be selected as relays. We assume D is not in the set N_S, otherwise S can simply transmit the data to D immediately. A relay is selected according to the following steps:

- S collects the $O(k)$ cave profiles of its neighbors $n_i \in N_S$ by exchanging messages.
- S estimates the TLS between its neighbor n_i and D by calculating $TLS_{i,D}$ through Eq. (2.6) and (2.7), which use D's $O(0)$ and n_i's $O(k)$ cave profiles.
- S selects relay node R_1 that has the strongest TLS with D, i.e., $TLS_{R_1,D} = \max\{TLS_{i,D}, n_i \in \{N_S \cup S\}\}$.
- R_1 forwards to another relay R_2 if $TLS_{R_2,D} = \max\{TLS_{i,D}, n_i \in \{N_{R_1} \cup R_1\}\}$.
- The message is forwarded to D by the selected relay nodes set $R = \{R_1, R_2, \ldots, R_k\}$.

Because the temporal locality similarity is based on mobility pattern similarity, a node is selected as a relay if it has the highest probability to appear at the same community site with D among all the neighbors of the message carrier. Since node's $O(k)$ cave profile is time-varying, we can only estimate the temporal locality similarity between a node and the destination by using D's $O(0)$ cave profile. As $O(0)$ cave profile represents the probability/preference of a node being each cave, estimation of $TLS_{i,D}$ indicates the possibility that node n_i will appear at same location with D. Therefore, TLS assisted strategy should be able to disseminate information to destination through fewer relay nodes within shorter time than random forwarding that randomly chooses a neighbor as relay.

We implement TLS-assisted data forwarding in OMNeT++ and INET-Framework [58]. Figure 2.23 shows TLS-assisted algorithm's delivery ratio comparing with random forwarding in both random mobility and time-space varying caveman (TSC) mobility scenarios. Under RWP mobility, TLS-assisted forwarding outperforms random forwarding by 10% in delivery ratio, while it performs much better performance than random forwarding under TSC mobility because nodes can take its advantages of node mobility patterns to assist data forwarding.

2.5.4 Assisting Clustering

We further utilize dual locality ratio in mobility-aware clustering, which is one of the most general applications of topology control, and show its benefit of lower clusterhead changing rate comparing with lowest-ID algorithm [44].

In lowest-ID clustering algorithm, each node is randomly assigned an ID, and the node with smallest ID among its neighbors acts as clusterhead. Because mobile users may move at high

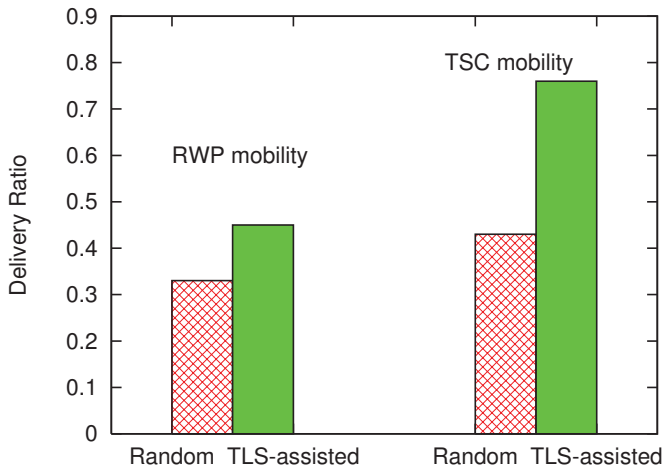

Figure 2.23: Delivery ratio: TLS-assisted versus random forwarding.

speed and travel different routes to their destinations, a node may be frequently disconnected with its randomly selected clusterhead, i.e., clusters based on lowest-ID are unstable. The stability of clusters could be improved by selecting clusterhead based on average DLRs of nodes with their neighbors.

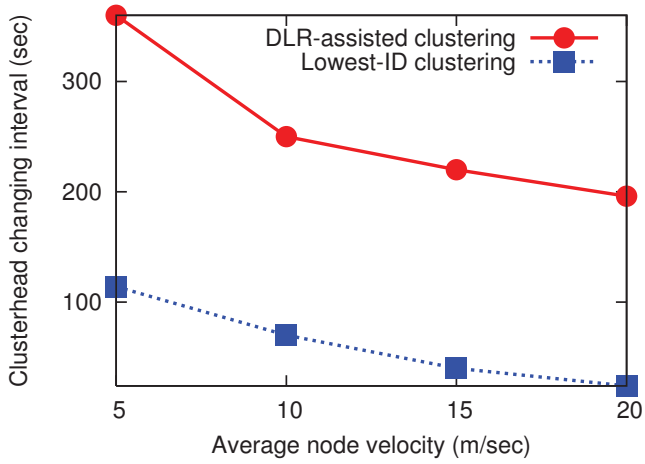

Figure 2.24: DLR assists clustering under TSV mobility.

Using the same simulation setting as in Section 2.4.2, the average clusterhead changing time is measured to indicate cluster stability. Figure 2.24 shows that under various node speeds, changing interval of clusterhead that is selected base on DLR is much longer than that based

on lowest-ID. A node with higher average DLRs not only moves closely with its neighbors, but also may share overlapping paths, therefore is much less likely to be disconnected from its cluster members if being selected as the clusterhead, i.e., changing rate of clusterhead is reduced. Therefore, DLR can be used to obtain stable clustering for topology control of vehicle-to-vehicle network.

2.6 Summary

In this chapter, we present our observations, measurements, and applications of mobility correlation in both spatial and temporal locality. In spatial domain, mobility of a node depends on its location and mobility correlation of two nodes exhibits in relative distance, speed, and moving direction. In temporal domain, mobility of a node shows difference preferences over different locations over time and the mobility correlation of two nodes is indicated by similarity in mobility patterns. By measuring the similarity in both spatial and temporal domains of mobility between two nodes, we propose a dual-locality ratio (DLR) metric to quantify the inter-node mobility correlation. DLR is shown to effectively identify node groups in real traces as well as in simulations. Furthermore, DLR is shown to have implications on link lifetime and group stability, and is utilized for evaluating group stability, providing conditions for group evolutions, assisting data forwarding and clustering in topology control.

Chapter 3

Analyzing Performance of Geocast Message Dissemination in Intermittently Connected VANETs

Vehicular ad hoc network (VANET) is one of the most promising large-scale applications of mobile ad hoc networks, which can facilitate the design of *intelligent transportation systems* (ITS). Many VANET applications, such as traffic collision warning, require message dissemination in certain geographic regions (i.e., *area of interest*), which is referred to as *Geocast*. The challenges of geocast come from highly dynamic environments on the road. Destination nodes in geocast are *dynamic* over time due to vehicle mobility, which undermines our existing results of dissemination latency and information propagation speed with *pre-determined* destinations. Moreover, the area affected by the dissemination is vital in geocast as it determines the dissemination latency for spreading the message to nodes located at certain places inside of the area of interest (AOI).

In this chapter, we study new questions presented by geocast: how far the dissemination can reach by time t (referred as *dissemination distance*) and how long the dissemination takes to inform nodes located at certain locations (referred as *hitting time*). As only nodes actively spreading a message contributes to dissemination (referred as disseminator), we first model message mobility that includes movement of disseminator and transmission between disseminators in order to shroud dynamic destination nodes of geocast. Then, analytic bounds of dissemination distance and hitting time are derived based on message mobility, which serve as the spatial and temporal limits of geocast. Analytic results are further validated by simulation results of several dissemination algorithms. Putting together, two application scenarios are provided to illustrate how our results serve as guidelines to choose or design appropriate dissemination methods for different vehicle-to-vehicle network applications.

3.1 Motivation and Related Work

Vehicular Ad hoc Networks (VANETs) have emerged as a radically new paradigm and served as a cornerstone to the design of the *Intelligent Transportation System* (ITS), which has attracted much attention from the government, car manufacturers, and researchers. In the US, FCC (Federal Communications Commission) assigned 75 MHz of spectrum for Dedicated Short Range Communications (DSRC) [67] in 2003. In Europe, national governments, car manufacturers, and the European Commission are pushing for a new research effort in this area in order to develop a network system assisting the drivers. More recently, IEEE also formed the new IEEE 802.11p task group [17] that focuses on providing wireless access for the vehicular environments to improve road safety, traffic efficiency, and driving convenience. A list of applications of VANETs can be found in *Car2Car Communication Consortium* (C2CCC) [18], which aims to standardize inter-vehicle communication. For instance, CarTalk2000 [68] promotes a driver assistance system based on inter-vehicle communication; FleetNet [69] focuses on developing vehicular ad hoc networks that satisfy the drivers' and passengers' needs for local dependent information and services.

Many VANET applications are heavily dependent on the message dissemination in specific geographic regions, i.e., *geocast*. Many researchers have studied the performance of message dissemination in various application scenarios, such as MANETs and VANETs. Essentially, most researchers are interested in: (i) how *fast* the information can be spread, and (ii) how *far* the dissemination can propagate. Hence, the metrics of message dissemination performance can be in time domain (e.g., dissemination latency and propagation speed), and space domain (e.g., propagation distance) [70].

In the time domain, message dissemination latency and information propagation speed have been studied in MANETs [19–22, 71, 72] as well as in VANETs [73–78]. Zheng [19] derived fundamental limits of the broadcast capacity and information diffusion rate of information dissemination in power-constrained large-scale wireless networks. Zheng showed that there is a constant upper bound on the information diffusion rate in large wireless networks. Xu and Wang [20] investigated the speed limit of information propagation in large wireless networks and showed that there exists a unified speed upper bound for broadcast and unicast communications. Paper [71] presented a graph based model to characterize connectivity properties and derived a scaling law for message delay in large scale disconnected ad hoc networks. Kong and Yeh [22] found that delay scales linearly with the Euclidean distance between the sender and the receiver when the network is in the subcritical phase, and the delay scales sub-linearly with the distance if the network is in the supercritical phase. Paper [21] analyzed the information propagation speed limits in large scale mobile and intermittently connected networks (i.e., Delay Tolerant Networks (DTNs)). Jacquet et al. [72] derived the asymptotic capacity and delay in large scale

mobile networks and showed a practical throughput-delay trade-off.

Because real D2D communication network has finite size and number of nodes, nonasymptotic results on performance of message dissemination are also needed. Specially, the latency and speed of message dissemination in VANETs have received lots of attention. Fracchia and Meo [77] analyzed the average delay, the probability that a vehicle is informed, and the average number of duplicate messages received by a vehicle within a area of interest in a 1-D highway scenario. Similarly, paper [74] developed upper and lower bounds for the time of information propagation between two nodes in a 1-D network and showed that more vehicles on the road does not necessarily promote the fast propagation of information. The relationship between latency and reliability is studied in [78]. Jacquet et al. [75] analyzed information propagation in bidirectional vehicular DTNs. The authors proved and computed a threshold of vehicle density, above which information speed increases dramatically over vehicle speed, and below which information propagation speed is on average equal to vehicle speed.

On the contrary, the space metric domain receives less attention [73, 76, 79, 80]. Paper [73] derived spatial propagation of information in a 1-D vehicle-to-vehicle ad-hoc network. Both analysis and simulation show that information propagation depends on vehicle traffic characteristics, such as vehicle density, average vehicle speed, and relative speed among vehicles. Resta et al. [79] derived lower bounds on the probability that a car at distance d from the source of the emergency message correctly receives the message within time t. In a 1-D static network scenario, the authors showed that besides d and t, this probability depends also on 1-hop channel reliability and the message dissemination strategy. Further considering node mobility, [80] analyzed the propagation distance in 1-D VANETs with constrained vehicle mobility. Jacquet et al. [76] showed that when the vehicle density is smaller than a threshold, routing using bidirectional traffic in bidirectional vehicular DTNs provides a gain in the propagation distance, which follows a sublinear power law as time elapses.

In addition to the least attention on spatial propagation [70], research on spatial propagation is mainly performed in 1-D VANETs with static or simple node mobility (e.g., constant speed or moving direction). Nonetheless, spatial propagation is an important performance metric as many D2D applications target information delivery in geographic regions (i.e., geocast), such as collision warning, traffic congestion warning, and regional weather forecast. By limiting destinations only to vehicles in AOI, geocast helps to avoid the broadcast storm problem [81] and enables the coexistence of multiple VANET applications. At the same time, geocast introduces challenges to algorithm and protocol design, system evaluation, and even performance metrics of VANETs. Because vehicles can move into and out from the AOI, geocast has *dynamic* destination nodes. Such a *dynamic* group of destinations is different from message dissemination that specifies destination nodes prior to transmissions in traditional MANETs. There still lacks understanding of the spatial propagation properties of geocast in 2-D networks with realistic

node mobility, which we will study in this chapter. Our contributions are three-fold.

1. We model message dissemination by message mobility, which focuses on where active messages reach rather than by which nodes carry and relay. Without specifying relay nodes on information propagation path, message mobility can shroud dynamic destination nodes of geocast and enable us to analyze the dissemination distance and hitting time.

2. Based on the formulation of active message mobility, we derive lower and upper bounds for the farthest distance that active messages reach at time t and the first hitting time that active messages reach devices at distance d from the original source location under L-copy ($L \geq 1$) direction-invariant and geographic-assisted dissemination strategies, respectively. Simulation results show that several well known dissemination algorithms, including stateless opportunistic forwarding (SOF) [82] and GPS-based broadcasting (GBB) [83], are well bounded by our analytic bounds. Both analysis and simulation evidence that regardless of the number of disseminators used in the dissemination, the upper bound on expected dissemination distance increases with the *square root* of time t in direction-invariant dissemination strategy, while it increases approximately *linearly* with time t in geographic-assisted dissemination strategy.

3. We apply our analytical and simulation results to two real VANET applications, i.e., *post-crash warning* and *emergency vehicle signal preemption*, to provide guidelines for dissemination algorithm design. Our results suggest that for VANET applications that target an area near the source location, dissemination algorithms with multiple disseminators are suitable, while for applications that target an area far from the source location, geographic-assisted dissemination strategy is preferable in order to satisfy the application requirements.

The rest of this chapter is organized as follows. In Section 3.2, we introduce our network and vehicle mobility models, define and classify dissemination strategies, and formulate message mobility. Lower and upper bounds on dissemination distance and hitting time are derived in Section 3.3. In Section 3.4, we validate our analytic bounds by comparing them with simulation results, and show how our results provide guidelines for choosing appropriate dissemination methods in VANET applications. We conclude this chapter in Section 3.5.

3.2 Models and Problem Formulation

In this section, we first introduce our network and mobility models and two general dissemination strategies, then formally define dissemination distance and hitting time with formulation of the mobility of active message.

3.2.1 Network and Mobility Models

Assume that at time 0, n nodes $\{\mathcal{X}(0)\} = \{X_1(0), \ldots, X_n(0)\}$ are uniformly distributed at random in a two-dimensional torus $\mathcal{B} = [0, B]^2$, where $B = \sqrt{n/\lambda}$ for some $\lambda > 0$. The random vector $X_i(0)$ denotes the location of node i at time 0. By definition [84], $\{\mathcal{X}(0)\}$ is a homogeneous Poisson point process. n nodes are Poisson distributed in the network with density $\lambda = n/B^2$ everywhere. The average number of neighbors per node is therefore smaller (or equal) than $\pi r^2 \frac{n}{\mathcal{B}}$. In [85], Xue and Kumar have shown that if the average number of neighbors is smaller than $0.074 \log n$, then the network is almost surely disconnected when n is large. In order to study the properties of intermittently connected VANETs, we further assume that λ is small such as to capture the intermittent connectivity in vehicle to vehicle networks due to node mobility and limited radio coverage.

Suppose that time is slotted and each node moves according to a given mobility model $M(t), t = 1, 2 \ldots$. In other words, the displacement of node i from its position $X_i(t-1)$ to $X_i(t)$ is distributed according to $M(t)$. Two nodes i and j can communicate with each other at time t if only if their distance is less or equal to transmission range r, i.e., $d_{i,j}(t) \triangleq ||X_i(t) - X_j(t)|| \leq r$. We consider a generic mobility model [86] defined as the following.

Definition 7 *(Generic Mobility) Given initial nodes' positions $\mathcal{X}(0)$ at $t = 0$, the spatial distribution $X_i(t)$ of node i at time slot t is around a point x_i^* by a non-increasing and direction-invariant function $\Psi_i(x) = \Psi(x - x_i^*)$. Assume that Ψ_i is non-zero in and only in a region characterized by a constant a; that is, $\Psi_i(x) = \Psi(x - x_i^*) > 0$ when $||x - x_i^*|| < a$ and $\Psi_i(x) = \Psi(x - x_i^*) = 0$ otherwise.*

This mobility model is very general such that it covers a wide range of possible scenarios of realistic mobility processes. The case of static nodes uniformly deployed over the area can be obtained by setting $\Psi_i(x) = \delta(x - X_i(0))$. The i.i.d. mobility model in [87] corresponds to the case when $\Psi(x)$ is a constant function independent of x and $a = \infty$. When $a < \infty$ and $x_i^* = X_i(0)$, we obtain the constrained i.i.d. mobility model used in [22].

To mimic vehicle mobility which is constrained by the speed limit and dependent on previous movements, we assume that $a < \infty$ and $x_i^* = X_i(t-1)$ at time slot t, thus $X_i(t)$ is uniformly distributed at random in $\mathcal{A}(X_i(t-1), a)$-the circular region centered at $X_i(t-1)$ with radius $a > 0$. The positions $X_i(t)$ are mutually independent among all nodes and only dependent on previous locations $X_i(t-1)$. This mobility process can also be interpreted as the following. At each time slot, a node chooses a random direction uniformly from $[0, 2\pi]$ and travels for a random length which is chosen from $[0, a]$ following certain distribution. Denote by $A(t)$ and $\theta(t)$ the length and angle of the movement step at time t, respectively. Movement vector of a node i at time t is $Y_M(t) \triangleq A(t)e^{j\theta(t)}$ (j is imaginary unit) with origin at $X_i(t)$ and endpoint uniformly distributed in $\mathcal{A}(X_i(t), a)$. Since both sequences $\{A(t)\}$ and $\{\theta(t)\}$ are i.i.d. and independent

from each other, $Y_M(t)$ $(t = 1, 2 \dots)$ are also i.i.d. random variables. This constrained i.i.d. mobility satisfies the following Lemma.

Lemma 1 *Under constrained mobility in Definition 7, n nodes are Poisson distributed in the network $\mathcal{B}$ with density λ everywhere at all times.*

Proof : Let the circular region $\mathcal{A}$ centered at a random selected location in the network with radius a be partitioned to $m+1$ concentric rings $\mathcal{A}_i$ indexed by $i, 0 \leq i \leq m$, each of which is with equal ring width ϵ. Note that $\mathcal{A}_0$ is a circle centered at the center of $\mathcal{A}$ with radius ϵ. Denote the area enclosed by ring $\mathcal{A}_i$ as $S_{\mathcal{A}_i}$, and define by $N_i(t)$ the number of nodes in $\mathcal{A}_i$ at time t. Since nodes are Poisson distributed in the network area initially, i.e., $N_i(0) \sim Pois(\lambda S_{\mathcal{A}_i})$.

Under constrained mobility model in Definition 7, when $t = 1$, number of nodes in interval $\mathcal{A}_0$ is

$$N_0(1) = N_0(0) \int_{\mathcal{A}_0} \frac{1}{\pi a^2} d\mathcal{A}_0 + \cdots + N_m(0) \int_{\mathcal{A}_m} \frac{S_{\mathcal{A}_0}}{S_{\mathcal{A}_m}} \frac{1}{\pi a^2} d\mathcal{A}_m. \tag{3.1}$$

Denote $p_i \triangleq \int_{\mathcal{A}_i} \frac{S_{\mathcal{A}_0}}{S_{\mathcal{A}_i}} \frac{1}{\pi a^2} d\mathcal{A}_i, i = 0, 1, \dots, m$, and $N_i^*(0) \triangleq p_i N_i(0)$. Accordingly,

$$\begin{aligned} P(N_i^*(0) = k^*) &= \sum_{k \geq k^*} \tbinom{k}{k^*} p_i^{k^*} (1-p_i)^{k-k^*} \cdot e^{-\lambda S_{\mathcal{A}_i}} \frac{(\lambda S_{\mathcal{A}_i})^k}{k!} \\ &= e^{-\lambda S_{\mathcal{A}_i}} \frac{(p_i \lambda S_{\mathcal{A}_i})^{k^*}}{k^*!} \sum_{k \geq k^*} \frac{[(1-p_i)\lambda S_{\mathcal{A}_i}]^{k-k^*}}{(k-k^*)!} \\ &= e^{-p_i \lambda S_{\mathcal{A}_i}} \frac{(p_i \lambda S_{\mathcal{A}_i})^{k^*}}{k^*!}. \end{aligned} \tag{3.2}$$

Clearly, $N_i^*(0)$ follows Poisson distribution with parameter $p_i \lambda S_{\mathcal{A}_i}$. Since $\{N_i^*(0), i = 0, \dots, m\}$ are independent Poisson random variables and $N_0(1)$ equals to the sum of $N_i^*(0)$,

$$N_0(1) \sim Pois(\sum_{i=0}^{m} p_i \lambda S_{\mathcal{A}_i}) \sim Pois(\lambda S_{\mathcal{A}_0}). \tag{3.3}$$

Similarly, number of nodes in any small circular region in the network at $t = 1$ follows the same Poisson distribution as at $t = 0$. Assume $N_i(t) \sim Pois(\lambda S_{\mathcal{A}_i}), 0 \leq i \leq m$, using the same method as above, we can show that $N_i(t+1) \sim Pois(\lambda S_{\mathcal{A}_i}), 0 \leq i \leq m$. Therefore, by induction, nodes are Poisson distributed in the network area at all times.

Remark 14 *We use the* constrained vehicle mobility *model because (i) it generally accounts for a wide range of realistic mobility processes in vehicular scenarios, including Manhattan mobility [12] and random walk, (ii) it reflects the speed limit of nodes (usually wireless devices carried by humans or installed on vehicles) that nodes can jump to adjacent locations with pre-assigned probabilities and each movement step is limited in a circular region around previous*

location, and (iii) n nodes are Poisson distributed in the network $\mathcal{B}$ with density λ everywhere at all times.

3.2.2 Dissemination Strategies

Dissemination performance, such as information propagation speed and dissemination latency, depends on how many nodes are recruited to disseminate the message (i.e., number of disseminators) and how the disseminators are chosen. By using as many disseminators as possible, full epidemic broadcast achieves best performance, but leads to network congestion. Hence, limited-copy dissemination (i.e., limited number of disseminators) is more feasible in order to save network resources and enable the coexistence of multiple applications. If being used for choosing disseminators, geographic information has the potential to enhance performance of geocast. But, geographic information may be unavailable for all vehicles in the network and exchanging geographic information consumes the limited network resources. Since number of disseminators and whether geographic information can be used in disseminator selection affect performance of dissemination strategies, we classify dissemination strategies according to these two factors, based on which we study geocast performance.

Definition 8 *(1-Copy Message Dissemination) Assume that at time* 0*, node v_0 initiates a message dissemination and acts as the disseminator. There is only one disseminator at each time slot. The disseminator will rebroadcast the message to its neighbors until it finds the next-hop disseminator. Disseminator is selected based on criteria imposed by applications. This process repeats until the dissemination completes.*

1-copy message dissemination is particularly useful in the following situations: 1) network has limited capacity; 2) network load is heavy; 3) nodes are computationally-constrained or energy-constrained devices. In these situations, the network could only support one disseminator in order to save network resources and enable the coexistence of multiple applications.

Definition 9 *(L-Copy Message Dissemination) Assume that node v_0 initiates a message dissemination at time* 0*. First, the message will be spread to L distinct disseminators. Then, each disseminator independently disseminates the message according to 1-copy message dissemination in Definition 8.*

As multiple disseminators actively rebroadcast the message at the same time, L-copy message dissemination likely increases the speed of message arriving AOI and enhance the probability of successful deliveries to destinations (i.e., dissemination reliability). Thus L-copy message dissemination could be favorable for time critical message dissemination of safety applications in VANETs.

Remark 15 *In a message dissemination, if neighbors' locations or speeds are unavailable, a disseminator chooses its next-hop disseminator isotropically (equally in all directions), which we refer to as* 1-copy *and* L-copy direction-invariant dissemination*; otherwise, geographic information can be used to assist disseminator selection such as to enhance dissemination performance (such as propagation speed), which we refer to as* 1-copy *and* L-copy geographic-assisted dissemination*.*

Direction-invariant dissemination can be implemented by imposing the receivers to probabilistically decide whether to become a disseminator; geographic-assisted dissemination can be achieved by scheduling receivers to broadcast their decision based on their locations. In either strategy, the feedback mechanism is used between two disseminators such that the previous disseminator stops broadcasting the message and the new disseminator starts spreading the message. Note that the traffic induced by feedback mechanism would not be overwhelming in intermittently connected networks.

3.2.3 Problem Formulation

In geocast, we are interested in the dissemination within the AOI, such as whether the message has reached vehicles in the AOI or how far the message is from the AOI. Hence, we study how far the dissemination has reached by time t and how long the dissemination takes to spread the message to certain location, i.e. the spatial and temporal limits.

In order to derive spatial and temporal limits of geocast, we do not consider the effects of buffering or congestion, and assume that a message can be transmitted instantaneously between two nodes in range (i.e., omit the transmission delay). Under these assumptions, we are able to derive spatial and temporal bounds of geocast since they correspond to an ideal scenario with that respect. Actually, previous assumptions have little impact on the accuracy of our results because information transmission occurs much faster than the speed of the mobile nodes and propagation delay is much smaller than the dissemination latency incurred by dynamic topology and intermittent connectivity in VANETs.

Dissemination Distance

In a dissemination, the message copy held by the disseminator is called *active* message, otherwise called *latent* message. Clearly, nodes holding latent messages contribute nothing to increase delivery ratio and decrease dissemination latency. Hence, we ignore the latent messages and their carriers and focus on active messages and disseminators.

Denote by $\mathcal{V}(t)$ the set of disseminators at time t and $1 \leq |\mathcal{V}(t)| \leq L$. Let us place a Cartesian coordinate system in the network with its origin at the source location. The *dissemination vector*

$D(t)$ is the vector from source point $X_0(0)$ to the location of the farthest disseminator at time t. The length of dissemination vector is called *Dissemination Distance*, which is defined as

$$|D(t)| \triangleq \max_{v_k \in \mathcal{V}(t)} \{||X_k(t) - X_0(0)||\}. \tag{3.4}$$

$|D(t)|$ by definition is the distance from the source location to the farthest location reached by disseminators by time t.

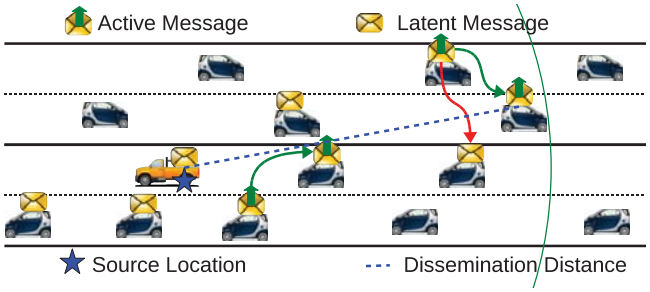

Figure 3.1: In the circular region centered at the source with $|D(t)|$ as diameter, nodes have at least partially received the message by time t.

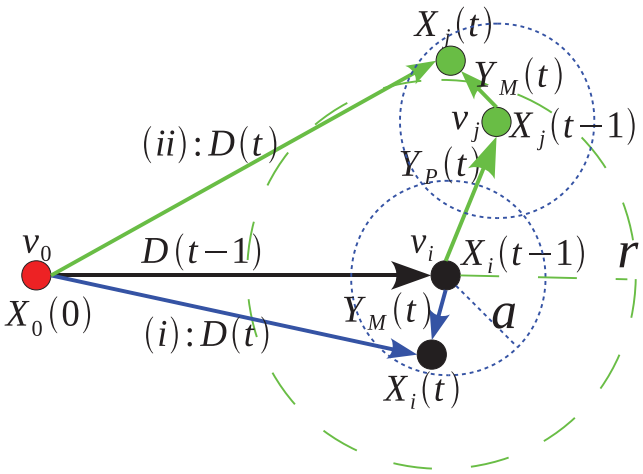

Figure 3.2: Dissemination distance varies due to movements of disseminators and jump of active message (i.e., transmission) from disseminator v_i (the black node) to next disseminator v_j (the green node).

To avoid specifying the relay nodes, we study $D(t)$ through the mobility of active messages.

In 1-copy message dissemination, denote by $Y(t) = D(t) - D(t-1)$ the progress of active message from time $t-1$ to t ($t = 1, 2, \ldots$.). As shown in Figure 3.2, suppose v_i (the black node) is the disseminator at time $t-1$, (i) if v_i is also the disseminator at time t, the active message moves with v_i, which means that $Y(t)$ equals to the movement step $Y_M(t)$ of v_i; (ii) if node v_j ($i \neq j$) (the green node) is selected as the next disseminator at time t, the active message jumps from v_i to v_j and moves with v_j, which means that $Y(t)$ equals to the propagation vector $Y_P(t)$ plus the movement step $Y_M(t)$ of v_j. $Y(t)$ is affected by constrained node mobility, limited transmission range, and disseminator availability.

In L-copy message dissemination, we number each active message from 1 to L and denote by $|D_k(t)|$ the farthest distance reached by the k^{th} active message. We assume that each active message is independent from each other. Hence, the progress of the k^{th} active message, $Y_k(t)$, equals either $Y_M(t)$ or $Y_M(t) + Y_P(t)$.

Remark 16 *For* direction-invariant *dissemination,* $E(Y_P^x(t)) = E(Y_P^y(t)) = 0$*; for* geographic-assisted *dissemination that utilizes geographic information to increase dissemination distance* $|D(t)|$, $E(Y_P^x(t)) \geq 0$ *when* $D_x(t) \geq 0$ *and* $E(Y_P^x(t)) \leq 0$ *when* $D_x(t) < 0$ *(the same for the y-component).*

Hitting Time

To find the first time of the dissemination reaching certain location, we define *hitting time* as

$$\tau(d) \triangleq \inf_{t>0}\{t : |D(t)| \geq d\}, \tag{3.5}$$

where d is a positive constant. Essentially, $\tau(d)$ is the message's fist hitting time of the region that is outside the circular region centered at $X_0(0)$ with radius d.

Dissemination distance and hitting time manifest the spatial and temporal limits of geocast, respectively. Dissemination distance reveals the size of the zone affected by the dissemination in which nodes are at least partially informed by the message. Hitting time uncovers the minimum latency of reaching nodes at certain locations. Putting together, they can be used to determine whether the dissemination has reached vehicles in the AOI and whether vehicle-to-vehicle communication can possibly satisfy the time requirements of time critical safety applications in VANETs. Intuitively, different dissemination strategies exhibit different performance in dissemination distance and hitting time. We expect our results to also provide guidelines on choosing appropriate dissemination methods according to application requirements.

3.3 Analysis of Dissemination Distance and Hitting Time

3.3.1 Lower bounds

To begin with, we derive lower bounds on dissemination distance and hitting time under a 1-copy message dissemination that the source will be the only disseminator. Originally, the source initiates a geocast with $|D(0)| = 0$, and actively spreads the message while all other recipients carry latent messages without retransmission. The dissemination is solely determined by the mobility of the source node. In other words, the progress of the active message from time $k-1$ to k ($k = 1, 2, \dots$) equals to the movement vector $Y_M(k)$. Hence, at time t, $D(t) = \sum_{k=1}^{t} Y_M(k)$.

Lower bound on dissemination distance

In order to find lower bound on $|D(t)|$, we first examine mobility vector $Y_M(t)$.

Lemma 2 *Under constrained i.i.d. mobility, movement vector $Y_M(t)$ satisfies that $E(|Y_M(t)|) = \frac{2a}{3}$ and $E\left\{|Y_M(t)|^2\right\} = \frac{a^2}{2}$, where $|Y_M(t)|$ is the length of vector $Y_M(t)$ and a is the maximum movement length per time slot.*

Proof : Suppose node i locates at $X_i(t-1)$ and $X_i(t)$ at time $t-1$ and t, respectively. The movement vector $Y_M(t)$ has its origin at $X_i(t-1)$ and endpoint at $X_i(t)$ that is uniformly distributed in $\mathcal{A}(X_i(t-1), a)$. The length of a movement step $|Y_M(t)| = A_i(t) = ||X_i(t) - X_i(t-1)||$.

Let the circular region $\mathcal{A}(X_i(t-1), a)$ be partitioned to m concentric rings indexed by $j, 0 \le j \le m-1$, each of which is with equal ring width ϵ. When ϵ is much smaller than a,

$$P(A_i(t) = x) \approx \frac{2\pi x\epsilon}{\pi a^2} = \frac{2}{a^2} x\epsilon. \tag{3.6}$$

Hence,

$$E(|Y_M(t)|) = E(A_i(t)) = \int_0^a \frac{2}{a^2} x^2 dx = \frac{2a}{3}, \tag{3.7}$$

$$E\left\{|Y_M(t)|^2\right\} = \int_0^a \frac{2}{a^2} x^3 dx = \frac{a^2}{2}. \tag{3.8}$$

Based on Lemma 2, we have

$$E(|D(t)|^2) = \sum_{k=1}^{t} E(|Y_M(k)|^2) = \frac{a^2 t}{2}. \tag{3.9}$$

Since better designed dissemination algorithm can spread out the message faster, $a^2t/2$ can serve as the lower bound of the mean square displacement (MSD) of dissemination distance.

Lower bound on hitting time

We have the following lower bound on probability distribution of hitting time $\tau(d)$.

Theorem 2

$$P(\tau(d) < t) \geq \max\{0, 1 - \frac{4(d+a)^2}{a^2 t}\}$$

Proof : Denote by $D_x(t)$ and $D_y(t)$ the x-component and y-component of dissemination distance vector $D(t)$, respectively. Then,

$$\tau(d) \leq \tau_x(d) \triangleq \inf_{t>0}\{t : |D_x(t)| \geq d\}. \tag{3.10}$$

And $D_x(t) = \sum_{k=1}^{t} Y_M^x(k)$, where $Y_M^x(k)$ is the x-component of mobility vector $Y_M(k)$ as shown in Figure 3.3.

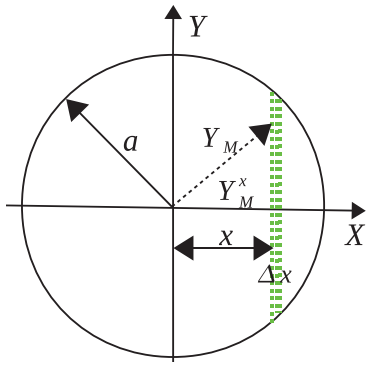

Figure 3.3: Probability distribution of x-component of mobility vector Y_M.

Based on constrained vehicle mobility model, for $-a < x < a$ and small Δx, we have

$$P(x \leq Y_M^x(k) \leq x + \Delta x) \approx \frac{2\sqrt{a^2 - x^2}\Delta x}{\pi a^2}, \tag{3.11}$$

i.e., the green area in Figure 3.3. Clearly, the probability distribution of $Y_M^x(k)$ is a even function, thus $E(Y_M^x(k)) = 0$. As $D_x(t)$ is a generalized 1-D random walk with independent and mean-zero increments $Y_M^x(k)$, $D_x(t)$ is a martingale process with respect to $(Y_M^x(k), k \geq 0)$ according to the definition of martingale. Further, based on Eq. (3.11),

$$E((Y_M^x(k))^2) = 2\int_0^a x^2 \frac{2\sqrt{a^2 - x^2}}{\pi a^2} dx = \frac{a^2}{4}. \tag{3.12}$$

According to Wald's Second Inequality, the stopping time of martingale $D_x(t)$ satisfies

$$E(\tau_x(d)) = \frac{E(D_x^2(\tau_x(d)))}{a^2/4} \leq \frac{4(d+a)^2}{a^2}. \tag{3.13}$$

Based on Markov inequality $P(\tau_x(d) < t) \geq 1 - \frac{E(\tau_x(d))}{t}$. Therefore,

$$P(\tau(d) < t) \geq P(\tau_x(d) < t) \geq \max\{0, 1 - \frac{4(d+a)^2}{a^2 t}\}. \tag{3.14}$$

3.3.2 Upper bounds on 1-Copy Message Dissemination

Next, we move to derive upper bounds of dissemination distance under dissemination strategies that use one or multiple disseminators with or without geographic information assistance in choosing disseminators, respectively. We begin with the baseline case: *1-copy message dissemination strategy* defined in Definition 8.

1-Copy Message Dissemination Distance

Originally, $D(0) = 0$. At time t, $D(t) = \sum_{k=1}^{t} Y(k)$, where $Y(k)$ is the progress of the active message from time $k-1$ to k ($k = 1, 2, \ldots$). As shown in Figure 3.2, $Y(k)$ equals either the movement vector $Y_M(k)$ or $Y_M(k)$ plus the propagation vector $Y_P(t)$. In order to derive dissemination distance $D(t)$, we study random variables $Y_M(k)$ and $Y_P(k)$, respectively in Lemmas 2 and 3.

Lemma 3 *Propagation vector $Y_P(t)$ satisfies that*

$$E(|Y_P(t)|) \leq r(1 - e^{-\lambda\pi r^2}), \tag{3.15}$$

$$E\left\{|Y_P(t)|^2\right\} \leq r^2 - \frac{1}{\lambda\pi}(1 - e^{-\lambda\pi r^2}), \tag{3.16}$$

where $|Y_P(t)|$ is the length of propagation vector $Y_P(t)$, r is transmission range of nodes, and λ is node density.

Proof : Assume node v_i locates at $X_i(t) = 0$ at time t and $N_i(t)$ is the set of v_i's neighbors. Denote by random variable

$$S(t) = \max_{v_j \in N_i(t)}\{||X_i(t) - X_j(t)||\} \tag{3.17}$$

the maximum distance between node v_i and its neighbors $N_i(t)$ at time t. Apparently, the length of propagation vector, denote by $|Y_P(t)|$, is equal or less than $S(t)$ and the equality holds if

only if v_i chooses its furthest neighbor as next-hop disseminator (in other words, active message jumps from v_i to its furthest neighbor). Hence, for $0 \leq x \leq r$,

$$P(|Y_P(t)| \geq x) \leq P(S(t) \geq x), \tag{3.18}$$

i.e., random variable $S(t)$ stochastically dominates $|Y_P(t)|$.

Let $A_m(x, y)$ denote the event that there exist m neighbors in area $\{s : x \leq ||s - X_i(t)|| \leq y\}$, $(0 \leq x < y \leq r)$. As n nodes are Poisson distributed in the network with density λ everywhere at all times, $P(A_m(x, y)) = e^{-\lambda\pi(y^2-x^2)} \frac{[\lambda\pi(y^2-x^2)]^m}{m!}$. For $0 \leq x \leq r$,

$$P(S(t) \geq x) = 1 - P(A_0(x, r)) = 1 - e^{-\lambda\pi(r^2-x^2)}. \tag{3.19}$$

Hence,

$$\begin{aligned} E(S(t)) &= r - e^{-\lambda\pi r^2} \int_0^r e^{\lambda\pi x^2} dx \leq r(1 - e^{-\lambda\pi r^2}) \\ E(S^2(t)) &= \int_0^r x^2 (e^{-\lambda\pi(r^2-x^2)})' dx = r^2 - \frac{1 - e^{-\lambda\pi r^2}}{\lambda\pi}. \end{aligned} \tag{3.20}$$

Combining (3.18) and (3.20) completes our proof.

Remark 17 *Lemma 3 shows that the largest expected length of propagation vector increases as transmission range r and node density λ increase, which provide better chance of finding the next active spreading node located far from the previous active spreading node.*

Because of the intermittent network connectivity in VANETs, the active message travels a journey in the network area through movements and transmissions of disseminators. Hence, dissemination distance $|D(t)|$ is affected not only by distributions of $Y_M(k)$ and $Y_P(k)$ $(1 \leq k \leq t)$ but also by number of jumps of an active message within time t, which is defined as $\mathcal{N}(t) = \sum_{k=1}^{t} 1_{Y(k)=Y_M(k)+Y_P(k)}$. As transmission between two disseminators occurs when the preceding disseminator meets its succeeding disseminator, $\mathcal{N}(t)$ is determined by intermittent connectivity of VANETs as well as dissemination algorithms. Rather than limiting our study on specific algorithms, we focus on the impact of intermittent connectivity. More specifically, we derive upper bound for $\mathcal{N}(t)$ by investigating the *first passage time*, which is defined as the time when two nodes first meet.

Lemma 4 *In a large scale network that L goes to infinity, suppose two independent nodes v_i and v_j move according to the constrained i.i.d. mobility model in Definition 7. Then, there exists constant $C > 0$ that the first passage time T_F of nodes v_i and v_j satisfies*

$$P(T_F > t) \geq Ct^{-\alpha}, \text{ for all sufficiently large } t, \tag{3.21}$$

where α is a constant determined by mobility model and $\alpha = \frac{1}{2}$ in constrained i.i.d. mobility.

Proof : We prove this lemma using the same methodology that is used to derive inter-meeting time of two nodes under 2-D isotropic random walk in [60]. According to Definition 7, the position of a node v at time t can be written as

$$X_v(t) - X_v(0) = \sum_{k=1}^{t} Y_M^v(k) = \sum_{k=1}^{t} A_v(k) e^{i\theta_v(k)}, \tag{3.22}$$

where $Y_M^v(k)$ is the movement vector at time k. Define by $C(t) = X_u(t) - X_v(t)$ the difference vector between the positions of nodes u and v at time t. Under 2-D constrained i.i.d. node mobility, we observe that

$$C(t) = \sum_{k=1}^{t} (A_u(k) e^{i\theta_u(k)} - A_v(k) e^{i\theta_v(k)}). \tag{3.23}$$

Under constrained i.i.d. mobility, $A_u(k)$ and $A_v(k)$ are all i.i.d. and so are $\theta_v(k)$ and $\theta_v(k)$. Thus, $A_u cos(\theta_u) - A_v cos(\theta_v)$ is symmetric and continuous (because uniform distribution is continuous). Accordingly, $[C(t)]_x$, sum of random variables $A_u(k) cos\theta_u(k) - A_v(k) cos\theta_v(k)$ for $1 \leq k \leq t$, is 1-D random walk. According to results in [60], $P(T_F > t) \sim t^{-1/2}$, thus completes our proof.

Now, we can derive upper bound on distribution of $\mathcal{N}(t)$.

Theorem 3 *$\mathcal{N}(t)$ is stochastically dominated by Poisson process with parameter α and $E(\mathcal{N}(t)) \leq \alpha t$, where α is a constant determined by mobility model and $\alpha = 1/2$ under 2-D constrained i.i.d. mobility.*

Proof : Denote by T the time interval that a disseminator transmits the active message to its next-hop disseminator. Clearly, random variable T stochastically dominates their first passage time T_F, which means $P(T > t) \geq P(T_F > t)$. Based on Lemma 4, there exists constant $C > 0$ such that

$$P(T > t) \geq Ce^{-\alpha t}, \text{ for all sufficiently large } t. \tag{3.24}$$

Therefore, number of transmissions among disseminators, $\mathcal{N}(t)$, is stochastically dominated by Poisson process with parameter α, and $E(\mathcal{N}(t)) \leq \alpha t$ accordingly, where $\alpha = 1/2$ under 2-D constrained i.i.d. mobility model.

As T_F corresponds to the residual life time of the *inter-meeting time* T_I, when T_I has a finite mean, T_F has the equilibrium distribution of T_I [88]. Extensive existing studies have shown that T_I exhibits exponential tail decay under existing mobility models (such as random direction, random waypoint and Brownian Motion) in a bounded domain, while power-law

decay in empirically traces as well as infinite domain (See [60] and references inside). Because Eq. (3.24) holds under exponential as well as power-law decayed T_F with α determined by specific mobility model, Theorem 3 likely holds under other mobility models and in realistic traces.

Equipped with previous results on movement vector $Y_M(t)$, propagation vector $Y_P(t)$ and $\mathcal{N}(t)$, we are ready to analyze dissemination distance for both *direction-invariant* dissemination and *geographic-assisted* dissemination that are mentioned in Remark 15 and Remark 16.

Theorem 4 *The mean of dissemination distance, $E(|D(t)|)$, is upper bounded by function $\sqrt{tf_1(r,\lambda,a,\alpha)}$ in 1-copy direction-invariant dissemination, by function $\sqrt{tf_2(r,\lambda,a,\alpha,t)}$ in 1-copy geographic-assisted dissemination, where $f_1(r,\lambda,a,\alpha)$ and $f_2(r,\lambda,a,\alpha,t)$ are shown in Eqs. (3.28) and (3.31).*

Proof : Denote $D(t) = (D_x(t), D_y(t)) = (\sum_{k=1}^{t} Y_x(k), \sum_{k=1}^{t} Y_y(k))$. Define by $Z(k)$ the event of active message jump at time k (i.e., $Y(k) = Y_M(k) + Y_P(k)$). Accordingly, $\mathcal{N}(t) = \sum_{k=1}^{t} 1_{Z(k)}$.

(i) Direction-invariant dissemination, which means $E(Y_P^x(k)) = E(Y_P^y(k)) = 0$.

Because $E(Y_M^x(k)) = E(Y_M^y(k)) = 0$ under constrained i.i.d. mobility and $Y_M(k)$ and $Y_P(k)$ are independent,

$$E\{D_x^2(t)\} = E\left\{\sum_{k=1}^{t} [Y_M^x(k)]^2 + \left[Y_P^x(k)\cdot 1_{Z(k)}\right]^2\right\} = tE\left\{|Y_M^x(k)|^2\right\} + E(\mathcal{N}(t))E\left\{|Y_P^x(k)|^2\right\}. \tag{3.25}$$

Similar results can be obtained for $E\{D_y^2(t)\}$.

$$E(|D(t)|^2) = E\{D_x^2(t) + D_y^2(t)\} = tE\left\{|Y_M(k)|^2\right\} + E(\mathcal{N}(t))E\left\{|Y_P(k)|^2\right\}. \tag{3.26}$$

Based on results in Lemmas 2, 3, and Theorem 3, we get

$$E^2(|D(t)|) \le E(|D(t)|^2) \le t[a^2/2 + \alpha r^2 - \frac{\alpha}{\lambda\pi}(1 - e^{-\lambda\pi r^2})]. \tag{3.27}$$

Therefore, by denoting

$$f_1(r,\lambda,a,\alpha) = a^2/2 + \alpha r^2 - \frac{\alpha}{\lambda\pi}(1 - e^{-\lambda\pi r^2}), \tag{3.28}$$

we have $E(|D(t)|) \le \sqrt{tf_1(r,\lambda,a,\alpha)}$.

(ii) Geographic-assisted dissemination, which means $E(Y_P^x(k)) = E(Y_P^y(k)) \neq 0$.

As $Y_M(k)$ and $Y_P(k)$ are independent and $E(Y_M^x(k)) = 0$,

$$E\left\{D_x^2(t)\right\} = E\left\{\sum_{k=1}^{t}\left[\left(Y_M^x(k)\right)^2 + \left(Y_P^x(k)\cdot 1_{Z(k)}\right)^2\right]\right\} + E\left\{\sum_{1\le k_1<k_2\le t}\left[Y_P^x(k_1)\cdot 1_{Z(k_1)}\cdot Y_P^x(k_2)\cdot 1_{Z(k_2)}\right]\right\}. \tag{3.29}$$

Similar result can be obtained for $E\left\{D_y^2(t)\right\}$. Hence,

$$E(|D(t)|^2) \le tE\left\{|Y_M(k)|^2\right\} + \left(E(\mathcal{N}(t)) + \frac{t(t-1)}{2}\right) E\left\{|Y_P(k)|^2\right\}. \tag{3.30}$$

In view of Lemmas 2, 3, and Theorem 3, and by denoting

$$f_2(r,\lambda,a,\alpha,t) = \frac{a^2}{2} + (\alpha + (t-1)/2)\left(r^2 - \frac{1-e^{-\lambda\pi r^2}}{\lambda\pi}\right), \tag{3.31}$$

$$E^2(|D(t)|) \le E(|D(t)|^2) \le tf_2(r,\lambda,a,\alpha,t). \tag{3.32}$$

Therefore, $E(|D(t)|) \le \sqrt{tf_2(r,\lambda,a,\alpha,t)}$.

Theorem 4 shows that the upper bound of $E(|D(t)|)$ depends on node *velocity* (indicated by maximum movement length a per time slot), *mobility model* (represented by α), node *transmission range* r, and node *density* λ. Furthermore, the expected dissemination distance can at most increase with the *square root* of time t in *direction-invariant* dissemination while approximately *linearly* with time t in *geographic-assisted* dissemination. The upper bound on dissemination distance under 1-copy geographic-assisted dissemination is approximately $\sqrt{t}$ times of that under 1-copy direction-invariant dissemination. In other words, comparing to direction-invariant dissemination, the increase in dissemination distance of utilizing geographic information accumulates as time goes by.

1-Copy Message Dissemination Hitting Time

As dissemination distance vector $D(t) = \sum_{k=1}^{t} Y(k)$, i.e., sum of i.i.d. random variables, we use martingale theory to study the hitting time $\tau(d)$.

Lemma 5 *Dissemination distance* $\{|D(t)|^2\}_{t\in\mathbb{N}}$ *is a submartingale with respect to Filtration* $\mathcal{F}_t$*, which is the* σ*-algebra generated by* $\{D(k); k \le t\}$ *for every* $t \in \mathbb{N}$.

Proof : We prove that $|D(t)|^2$ is a submartingale by proving that $(D_x^2(t), D_y^2(t))$ is a 2D sub-martingale according to the following definition. A sub-martingale is defined as an integer-time stochastic process $\{Z_n; n \geq 1\}$ with the properties that $E[|Z_t|] < \infty$ for all $t \geq 1$ and

$$E[Z_t|Z_{t-1}, Z_{t-2}, \ldots, Z_1] \geq Z_{t-1}; \text{ for all } t \geq 2. \tag{3.33}$$

Denote dissemination vector $D(t) = (D_x(t), D_y(t)) = (\sum_{k=1}^{t} Y_x(k), \sum_{k=1}^{t} Y_y(k))$, where $Y(k) = (Y_x(k), Y_y(k))$ equals either $Y_M(k)$ or $Y_P(k) + Y_M(k)$.

(i) Due to limited transmission range r and movement length a, $|Y_x(k)| \leq r + a < \infty$. Then, for any $k \in \mathbb{N}$.

$$|D_x(t)| \leq |Y_x(1)| + \cdots + |Y_x(t)| \leq t \times (r + a) < \infty. \tag{3.34}$$

Similarly, $|D_y(t)| \leq t \times (r + a) < \infty$.

(ii) Then we prove Eq. (3.33), which distinguishes martingale from other processes. Assume node v_i is the disseminator at time $t-1$. Denote filtration $\mathcal{F}$ of process $\{D(t)\}$ as $\mathcal{F}_t = \sigma$-algebra generated by $\{D(k); k \leq t\}$ for every t. First, for any $t \in \mathbb{N}$, it holds that $E(D_x(t)|\mathcal{F}_{t-1}) = D_x(t-1) + E(Y_x(t))$ and $E(D_y(t)|\mathcal{F}_{t-1}) = D_y(t-1) + E(Y_y(t))$.

(a) When $Y(t) = Y_M(t)$, $Y_x(t) = Y_M^x(t) = A(t)cos(\theta(t))$ and $Y_y(t) = Y_M^y(t) = A(t)sin(\theta(t))$. In constrained i.i.d. mobility, $P(Y_M^x(t) = z) = P(Y_M^x(t) = -z)$ and $P(Y_M^y(t) = z) = P(Y_M^y(t) = -z)$ $(0 \leq z \leq a)$. Hence, $E(Y_x(t)) = E(Y_M^x(t)) = 0$ and $E(Y_y(t)) = E(Y_M^y(t)) = 0$.

(b) When $Y(t) = Y_P(t) + Y_M(t)$, for direction-invariant dissemination, $E(Y_P^x(t)) = E(Y_P^y(t)) = 0$; for geographic-assisted dissemination, $E(Y_P^x(t)) \geq 0$ if $D_x(t) \geq 0$ and $E(Y_P^x(t)) \leq 0$ if $D_x(t) \leq 0$ (the same for $Y_P^y(t)$).

From (a) and (b), when $D_x(t) \geq 0$, $E(D_x(t)|\mathcal{F}_{t-1}) \geq D_x(t-1)$, which proves that $D_x(t)$ is submartingale. When $D_x(t) < 0$, $E(-D_x(t)|\mathcal{F}_{t-1}) \geq -D_x(t-1)$, which means that $-D_x(t)$ is submartingale. As $D_x^2(t) = D_x(t) * D_x(t) = (-D_x(t)) * (-D_x(t))$ and square function is convex, $D_x^2(t)$ is a submartingale. Similarly $D_y^2(t)$ is also a submartingale. Therefore, $|D(t)|^2 = D_x^2(t) + D_y^2(t)$ is a submartingale.

Based on Lemma 5, we have the following theorem.

Theorem 5 *In a geocast, the probability of $\tau(d) \leq t$ satisfies, i) for direction-invariant dissemination,*

$$P(\tau(d) \leq t) \leq \frac{E(|D(t)|^2)}{d^2} \leq \frac{t}{d^2} f_1(r, \lambda, a, \alpha); \tag{3.35}$$

ii) for geographic-assisted dissemination,

$$P(\tau(d) \leq t) \leq \frac{E(|D(t)|^2)}{d^2} \leq \frac{t}{d^2} f_2(r, \lambda, a, \alpha, t), \tag{3.36}$$

where $f_1(r, \lambda, a, \alpha)$ and $f_2(r, \lambda, a, \alpha, t)$ are shown in Eqs. (3.28) and (3.31), respectively.

Proof : We proceed to find distribution of hitting time $\tau(d)$ using *Doob's Submartingale Maximal Inequality*, which is that for $(|D(k)|^2)_{k\in\mathbb{N}}$ being a non-negative sub-martingale with respect to a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$, for any $d > 0$,

$$P(\max_{1\leq k\leq t}|D(k)|^2 \geq d^2) \leq \frac{1}{d^2}E(|D(t)|^2). \tag{3.37}$$

Based on definitions of dissemination distance in Eq. (3.4) and hitting time in Eq. (3.5), we have $\{\max_{1\leq k\leq t}|D(k)|^2 \geq d^2\} = \{\tau(d) \leq t\}$. By applying Doob's Submartingale Maximal Inequality to the submartingale $|D(t)|^2$, we have

$$P(\tau(d) \leq t) \leq \frac{1}{d^2}E(|D(t)|^2). \tag{3.38}$$

(i) Direction-invariant dissemination: the proof of Theorem 4 shows that $E[|D(t)|^2] \leq tf_1(r,\lambda,a,\alpha)$, where $f_1(r,\lambda,a,\alpha)$ is shown in Eq. (3.28). Hence, we prove Eq. (3.35).

(ii) Geographic-assisted dissemination: the proof of Theorem 4 shows that $E(|D(t)|^2) \leq tf_2(r,\lambda,a,\alpha,t)$, where $f_2(r,\lambda,a,\alpha,t)$ is shown in Eq. (3.31). Thus, we prove Eq. (3.36).

Remark 18 *The probability that a 1-copy message dissemination reaches nodes or infrastructures located distance d from the source within time t is upper bounded by a function proportional to $E(|D(t)|^2)$ and inversely proportional to d^2.*

3.3.3 Upper bounds on L-Copy Message Dissemination

In this section, we extend our results from 1-copy message dissemination to the more general message dissemination strategy (i.e., L-copy message dissemination).

L-Copy Message Dissemination Distance

In L-copy message dissemination, number of disseminator is equal or less than L, i.e., $|\mathcal{V}(t)| \leq L$. Denote by $|D_i(t)|$ the distance between the source location and the location of the i^{th} disseminator at time t and denote by dissemination distance $|D^L(t)|$ the maximum of $|D_i(t)|$.

Theorem 6 *For L-copy direction-invariant dissemination, $E(|D^L(t)|) \leq (\sqrt{L-1}+1)\times\sqrt{tf_1(r,\lambda,a,\alpha)}$; for L-copy geographic-assisted dissemination, $E(|D^L(t)|) \leq (\sqrt{L-1}+1)\times\sqrt{tf_2(r,\lambda,a,\alpha,t)}$, where $f_1(r,\lambda,a,\alpha)$ and $f_2(r,\lambda,a,\alpha,t)$ are shown in Eqs. (3.28) and (3.31), respectively.*

Proof : To analyze $|D^L(t)|$), which equals the maximum of several random variables, we introduce Aven's [89] upper bound on the mean of the maximum of a number of random variables

$\{Z_i, 1 \leq i \leq L\}$ with general distributions (not necessarily independent and identically distributed).

$$E(\max_{1\leq i\leq L} Z_i) \leq \max_{1\leq i\leq L} E(Z_i) + \sqrt{\frac{L-1}{L}} \left(\sum_{i=1}^{L} \mathrm{Var}(Z_i)\right)^{1/2}. \tag{3.39}$$

Applying the above equation to $|D^L(t)|$,

$$E(|D^L(t)|) = E(\max_{v_i\in\mathcal{V}(t)} \{|D_i(t)|\}) \leq \max_{1\leq i\leq L} E(|D_i(t)|) + \sqrt{\frac{L-1}{L}} \left(\sum_{i=1}^{L} E(|D_i(t)|^2)\right)^{1/2} \tag{3.40}$$

Based on the proof and results of Theorem 4, for L-copy direction-invariant dissemination,

$$E(|D^L(t)|) \leq (\sqrt{L-1}+1)\sqrt{tf_1(r,\lambda,a,\alpha)}; \tag{3.41}$$

for L-copy geographic-assisted dissemination,

$$E(|D^L(t)|) \leq (\sqrt{L-1}+1)\sqrt{tf_2(r,\lambda,a,\alpha,t)}. \tag{3.42}$$

Remark 19 *Clearly, using multiple disseminators can increase the dissemination distance. Theorem 6 shows that the upper bounds on the expected dissemination distance in L-copy message dissemination are $\sqrt{L-1}$ times larger than those in corresponding 1-copy message dissemination in Theorem 4.*

Regarding 1-copy direction-invariant dissemination as a base line, the upper bound of dissemination distance at time t increases $\sqrt{L-1}$ times under L-copy direction-invariant dissemination, increases approximately $\sqrt{t}$ times under 1-copy geographic-assisted dissemination. This means that using multiple disseminators may benefit dissemination more at the beginning while as time goes on, utilizing geographic information tends to gain more benefits. Hence, multiple disseminators should be used for geocast applications with AOI near the source, while for AOI far from the the source, geographic information should be used to choose relays in order to enhance dissemination distance.

Remark 20 *In order to increase the dissemination distance, multiple disseminators should be used at the beginning, and geographic information is used preferably as time goes on.*

L-Copy Message Dissemination Hitting Time

The hitting time $\tau^L(d)$, i.e., the first time that $|D^L(t)|^2 \geq d^2$, satisfies the following theorem.

Theorem 7 *For L-copy direction-invariant dissemination, $P(\tau^L(d) \leq t)$ is upper bounded by*

$$\frac{t}{d^2}\left(f_1(r,\lambda,a,\alpha) + \sqrt{L-1}(r+a)\sqrt{f_1(r,\lambda,a,\alpha)}\right); \tag{3.43}$$

for L-copy geographic-assisted dissemination, $P(\tau(d) \leq t)$ is upper bounded by

$$\frac{t}{d^2}\left(f_2(r,\lambda,a,\alpha,t) + \sqrt{L-1}(r+a)\sqrt{f_2(r,\lambda,a,\alpha,t)}\right). \tag{3.44}$$

Proof : According to L-copy message dissemination in Definition 9, the message is first spread to L distinct disseminators and then each of disseminators independently disseminates the message according to 1-copy message dissemination. Define $|D^{L^*}(t)|$ as the dissemination distance that L disseminators start to independently disseminate the message according to 1-copy message dissemination from time $t = 0$. As $|\mathcal{V}(t)| \leq L$, $|D^L(t)|^2 = \max_{i\in\mathcal{V}(t)}\{|D_i(t)|^2\} \leq |D^{L^*}(t)|^2 = \max_{i=1}^{L}\{|D_i^*(t)|^2\}$. Thus,

$$\tau^L(d) = \inf_{t\geq 1}\{|D^L(t)| \geq d\} \geq \tau^{L^*}(d) = \inf_{t\geq 1}\{|D^{L^*}(t)| \geq d\}. \tag{3.45}$$

Upon Lemma 5, $\{|D_i^*(t)|^2, 1 \leq i \leq L\}$ are independent sub-martingales. Hence, $|D^{L^*}(t)|^2 = \max_{i=1}^{L}\{|D_i(t)|^2\}$ is a sub-martingale. Based on Doob's Submartingale Maximal Inequality in Eq. 3.38,

$$P(\tau^{L^*}(d) \leq t) = P(\max_{1\leq k\leq t}|D^{L^*}(k)|^2 \geq d^2) \leq \frac{1}{d^2}E(|D^{L^*}(t)|^2). \tag{3.46}$$

Using Aven's [89] upper bound on the mean of the maximum of a number of random variables in Eq. (3.39), we have

$$E(|D^{L^*}(t)|^2) = E(\max_{1\leq i\leq L}\{|D_i^*(t)|^2\}) \leq \max_{1\leq i\leq L}E(|D_i^*(t)|^2) + \sqrt{\frac{L-1}{L}}\left(\sum_{i=1}^{L}\mathrm{Var}(|D_i^*(t)|^2)\right)^{1/2}. \tag{3.47}$$

Denote $Z = \frac{|D_i^*(t)|^2}{(r+a)^2 t}$. Clearly, $0 \leq Z \leq 1$. Thus,

$$\mathrm{Var}(Z) = E(Z^2) - E^2(Z) \leq E(Z)(1-E(Z)) \leq E(Z), \tag{3.48}$$

$$\mathrm{Var}(|D_i^*(t)|^2) = (r+a)^4 t^2 \mathrm{Var}(Z) \leq (r+a)^2 t E(|D_i^*(t)|^2). \tag{3.49}$$

Based on the proof of Theorem 4 and combining Eqs. (3.45), (3.46), and (3.49), we complete our proof.

Remark 21 *Compared with 1-copy dissemination, L-copy dissemination can reduce dissemination latency as it can increase the probability that reaches nodes or infrastructures located d distance from the source within time t. Note that it is not known whether these upper bounds*

are achievable. The upper bounds may not be achievable by any algorithm in reality since the analysis is not based on realistic network and mobility models.

3.4 Simulation Results and Applications

Many VANET applications require position-based multicasting (e.g., for disseminating traffic information to vehicles approaching the current position of the source). A natural match for this type of routing is the geocasting protocols that forward messages to all nodes within a Area of Interest (AOI). Previous research work on geocast schemes for vehicular networks has mostly proposed various flooding schemes. One problem with a pure flooding-based geocasting protocol is that the flooding can cause network congestion [81]. Therefore, selective flooding may be used in which the forwarding is based on an intelligent decision that should maximize the spreading of the message at the same time as it minimizes the network load caused by the message spreading. Apparently, *limiting number of disseminators* and *geographic information exchanges* are two effective methods to reduce network load caused by the message spreading.

However, there is a trade-off between minimizing network load by limiting number of disseminators and geographic information exchanges and maximizing the spreading of the message through increasing number of disseminators and exchanging geographic information to select nodes with the most forward progress towards the destination as relays. It is not clear *how many disseminators are needed and whether geographic information exchanges should be used* for selecting relays, which depend on dissemination mechanism performance as well as application requirements.

Therefore, we perform simulations using four dissemination algorithms that use one or multiple disseminators and choose disseminators randomly or based on geographic information. Along with the simulation results and application requirements, we intend to provide guidelines on design of dissemination algorithms for geocast in VANET. In this section, we first present and implement four dissemination algorithms in OmNet++ and compare simulation results with the analytic bounds presented in the previous sections. Then, we present two geocast scenarios to demonstrate the applications of our results.

3.4.1 Simulation Results

Dissemination Algorithms

In *stateless opportunistic forwarding (SOF)* [82], a disseminator will choose next disseminator from its available neighbors at random. Stateless opportunistic forwarding has been suggested to be useful in intermittently connected networks [90–93]. It is particularly useful in vehicular ad hoc network as its global network topology is not known and rapidly varying due to high

vehicle mobility and the presence or availability of the next-hop neighbors is not easily controllable. Clearly, SOF chooses next disseminator isotropically, thus is a type of *direction-invariant* dissemination. The SOF with one disseminator at each time slot is referred to as *1-copy SOF*. Similarly, dissemination algorithm that first sprays active messages to L disseminators and then each disseminator performs SOF independently, is referred to as *L-copy SOF*.

In *GPS-based broadcasting (GBB)* [83], a disseminator will choose its farthest neighbor as next disseminator so that the message can be spread out as fast as possible to certain locations (e.g., police station). GBB is useful for disseminating time-critical message (such as emergency warning) in VANETs. Apparently, GBB is an example of *geographic-assisted* dissemination. The GBB with one disseminator is referred as *1-copy GBB*. Similarly, in *L-copy GBB*, source node first sprays active messages to L disseminators and then each disseminator performs GBB independently.

Constrained vehicle mobility

Because currently there is no single benchmark of D2D communication network scenarios to evaluate its performance [94], we choose simulation parameters that are close to realistic network scenario and IEEE 1609/802.11p standards [95] for VANETs. In a 10 km$\times$10 km network area, size of a university, 5000 nodes move according to constrained i.i.d. mobility model. Node density is $\lambda = 5 \times 10^{-5}$ vehicle/m^2. The average number of neighbors for each vehicle is equal to or less than $\pi * 200 * 200 * 5000/(10000 * 10000) = 2\pi$. As simulations in [96] suggested that six to eight neighbors can make a small size network connected with high probability, the above simulation settings will produce a intermittently connected network. Using the standard simulator 802.11p in OMNeT++ INET framework, we set *carrierFrequency = 5.9GHz*, *wlan.opMode = "p"*, *bitrate = 27Mbps*, and *messageLength = 512B*. The transmission range of node is $R = 200$m. Each time slot is 1 second and the maximum movement length $a = 20$m per time slot, which means that speed limit is about 45mph. $L = 1$ or 4 for L-copy SOF and GBB.

As shown in Figure 3.4, average dissemination distances of 1-copy SOF and 1-copy GBB are upper bounded by bounds of expected dissemination distances of 1-copy direction-invariant and geographic-assisted dissemination in Theorem 4, respectively. Similarly, Figure 3.5 shows that average dissemination distance of 4-copy SOF and 4-copy GBB are upper bounded by bounds of expected dissemination distances of L-copy direction-invariant and geographic-assisted dissemination in Theorem 6, respectively. In a word, the average dissemination distances of above four algorithms are well upper bounded by their corresponding analytic upper bounds.

Since more sophisticated algorithms could achieve better performance, the simulation results of SOF and GBB could serve as the lower bounds for direction-invariant and geographic-assisted dissemination, respectively. As shown in Figures 3.4 and 3.5, the upper bounds of expected dis-

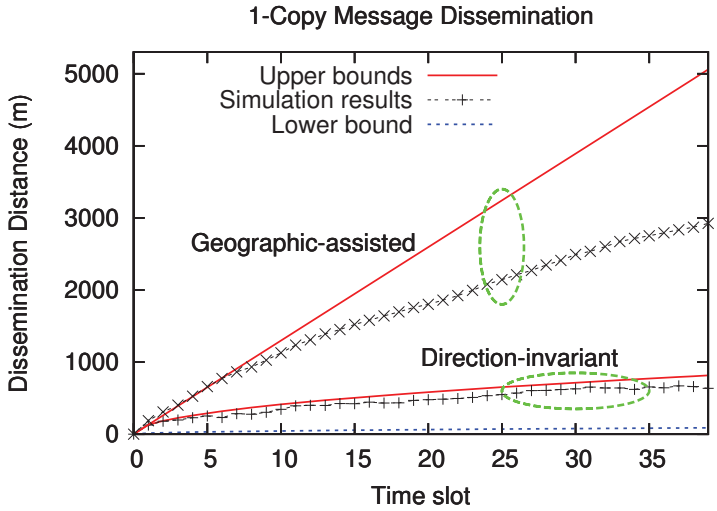

Figure 3.4: Dissemination distance $|D(t)|$ of 1-copy direction-invariant and geographic-assisted message dissemination, respectively.

semination distance under direction-invariant dissemination is tight, while there is a wide gap between the performance of GBB and the upper bounds under geographic-assisted dissemination. The gap could be lessened by more sophisticated algorithms, such as those choosing nodes that move away from the message source as relays.

Examining more closely, Figures 3.6 and 3.7 show that the expected dissemination distances of direction-invariant dissemination algorithms (e.g., 1-copy and 4-copy SOF) exhibit square root increase with time t, while that of geographic-assisted dissemination algorithms (e.g., 1-copy and 4-copy GBB) achieve approximately *linear* increase as time eclipses.

In addition, both Figures 3.4 and 3.5 reveal that comparing to direction-invariant dissemination, geographic-assisted dissemination significantly increases dissemination distance by utilizing nodes' geographic information. Increasing number of disseminators, although may benefit the dissemination reliability, is less effective than incorporating geographic-assisted dissemination in terms of enhancing dissemination distance.

Figures 3.8 and 3.9 show that simulation results of $P(\tau(d) \leq t)$ of four dissemination algorithms are well bounded by corresponding analytic bounds in Theorem 5 and Theorem 7. Both figures demonstrate benefits of utilizing geographic information in greatly reducing hitting time. However, geographic information becomes less effective in reducing hitting time when multiple disseminators are used than when one disseminator is used. In other words, increasing number of disseminators seems reduce hitting time in direction-invariant dissemination more dramatically than that in geographic assisted dissemination.

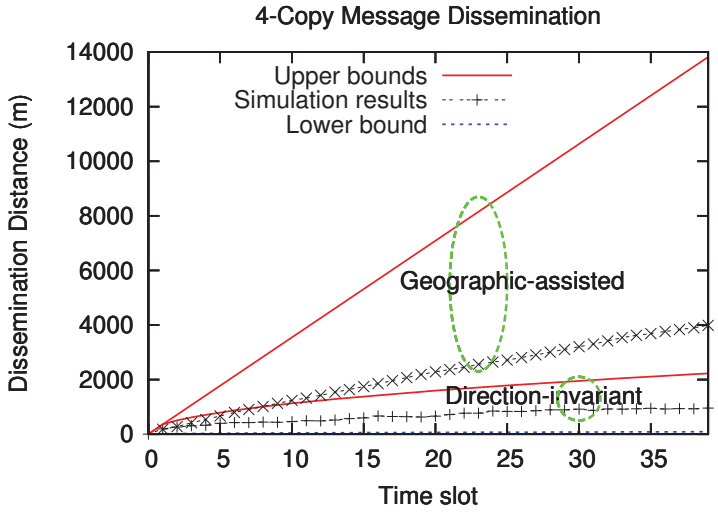

Figure 3.5: Dissemination distance $|D(t)|$ of L-copy ($L = 4$) direction-invariant and geographic-assisted message dissemination, respectively.

Highway mobility

In a 20 km×16 m rectangular network area, 200 nodes move according to highway mobility model. This setting characterizes a stretch of highway that has 4 lanes with each lane width 4 meters. As message dissemination on highway may affect several miles to dozens of miles, 20 km network length is suitable for studying message dissemination in highway scenario. Each time slot is 1 second. Each vehicle chooses a moving direction (left or right) at time 0 and will not change its moving direction at any time $t > 0$. The vehicle speed is uniformly distributed in $[25m/s, 35m/s]$ (approximately 55∼80 mph). The transmission range $R = 250$ meters. The average number of neighbors for each vehicle is equal to or less than $2*250*200/20000 = 5$. Such simulation settings will produce an intermittently connected network with high probability [96].

Figure 3.10 shows the simulation results and theoretical bounds on dissemination distance of 1-copy message dissemination strategy in a highway scenario. Dissemination distances of stateless opportunistic forwarding and GPS-based broadcasting are upper bounded by the results on that of direction-invariant strategy and geographic-assisted strategy, respectively. Moreover, a close look at the results on stateless opportunistic forwarding in Figure 3.11 shows that dissemination distance increases with square root of time t without using geographic information for disseminator selection. If Geographic information is used to speed up the dissemination, the dissemination distance increases linearly with time t, as shown in Figure 3.10. These results are consistent with our theoretical analysis as well as results in a scenario with constrained vehicle mobility.

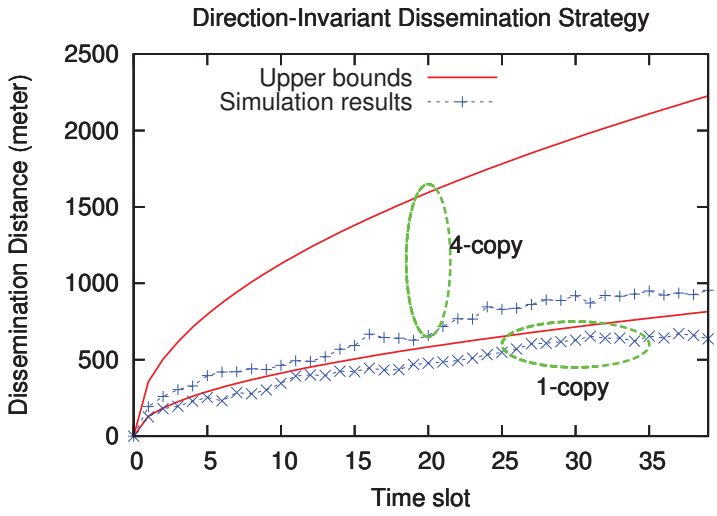

Figure 3.6: Dissemination distance of 1-copy/4-copy direction-invariant dissemination strategies.

In the following, by setting parameters according to specific realistic applications, we further demonstrate how our analytical bounds and simulation results serve as guidelines to choose dissemination strategy and decide number of disseminators such as to satisfy application requirements.

3.4.2 Applications

We give two important applications of VANETs, which are *post-crash warning* and *emergency vehicle signal preemption* to show how our results could be used in guiding the network design. The application requirements are obtained from vehicle safety communication project report [97] by National Highway Traffic Safety Administration in Department of Transportation of US. In the following, we assume the time is slotted and each time interval is 1 sec. The node density and node transmission range are assumed to be the same as our simulation settings in the previous subsection.

Post-Crash Warning

In the application of post-crash warning, a disabled vehicle (due to an accident or mechanical breakdown) will warn approaching vehicles of its position and will discontinue broadcast when the accident is cleared. According to report [97], the allowable latency for this application is approximately 5 seconds.

Suppose vehicle speed is about 20m/s and drivers need about 1.5sec to react and 3sec to

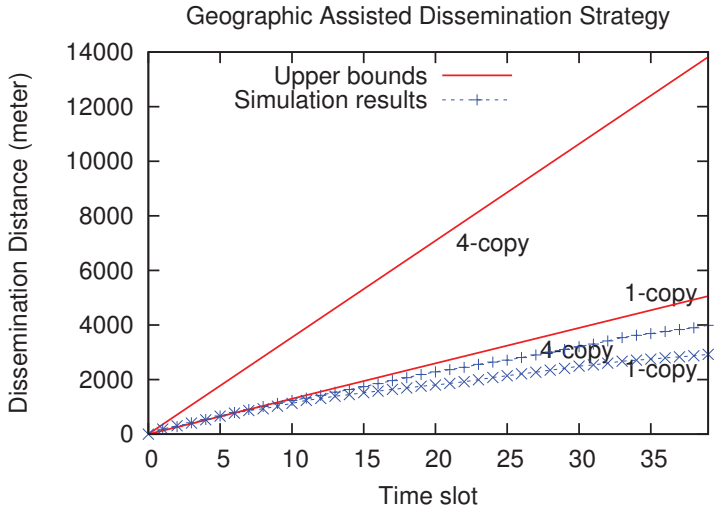

Figure 3.7: Dissemination distance of 1-copy/4-copy geographic-assisted dissemination strategies.

brake if it is necessary, which mean that the geocast needs to target vehicles within range of about 100m. Simulation results (Figures 3.4 and 3.5) show that in 5 time slots, dissemination distances are about 252m, 396m, 660m, and 669m for 1-copy and 4-copy SOF, 1-copy and 4-copy GBB, respectively. Since more sophisticated algorithms could achieve better performance than SOF and GBB algorithms, both direction-invariant and geographic-assisted algorithms could disseminate the message fast enough to reach the borders of targeted area.

As propagation speed is fast enough, dissemination strategy should focus on achieving the reliability requirement of this safety application. As simultaneous rebroadcasting of multiple disseminators can enhance the probability of vehicles receiving this warning, L-copy message dissemination strategy could be a good candidate in fast and reliably post-crash warning dissemination.

Emergency Vehicle Signal Preemption

Emergency vehicle signal preemption allows the emergency vehicles to override traffic signals. When an emergency vehicle is approaching an intersection, it initiates a geocast targeting vehicles around that intersection. After receiving the message and verifying that the request has been made by an authorized source, the vehicles around the intersection should prepare to stop and provide the right of way to the emergency vehicle.

As an example, we give specific and reasonable data to illustrate a scenario of this application. We assume that the geocast targets vehicles in the circular region around the intersection

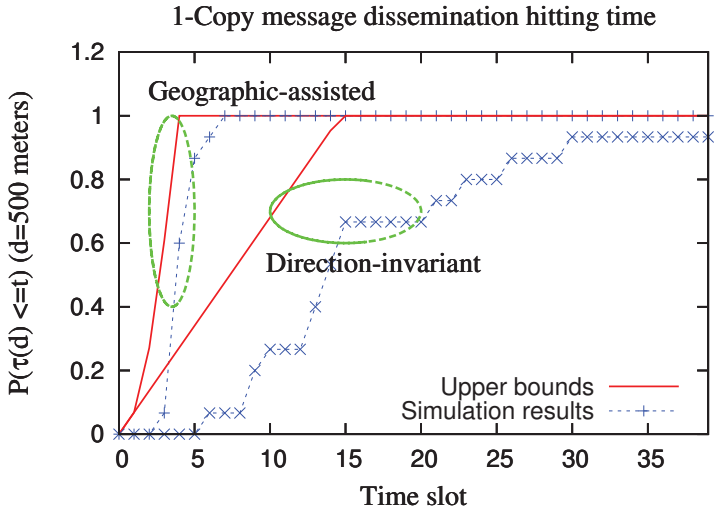

Figure 3.8: $P(\tau(d) \leq t)$ ($d = 500$m) in 1-copy direction-invariant and geographic-assisted message dissemination, respectively.

with radius 100m. Suppose the emergency vehicle moves at speed about 20m/s. In order to inform targeted vehicles about 15sec before the emergency vehicle enters that region around the intersection, the emergency vehicle should initiate the dissemination when it is about 300m away from that region. The allowable latency for this application is approximately 1sec according to report [97]. Hence, the message should at least hit the farthest locations of targeted region, which is 500m from the source, within 10 time slots.

From Figures 3.8 and 3.9, the upper bounds of the probability of reaching 500 meters in 10 time slots are about 70% for 1-copy direction-invariant dissemination while 100% for other three dissemination strategies. That means that 1-copy direction-invariant message dissemination is incapable of serving this application, while dissemination methods assisted by geographic information or using multiple disseminators could satisfy requirements for this application scenario. Furthermore, we can see that the probability of reaching 500 meters in 10 time slots is about 30% and 40%, 100% and 100% for 1-copy and 4-copy direction-invariant dissemination, 1-copy and 4-copy geographic-assisted dissemination, respectively. Therefore, geographic-assisted dissemination better serves the application requirements for the above scenario.

Remark 22 *Dissemination strategies that use multiple disseminators are suitable for applications like post-crash warning, which geocast targets an area near the source location. Dissemination strategies that utilize geographic information to choose relays are needed for applications like emergency vehicle signal preemption, which geocast targets an area far from the source location.*

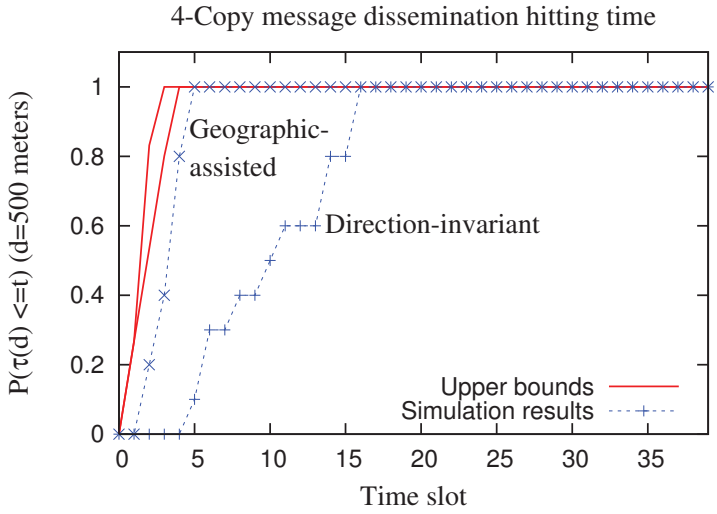

Figure 3.9: $P(\tau(d) \leq t)$ ($d = 500$m) in L-copy ($L = 4$) direction-invariant and geographic-assisted message dissemination, respectively.

3.5 Summary

In this chapter, we studied the spatial and temporal limits of geocast in VANETs. By focusing on movement of active messages rather than specifying relays on information propagation paths, we derived lower and upper bounds for the dissemination distance and hitting time. Simulation results of four dissemination algorithms validate our analysis. Two applications are presented to show that dissemination algorithms with multiple disseminators are suitable for geocast with area of interest near to source location while dissemination algorithms assisted by geographic information are suitable for geocast with area of interest far from the source.

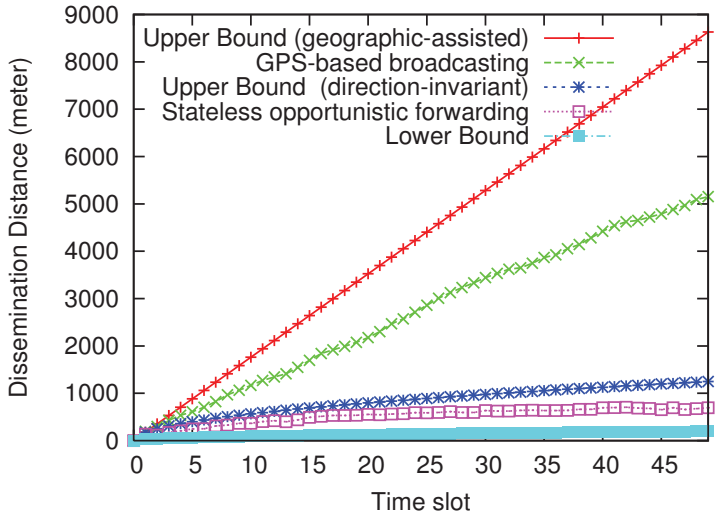

Figure 3.10: Dissemination distance of 1-copy message dissemination strategies in a highway scenario.

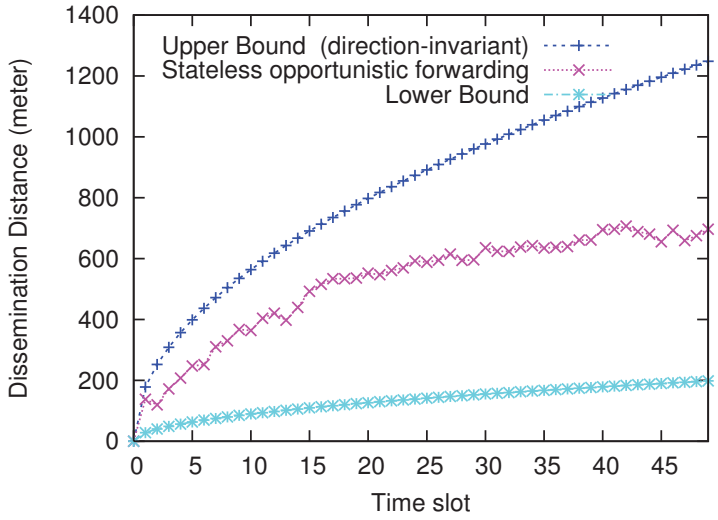

Figure 3.11: Dissemination distance of 1-copy direction-invariant dissemination strategies in a highway scenario.

Chapter 4

Evaluating the Feasibility of Mobile Cloudlets

With the emerging cloud computing [98] and the explosive growth of mobile applications, mobile cloud computing (MCC) has become a promising technology for mobile services. In MCC, mobile devices, such as smart phones and tablets, can offload data storage and computing onto the cloud through wireless communications, thereby overcoming their limited capabilities regarding process power, storage capacity, and battery lifetime [23]. Besides relying on a distant "cloud", *cloudlet* [99] is proposed to instantiate software in real-time on nearby computing resources (e.g., laptops, desktops) using virtual machine technology. In addition, with the increasing memory and computational power of mobile devices [25], nearby mobile devices can form a *mobile cloudlet* and share their computing resources to speedup computing and conserve energy. Open questions are when/whether mobile cloudlet is able to support mobile applications. In other words, whether the computation offloading generates communication traffic between mobile devices and infrastructures in cellular and WiFi networks or between peering mobile devices through opportunistic communications.

First of all, we investigate the basic scenario where a mobile device needs to offload computational tasks to remote cloud through cloudlets infrastructures. We discover that the cloudlet access probability is determined by mean connection time μ_{T_C} and mean inter-connection time μ_{T_I} between the mobile device and the cloudlet. To find out whether or when mobile cloudlet is feasible for MCC, we further study the computing performance of a mobile cloudlet through investigating the properties of a mobile cloudlet with respect to *cloudlet size*, cloudlet node's *lifetime* and *reachable time*. We demonstrate through traces and mathematical analysis that 1) the more frequently mobile devices meet, the larger the pool of computing resources an initiator can access; 2) intermittent connection between devices has little adverse effect on the optimal computing performance of mobile cloudlet in the long run; 3) the ratio $E(T_C)/[E(T_I) + E(T_C)]$

indicates the connection likelihood of two nodes, where T_C and T_I are their contact and inter-contact time. Then, we derive upper and lower bounds on computing capacity and computing speed of a mobile cloudlet. An initiator can use both bounds to decide whether to offload its task to local mobile cloudlets for mobile application services.

4.1 Motivation and Related Work

The rapid development of processing and storage technologies and the success of wired/wireless networks have enabled the *cloud computing* model [98], in which a shared pool of configurable computing resources can be accessed conveniently. Cloud computing has been widely recognized as the next generation computing infrastructure. In cloud computing, users can use infrastructures (e.g., servers and storage), platforms (e.g., operating systems), and software (e.g., application programs) provided by cloud providers at low cost in an on-demand fashion. Therefore, computing resources can be rapidly provisioned and released with minimal management cost or service provider interaction.

As mobile devices (e.g., smart phone, tablet, etc.) increasingly become an essential part of human life for convenient communication and various mobile applications, mobile cloud computing (MCC) is introduced to apply cloud computing to mobile services [23]. In MCC, the data processing and storage of mobile applications are moved from the mobile devices to powerful and centralized computing platforms in clouds, thereby extending battery lifetime, improving data storage capacity and processing power of mobile devices. The general architecture of MCC can be shown in Figure 4.1. Mobile devices are connected to the mobile cloud via base stations, access points, or satellite.

MCC faces many challenges on both computing side and wireless communication side [23]. On one hand, issues in computing side include computing offloading, security, and efficiency of data access. For instance, Chun et al. [100] design and implement the CloneCloud, a system that enables unmodified mobile applications running in an application-level virtual machine to seamlessly off-load part of their execution from mobile devices onto device clones operating in a computational cloud. Many solutions (see survey [101]) are proposed to address security issues on the cloud.

On the other hand, challenges for MCC in communication side are low bandwidth, service availability, and network heterogeneity. Paper [99] points out that accessing remote cloud through wireless communication is costly because of long WAN latencies. Rather than relying on a distant "cloud", the authors propose the use of *cloudlets*, as shown in Figure 4.2. Paper [102] presents a cloudlet architecture and a prototype implementation, showing the advantages and capabilities of cloudlet for a mobile real-time augmented reality application. The authors also point out that cloudlets do not have to be fixed infrastructure close to the wireless access point,

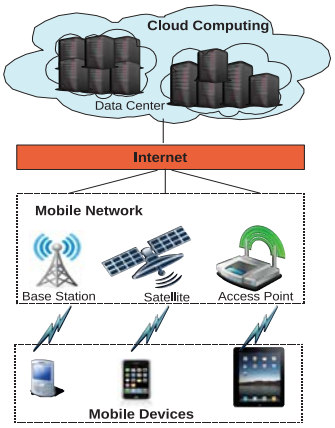

Figure 4.1: MCC uses remote cloud.

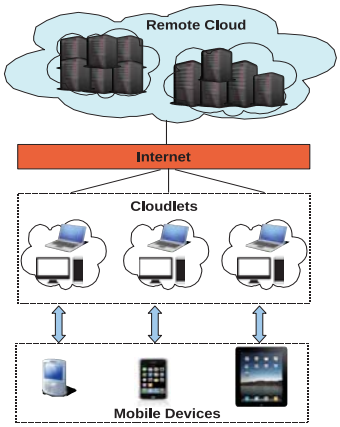

Figure 4.2: MCC uses cloudlet.

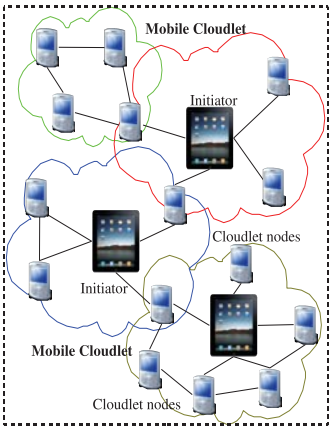

Figure 4.3: MCC uses mobile cloudlet.

but can be formed dynamically with any resource-rich devices in the LAN network. Nowadays, mobile devices have increasing memory and computational power [25]. Paper [27] proposes a virtual cloud computing platform using mobile phones with pervasiveness and high computational capabilities. The authors of paper [29] also show that it is possible for a mobile computation initiator to use computing resources in other encountered mobile devices to speedup computing and conserve energy. As shown in Figure 4.3, mobile devices can be service clients and cloudlet nodes, providing hardware and software resources.

The benefits of utilizing cloudlet and mobile cloudlet are the omnipresent and fast service access, the support of mobility and locality, the freedom of deployment and use of new services as well as the reduced hardware maintenance costs [27]. First, as the computation and information reside on nearby devices, users can get direct access instantly through interactions among one another, eliminating the communication latency introduced by the cellular radio. Second, service performance can be enhanced if the execution sequence of an application can be reordered for increasing the level of parallelism. Third, offloading to nearby mobile devices not only saves monetary cost due to expensive data charging in roaming situations, but also mitigates pressure of cellular networks and WiFi networks by serving user requests on local devices. Finally, nearby mobile users often tend to pursue the same task in social activities. This is especially true in group activities, such as visiting a museum, performing archaeological expeditions, and attending a conference. By dividing the task (e.g. construct augmented reality tourist guide) among a group of users, only a portion of the task is executed locally, thus nodes can save energy compared with a complete local execution.

Therefore, in this chapter, we investigate the cloudlet spectrum, where a mobile device's contacts with other devices can be exploited for computing. In this scenario, an initiator mobile

device tries to use the available computing resources on nearby, potentially intermittently connected devices in order to improve computing performance. The set of devices that can provide computing resources for the initiator form a cloudlet. The computing performance of a cloudlet is determined by its properties, such as number of nodes in a cloudlet and the time that a node can compute task for the initiator. Therefore, we study cloudlet properties and performance for a deeper understanding of cloudlet in MCC. Our contributions are three-fold.

1. We study the impact of mobility on mobile cloud computing in a network where nodes access cloudlets located at community sites. We find that node mobility affects not only cloudlet access probability but also cloudlet computing performance and its impact can be represented by $\mu_{T_C}/(\mu_{T_C} + \mu_{T_I})$.

2. We further study the properties and computing performance of mobile cloudlet, where mobile devices share computing resources to execute a task. Traces and analysis results together prove that the mobile cloudlet size follows negative exponential growth with parameter $1/E(T_I)$, and the expected lifetime and reachable time grow linearly with τ with the increase rates 1 and $E(T_C)/[E(T_C) + E(T_I)]$, where $E(T_C)$ and $E(T_I)$ are the expectations of contact and inter-contact time between two nodes.

3. Based on the above properties of mobile cloudlet, we study the computing capacity and speed of a mobile cloudlet. The bounds on computing capacity and speed of mobile cloudlet can both be used for an initiator to decide whether to execute the task in mobile cloudlet or remote cloud.

The rest of this chapter is organized as follows. In Section 4.2, we introduce both cloudlet and mobile cloudlet models. We study the impact of mobility on cloudlet performance in Sections 4.3. In Sections 4.4, 4.5, and 4.6, we analyze the properties and computing performance of mobile cloudlet through traces and mathematical analysis. Finally, we conclude in Section 4.7.

4.2 Models and Problem Statement

4.2.1 Cloudlet Models

Assume that a mobile device is moving in a network Ω_m with m cloudlets. The locations of cloudlets can be community locations extracted from real map or points generated according to a random process. The network is partitioned into a Voronoi diagram with m Voronoi cells, and there is one cloudlet in each region. Figure, 4.4 shows 19 cloudlets in the network and the mobile device is connected to cloudlet C_7 for mobile application computing.

Suppose the computational task on the mobile device requires C instructions. Let S_i be the computing speed, in instructions per time slot (e.g., second), of the cloudlet C_i, $\forall i = 1, 2, \ldots, m$.

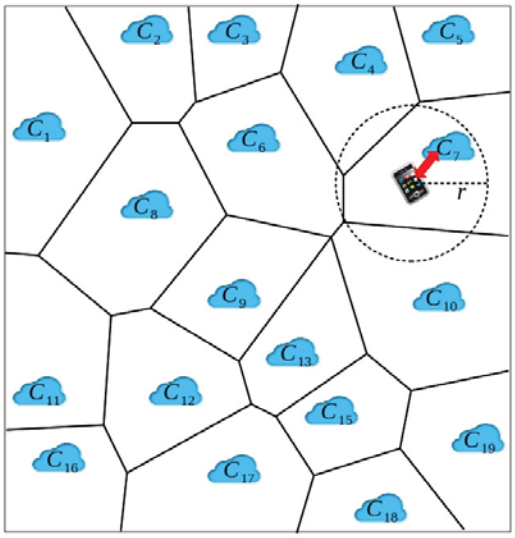

Figure 4.4: Cloudlet network model

This task thus takes C/S_i time slots to compute on cloudlet C_i, $\forall i = 1, 2, \ldots, m$. Denote B as the network bandwidth. If the mobile device needs to send D_{out} bytes of task data to the cloudlet while the cloudlet needs to send back D_{in} bytes of executed task data to the mobile device, it takes D_{out}/B and D_{in}/B time slots to transmit and receive data, respectively. Define $\delta_i = C/S_i + (D_{in} + D_{out})/B$ as the *task completion time*, which is the sum of task transmission time and task computing time at cloudlet C_i.

Cloudlet Connection Model: When cloudlet C_i ($1 \leq i \leq m$) is within the mobile device's transmission range, the mobile device can access the computing resources in C_i; otherwise, the mobile device is disconnected from C_i. An example is shown in Figure 4.5. Suppose mobile user Bob needs to do mobile commerce (e.g., mobile transactions and payments, mobile ticketing) using his smartphone. In order to avoid data overage charge and preserve battery on his phone, Bob offloads the task to a nearby cloudlet that includes resource-rich devices, such as desktops, laptops, even tablets and high-end smartphones. As Bob moves around, he exploits different cloudlets during different periods of time.

Formally, suppose mobile device is moving in the network according to a mobility process $\mathcal{M}$. Denote by $X(t)$ and $X_{C_i}(t)$ the locations of the mobile device and cloudlet C_i, respectively. Let the transmission range of the mobile device be r. Connection to C_i is available at time t if and only if $||X(t) - X_{C_i}(t)|| \leq r$, where $||\cdot||$ is the Euclidean norm in 2-Dimension. Further, the *connection and inter-connection time* between a mobile device and a cloudlet are defined as follows.

Definition 10 *The connection time T_C of the mobile device and cloudlet C_i ($\forall i = 1, 2, \ldots, m$)*

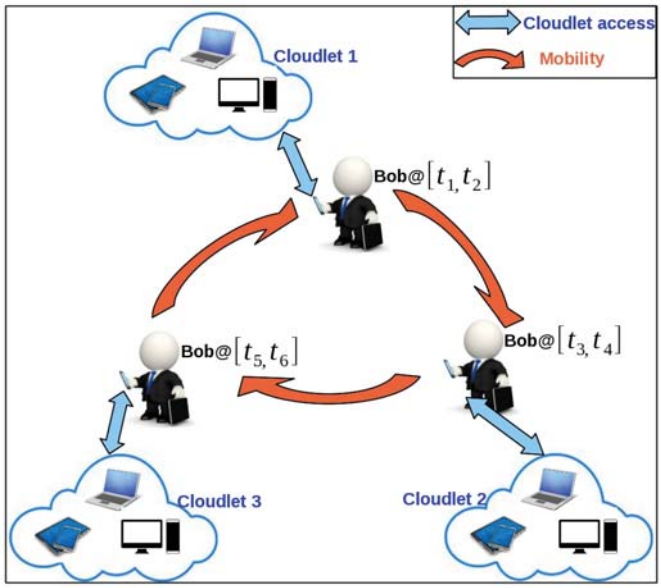

Figure 4.5: Mobile cloud computing through cloudlets in the vicinity of a mobile device: Bob uses cloudlet 1 during $[t_1, t_2]$, cloudlet 2 during $[t_3, t_4]$, and cloudlet 3 during $[t_5, t_6]$ to execute mobile applications on his phone.

is defined as

$$T_C^i \triangleq \inf_{t>0}\{t : ||X(t) - X_i(t)|| > r\}, \tag{4.1}$$

given that $||X(0) - X_i(0)|| > r$ *and* $||X(0^+) - X_i(0^+)|| \leq r$. *The inter-connection time (i.e., time between two consecutive connections) of the mobile device and cloudlet* C_i *is defined as*

$$T_I^i \triangleq \inf_{t>0}\{t : ||X(t) - X_i(t)|| \leq r\}, \tag{4.2}$$

given that $||X(0) - X_i(0)|| \leq r$ *and* $||X(0^+) - X_i(0^+)|| > r$.

Let $F_{T_C}^i$ and $F_{T_I}^i$ denote the distribution functions of the connection time T_C^i and inter-connection time T_I^i, respectively, and suppose that they have finite expectations $\mu_{T_C}^i$ and $\mu_{T_I}^i$ and their density functions $f_{T_C}^i$ and $f_{T_I}^i$ exist and are continuous on $(0, \infty)$, respectively. In reality, distributions of T_C^i and T_I^i can be estimated based on movement history of mobile users as human tend to repeat their everyday schedules [53].

In this chapter, we study the performance of using cloudlet for mobile applications. Due to node mobility, cloudlet connection is intermittent, which poses challenges for utilizing cloudlet computing. In order to identify power and node mobility of cloudlet computing, we examine the probability that a mobile device can connect to at least one cloudlet, which is called *cloudlet*

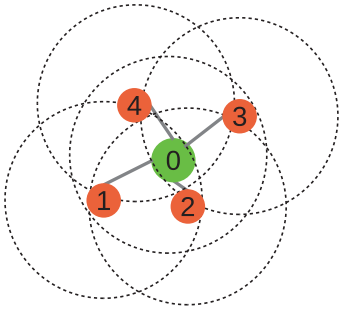

Figure 4.6: The initiator device 0 can distribute tasks to cloudlet nodes $1, 2, 3, 4$ through one-hop communications.

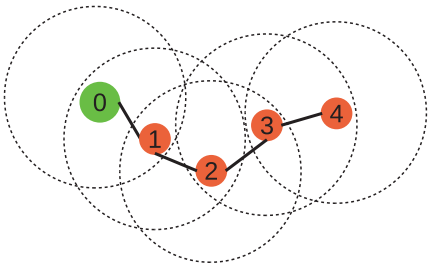

Figure 4.7: The initiator device 0 can distribute tasks to cloudlet nodes $1, 2, 3, 4$ through multi-hop communications.

access probability, the *success rate of task execution*, the total number of tasks executed by cloudlets and *average task execution speed* over time t.

4.2.2 Mobile Cloudlet Models

When accessing to cloud directly or through cloudlets is unavailable or too expensive, we consider mobile cloudlet for a mobile cloud computing network of n mobile nodes on a torus surface $\Omega_n = [0, \sqrt{\frac{n}{\lambda}}]$, where λ is the spatial density of mobile users. Suppose each mobile device has a transmission radius r. Denote by $\mathcal{X}_t = \{X_1(t), \ldots, X_n(t)\}$ the positions of users at time t, two nodes are in contact if $||X_i(t) - X_j(t)|| \leq r$ and out of contact otherwise. We assume that the mobility process of a node is stationary and ergodic that a node's location $X_i(\cdot)$ has uniform stationary distribution in the network area. Mobility processes of nodes are independent and identically distributed (i.i.d.).

Without loss of generality, we assume that a mobile user needs to offload a task to nearby mobile devices at time 0. As shown in Figures 4.6 and 4.7, initiator 0 can connect to cloudlet nodes $1, 2, 3, 4$ by direct communication links or multi-hop communication paths. In the *one-hop mobile cloudlet* (Figure 4.6), direct connections between the initiator and the cloudlet nodes ensure short delay in task transfer and easy management of task distribution and retrieval. In the *multi-hop mobile cloudlet* (Figure 4.7), employing mobile devices in *multi-hop* range provides the potential to utilize more devices in a large area. However, multi-hop communications incur longer delay and unreliable task dissemination and retrieval due to node mobility. Fesehaye et al. [103] show that when the maximum number of wireless hops in a cloudlet is larger than two, accessing cloudlet nodes incurs longer data transfer delay than directly accessing remote cloud through 3G/4G network. Hence, we only consider the more practical one-hop mobile cloudlet in this chapter.

Apparently, node mobility affects structure of mobile cloudlet. Specially, how frequent nodes meet and how long they stay in contact affect the size and stability of a mobile cloudlet. In turn, contact and inter-contact time between nodes influence the computing capacity and performance of a mobile cloudlet as tasks can only be distributed and retrieved when there are communication links between an initiator and cloudlet nodes.

Contact time T_C is also called link lifetime or link duration. Zhao et al. [104] find that PDF of link lifetime can be approximated by exponential distribution with parameter characterized by the ratio of average node speed to effective transmission range. Hence, in this chapter, we assume that T_C follows *exponential* distribution with parameter λ_C. Inter-contact time T_I has shown to exhibit exponential tail decay under many mobility models (such as random waypoint and Brownian motion) [88]. Analysis of a diverse set of mobility traces [105] also reveals that T_I follows a power law decay up to a characteristic time, beyond which T_I's distribution decays exponentially. For the simplicity of analysis, we assume that T_I has an *exponential* distribution with parameter λ_I. Note that our analysis and results can be easily extended to the case when T_C and T_I follow other distributions.

Suppose the delay requirement of an initiator's task is τ, mobile devices that meet the initiator before the task expires have the potential to provide computing services, thus can form a mobile cloudlet for the task computation. We assume that all nodes are willing to support cloudlet computing. Hence, *a mobile cloudlet is dynamically formed by the nodes that the initiator encounters over a period of time* τ. Formally, we define a mobile cloudlet as follows.

Definition 11 *(*Mobile Cloudlet*) For* $\tau \in \mathbb{R}_+$*, let* $\mathcal{C}_\tau$ *be the mobile cloudlet for an initiator* v_i *with a task to compute within delay* τ*.* $\mathcal{C}_\tau$ *is the set of nodes* v_i *encountered within time* τ*, where cloudlet node* $v_j \in \mathcal{C}_\tau$ *if and only if* $v_i \neq v_j$ *and there exists a link between* v_i *and* v_j *at a time* $0 \leq t \leq \tau$*.*

The task dispatching, computing, and retrieving can only be performed after the first contact between an initiator and a cloudlet node and before their last contact within time τ. We find the following definition useful.

Definition 12 *(*Lifetime*) For any cloudlet node* $v_j \in \mathcal{C}_\tau$ *for an initiator* v_i*,* v_j*'s* lifetime $LT(\tau) = exit - entr$*, where its* entrance time *to* $\mathcal{C}_\tau$

$$entr \triangleq \inf_{0 \leq t \leq \tau} \{t : ||X_i(t) - X_j(t)|| \leq r\},$$

and its exit time *from* $\mathcal{C}_\tau$

$$exit \triangleq \inf_{0 \leq t \leq \tau} \{t : \forall t' >= t \text{ and } t' \leq \tau, ||X_i(t') - X_j(t')|| > r\}.$$

In an optimal situation, an initiator utilizes a cloudlet node's whole lifetime for computing. A cloudlet node receives tasks at its entrance time; then it can compute the task during its lifetime even when it is not in contact with the initiator; it sends back the tasks right before its exit time. Hence, the lifetime of a cloudlet node can be used to provide an *upper bound* on the computing capacity of mobile cloudlet.

Nevertheless, the task dissemination and retrieval can only be performed during the contact period of an initiator and a cloudlet node. The percentage of time that a cloudlet node is in contact with the initiator shows how likely the initiator can reach it . In order to study the reachability of cloudlet nodes and reliability of a mobile cloudlet, we define the *reachable time* $RT(\tau)$ *as the total contact duration between a cloudlet node and an initiator within time* τ. Based on these mobile cloudlet properties, we can study mobile cloudlet computing performance and find out whether and when a mobile cloudlet can serve mobile applications.

4.3 Impact of Mobility on Cloutlet Performance

Because of node mobility, the connection between a mobile device and a cloudlet can be intermittent. In order to study cloudlet computing performance, we start with modeling the connection and inter-connection process between a mobile device and a cloudlet.

Definition 13 *Let* $\{\eta(t), 0 \leq t < \infty\}$ *be a stochastic process with state space* $\{0, 1\}$. *If a mobile device can connect to a cloudlet at time* t, $\eta(t) = 1$; *otherwise,* $\eta(t) = 0$. *Denote by* $\alpha_1, \beta_1, \alpha_2, \beta_2, \ldots$ *the lengths of successive intervals spent in states* 0 *and* 1, *respectively, in time* $(0, \infty)$, *where* $\alpha_1, \alpha_2, \ldots$ *are i.i.d. and* $\beta_1, \beta_2, \ldots$ *are i.i.d.. The process* $\{\eta(t)\}$ *assumes the states* 0 *and* 1 *alternately, as shown in Figure 4.8. The process* $\{\eta(t)\}$ *is called* alternating renewal process.

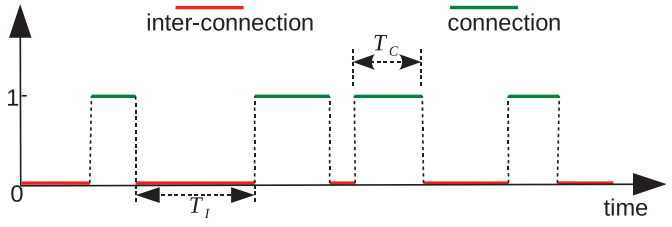

Figure 4.8: The connection and inter-connection process of a mobile device and a cloudlet is an alternating renewal process.

With only a slight loss of generality, we assume that the time origin of the process $\{\eta(t)\}$ is an arbitrary connection or inter-connection. When $\eta(0) = 0$, the mobile device is initially disconnected from the cloudlet; when $\eta(0) = 1$, the mobile device is initially connected with the cloudlet. If $\eta(0) = 0$, $\alpha_i = T_I^i$ and $\beta_i = T_C^i$, i.e., ith inter-connection and connection time, respectively; if $\eta(0) = 1$, $\alpha_i = T_C^i$ and $\beta_i = T_I^i$, i.e., ith connection and inter-connection time, respectively. The former case is shown in Figure 4.8. Based on the alternating renewal process of the connection between a mobile device and a cloudlet, we study the cloudlet access probability, task success rate and execution speed.

4.3.1 Cloudlet access probability

A mobile device's connection to cloudlets is intermittent due to node mobility. If there is no cloudlet in the vicinity of mobile device, cloudlet computing is unavailable. Hence, it is important to find out the *cloudlet access probability*, which is defined as the probability that a mobile device can connect to at least one cloudlet in the network.

Note that if the mobile device can connect to multiple resource-rich devices at the same time, these devices can be seen as belonging to one cloudlet. In other words, a mobile device can connect to at most one cloudlet at any time t, and a mobile device's connections with different cloudlets are exclusive. This assumption is reasonable because different cloudlets probably locate at different community sites. Based on this assumption and the connection and inter-connection process of a mobile device and a cloudlet, we have the following theorem for the cloudlet access probability.

Theorem 8 *The limiting cloudlet access probability is*

$$CA = \sum_{i=1}^{m} \frac{\mu_{T_C}^i}{\mu_{T_C}^i + \mu_{T_I}^i}. \tag{4.3}$$

where $\mu_{T_C}^i$ and $\mu_{T_I}^i$ are expectations of connection time T_C^i and inter-connection T_I^i between the mobile device and cloudlet C_i ($i = 1, 2, \ldots, m$), respectively.

Proof : The movements of mobile device result in alternating connection and inter-connection with a cloudlet $C_i, 1 \leq \forall i \leq m$, which is modeled in Definition 13. The probability that the connection between the mobile device and cloudlet C_i is available at time t, conditional on the initial state, is given by Cox in Renewal Theory (1962, p.83) [106]. When the mobile device is initially connected to cloudlet C_i,

$$CA_1^i(t) = 1 - F_{T_C}^i(t) + \int_0^t h_1^i(u)[1 - F_{T_C}^i(t-u)]du,$$

where $h_1(u)$ is the inverse Laplace transform of

$$H_1^i(s) = \frac{f_{T_C}^i(s) f_{T_I}^i(s)}{s(1 - f_{T_C}^i(s) f_{T_I}^i(s))};$$

when the mobile device is disconnected from cloudlet C_i at $t = 0$,

$$CA_0^i(t) = F_{T_I}^i(t) + \int_0^t h_1^i(u)[1 - F_{T_I}^i(t-u)]du.$$

It is reasonable to assume that the process of connection and inter-connection between nodes has been running for a long time before it is first observed. The limiting connection probability of cloudlet C_i is

$$CA^i = \lim_{t\to\infty} CA_1^i(t) = \lim_{t\to\infty} CA_0^i(t) = \frac{\mu_{T_C}^i}{\mu_{T_C}^i + \mu_{T_I}^i}. \tag{4.4}$$

As the mobile device's connections to different cloudlets are exclusive, the cloudlet access probability is $CA = \sum_{i=1}^m CA^i$. Thus, we finish our proof.

Remark 23 *Connection probability CA^i of a cloudlet C_i is determined by the average connection and inter-connection time, i.e., mobility pattern of a mobile user. The more frequent visit and the longer sojourn time at the location of a cloudlet, the more likely a mobile user can connect to this cloudlet. The mobile device's isolation probability is $1-CA$, which is determined by the percentage of time that the mobile user is at locations without any cloudlet (i.e., user mobility pattern).*

4.3.2 Task Success Rate

Cloudlets available at a point in time can not guarantee a successful task execution. In order to successfully compute a task, a cloudlet has to maintain a connection with the mobile device during the task transmissions and computation. In other words, cloudlet C_i can successfully execute a task for a mobile device if a connection is available between them for at least δ_i period of time, where $\delta_i = C/S_i + (D_{in} + D_{out})/B$. We derive the task success rate by applying results on interval availability of an alternating renewal process [107].

Theorem 9 *The task success rate is*

$$SR = \sum_{i=1}^m CA^i \left(1 - \int_0^{\delta_i} [1 - F_{T_C}^i(x)]/\mu_{T_C}^i dx\right), \tag{4.5}$$

where cloudlet C_i's access probability $CA^i = \frac{\mu_{T_C}^i}{\mu_{T_C}^i + \mu_{T_I}^i}$.

Proof : Define $SR^i(t,\delta_i)$ as the probability that a connection between the mobile device and cloudlet C_i is available at least δ_i period of time starting at t. Based on the interval availability of an alternating renewal process derived by Barlow and Hunter (1961) [107], we have that conditioning on initial state $\eta(0)=1$,

$$SR_1^i(t,\delta_i) = 1 - F_{T_C}^i(t+\delta_i) + \int_0^t h_1^i(u)[1 - F_{T_C}^i(t+\delta_i-u)]du;$$

while conditioning on initial state $\eta(0)=0$,

$$SR_0^i(t,\delta_i) = \int_0^t h_2^i(u)[1 - F_{T_C}^i(t+\delta_i-u)]du,$$

where $h_2^i(u)$ is the inverse Laplace transform of

$$H_2^i(s) = \frac{f_{T_I}^i(s) f_{T_I}^i(s)}{s(1 - f_{T_C}^i(s) f_{T_I}^i(s))}.$$

It is readily seen

$$\begin{aligned} SR^i &= \lim_{t\to\infty} SR_1^i(t,\delta_i) = \lim_{t\to\infty} SR_0^i(t,\delta_i) \\ &= \frac{\int_{\delta_i}^{\infty}(1 - F_{T_C}^i(u))du}{\mu_{T_C}^i + \mu_{T_I}^i} = \frac{\mu_{T_C}^i(1 - \int_0^{\delta_i}[1 - F_{T_C}^i(u)]/\mu_{T_C}^i du)}{\mu_{T_C}^i + \mu_{T_I}^i}. \end{aligned} \tag{4.6}$$

Note that SR^i is the product of the limiting cloudlet access probability CA^i and the limiting probability that it survives an interval of duration δ_i. As the limiting success rate is $SR = \sum_{i=1}^m SR^i$, we complete our proof.

The connection time between a mobile device and a cloudlet is also called contact time or link lifetime, which has been formally described to be *exponential* random variable under various mobility models [104, 108]. When T_C^i $(i = 1, 2, \ldots, m)$ follows exponential distribution with parameter $1/\mu_{T_C}^i$, we have the following corollary.

Corollary 1 *When* $\{T_C^i, i = 1, 2, \ldots, m\}$*, are exponential random variables with rates* $\{1/\mu_{T_C}^i, i = 1, 2, \ldots, m\}$*, the limiting task success rate is*

$$SR = \sum_{i=1}^{m} CA^i e^{-\frac{\delta_i}{\mu_{T_C}^i}}. \tag{4.7}$$

Remark 24 *The probability that a task can be executed successfully by cloudlet* C_i *not only depends on the* cloudlet access probability CA^i and the *probability distribution of connection time* T_C^i but also depends on the *task completion time* δ_i, which is determined by computation

demand C, sizes of task data D_{in} and D_{out}, cloudlet computing speed S_i, and channel bandwidth B.

4.3.3 Task Execution Speed

In mobile cloud computing, mobile applications, such as mobile learning, health monitoring, and map navigation, require recurrent services. For example, in mobile gaming, mobile users offload game engine (e.g., graphic rendering) to the servers in the cloudlet and users need to access the cloudlet repeatedly as the game refreshes during the game playing time. In general, a mobile device has a large amount of tasks to compute, and each task is sent to a cloudlet for computing after the previous task is finished. It is important to find out how many tasks can be executed successfully over time t and what is the average task execution speed.

Because of node mobility, the connection between a mobile device and a cloudlet is unstable. In order to maintain high cloudlet computing reliability, it is reasonable to assume that for recurrent task computing, a mobile device only utilizes a cloudlet when they are connected. Accordingly, the average number of executed tasks over a fixed time t depends on the total connection time between a mobile device and its encountered cloudlets as well as task completion time. We derive the following theorem using renewal theory.

Theorem 10 *The average number of executed tasks over time t, denoted as $N(t)$, satisfies*

$$E(N(t)) = \sum_{i=1}^{m} \left\lfloor \frac{E(N_C^i(t))\mu_{T_C}^i}{\delta_i} \right\rfloor, \tag{4.8}$$

where $\lfloor \cdot \rfloor$ is the floor function, and $E(N_C^i(t))$ is the number of connections between a mobile user and cloudlet C_i within time t. Formally, $E(N_C^i(t))$ is the inverse Laplace transform of $F_{T_C+T_I}^i(s)/[s(1-F_{T_C+T_I}^i(s))]$ and $F_{T_C+T_I}^i(s)$ is the Laplace transform of random variable $T_I^i + T_C^i$.

Proof : In the connection and inter-connection process of a mobile device and cloudlet C_i, define $S_0^i = 0$ and $S_n^i = \alpha^{i1} + \beta^{i1} + \alpha^{i2} + \beta^{i2} + \cdots + \alpha^{in} + \beta^{in}$ for $n \geq 1$. The process $N_C^i(t) = \max_{n \geq 0}\{n | S_n^i \leq t\}$ is the number of renewals over time t. The total connection time between a mobile device and cloudlet C_i over time t is approximately

$$CT_i(t) \approx \sum_{k=1}^{N_C^i(t)} T_C^{ik}, \tag{4.9}$$

where T_C^{ik} is the kth connection time between a mobile device and cloudlet C_i. As $\{T_C^{ik}, k =$

$1, 2, \dots\}$ are i.i.d. and have the same distribution as T_C^i,

$$E(CT_i(t)) = E(N_C^i(t))E(T_C^i) = E(N_C^i(t))\mu_{T_C}^i, \tag{4.10}$$

where $E(N_C^i(t))$, by renewal theory, is the inverse Laplace transform of $F_{T_C+T_I}^i(s)/[s(1 - F_{T_C+T_I}^i(s))]$ and $F_{T_C+T_I}^i(s)$ is the Laplace transform of random variable $T_I^i + T_C^i$. The number of tasks executed by cloudlet C_i is $\lfloor E(CT_i(t))/\delta_i \rfloor$. Accordingly, the total number of executed tasks over time t is sum of $\lfloor E(CT_i(t))/\delta_i \rfloor$ over all cloudlets C_i, $i = 1, 2, \dots, m$.

Similar to connection time T_C, inter-connection time T_I, also called inter-contact time, has been shown to exhibit exponential tail decay under many mobility models [88]. Under the special case when T_C^i and T_I^i $(i = 1, 2, \dots, m)$ are exponential random variables, we can derive the closed form for the average number of renewals $E(N_C^i(t))$ over time t, thus $E(N(t))$ in the following corollary.

Corollary 2 *If T_C^i and T_I^i $(1 \leq \forall i \leq m)$ are exponential random variables with rates $1/\mu_{T_C}^i$, $1/\mu_{T_I}^i$, respectively,*

$$E(N(t)) = \sum_{i=1}^{m} \Big\lfloor \frac{CA^i t + CA^i(1 - CA^i)\mu_{T_C}^i \left(1 - e^{-\frac{t}{CA^i \mu_{T_I}^i}}\right)}{\delta_i} \Big\rfloor, \tag{4.11}$$

where $CA^i = \mu_{T_C}^i/(\mu_{T_C}^i + \mu_{T_I}^i)$.

Proof : When T_C^i and T_I^i are exponentially distributed with rates $1/\mu_{T_C}^i$ and $1/\mu_{T_I}^i$, respectively, $T_I^i + T_C^i$ has density function $\frac{1}{\mu_{T_I}^i - \mu_{T_C}^i}(e^{-t/\mu_{T_I}^i} - e^{-t/\mu_{T_C}^i})$, which gives the Laplace transform $L_{T_I+T_C}(s) = \frac{1}{\mu_{T_C}^i \mu_{T_I}^i (s+1/\mu_{T_C}^i)(s+1/\mu_{T_I}^i)}$. Then,

$$\mathcal{L}(E(N_c^i(t)), s) = \frac{1}{\mu_{T_C}^i \mu_{T_I}^i s^2 (s + \frac{1}{\mu_{T_C}^i} + \frac{1}{\mu_{T_I}^i})}.$$

Performing inverse Laplace transform, we have

$$E(N_c^i(t)) = \frac{t}{\mu_{T_C}^i + \mu_{T_I}^i} + CA^i(1 - CA^i)(1 - e^{-(\frac{1}{\mu_{T_C}^i} + \frac{1}{\mu_{T_I}^i})t}).$$

Substituting this equation into Eq. (4.8), we finish the proof.

To understand how the number of tasks executed by cloudlets increases over time t, we give some numerical results in Figure 4.9. We set the scenario that a mobile user mainly stays at work place and home. Let there be two cloudlets in the network (i.e., $m = 2$). Cloudlet

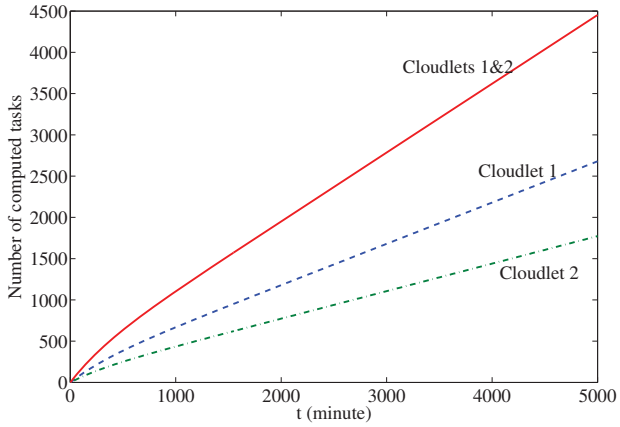

Figure 4.9: Number of tasks computed by cloudlet 1, cloudlet 2, and both cloudlets over time t.

C_1 locates at the mobile user's home, and Cloudlet C_2 locates at the mobile user's office. The mean connection and inter-connection time between the mobile user with these two cloudlets are $\mu^1_{T_C} = 12$ hours and $\mu^1_{T_I} = 12$ hours, $\mu^2_{T_C} = 8$ hours and $\mu^2_{T_I} = 16$ hours, respectively. Let the task completion time $\delta_1 = \delta_2 = 1$ minute. Figure 4.9 shows that number of tasks executed by cloudlet C_1 (or C_2) increases linearly with rate approximately equal to cloudlet C_1's (or C_2's) access probability $CA^1 = \frac{1}{2}$ (or $CA^2 = \frac{1}{3}$). The total number of executed tasks by cloudlets in the network increases linearly with rate approximately equal to cloudlet access probability $CA = CA^1 + CA^2 = \frac{5}{6}$.

Remark 25 *Average number of executed tasks over time t grows linearly with rate approximately equal $\sum_{i=1}^{m} CA^i/\delta^i$. In other words, number of executed tasks is mainly determined by cloudlet access probability and task completion time.*

Theorem 11 *The limiting average speed of task execution satisfies*

$$CS = \lim_{t\to\infty} \frac{E(N(t))}{t} = \sum_{i=1}^{m} \frac{CA^i}{\delta_i}. \tag{4.12}$$

Proof : The average speed of task execution is $CS(t) = E(N(t))/t$. Based on Theorem 10, we have limiting task execution speed when $t \to \infty$,

$$CS = \lim_{t\to\infty} \frac{E(N(t))}{t} = \sum_{i=1}^{m} \lim_{t\to\infty} \frac{E(N_c^i(t))\mu^i_{T_C}}{t}.$$

According to the elementary renewal theorem,

$$\lim_{t\to\infty} \frac{E(N_c^i(t))}{t} = \frac{1}{(\mu_{T_C}^i + \mu_{T_I}^i)}.$$

Hence, we complete our proof.

Remark 26 *The higher the cloudlet access probability CA^i $(1 \leq \forall i \leq m)$ and the shorter the task completion time δ_i $(1 \leq \forall i \leq m)$ are, the faster the task execution speed CS is. Findings in this chapter reveal that mobility pattern of a mobile user determines its connection and inter-connection time to cloudlets, which in turn affect not only the cloudlet access probability, but also success rate and speed of task execution.*

4.4 Mobile Cloutlet Properties in Traces

The cloudlet size and cloudlet node's lifetime and reachable time are determined by contacts and inter-contacts between the initiator and cloudlet nodes, which have been studied using mobility traces in mobile wireless networks. Hence, we start examining the mobile cloudlet properties using mobility traces.

4.4.1 Mobility Traces

Mobility traces record mobile users' access to base stations or access points (i.e., infrastructure based traces), or GPS locations (i.e., GPS based traces), or contact and inter-contact time (i.e., direct contact based traces). Because mobile cloudlet exploits contacts among nodes for computing, we choose the direct contact based traces. Moreover, since mobile cloudlet is promising for a social group sharing common tasks, mobility traces of users in social groups are preferred. Therefore, we select the *Cambridge/haggle2009* dataset [109] that includes several traces of Bluetooth sightings by groups of users carrying small devices (iMotes) for several days in office and conference.

In Cambridge/haggle2009 data collection, experiment 2 distributed iMotes to 19 graduate students from the System Research Group at University of Cambridge for around 5 days in 2005. Number of contacts, contact and inter-contact time among nodes were collected. Only 12 iMotes were used to produce trace file *Exp2*, while others were discarded because of hardware resets. Similarly, experiment 3 distributed iMotes to 50 students attending the student workshop at the IEEE Infocom Conference in Grand Hyatt Miami from March 7th to March 10th, 2005. Only 41 iMotes delivered useful contact information for trace file *Exp3*.

Exp2 and Exp3 represent node contact on campus and in conference environments, respectively. In both scenarios, mobile users are likely to work on a common task due to their

common social activities (i.e., working in the same lab and attending the same conference). Nearby mobile users can create computing communities in which mobile devices can collaboratively execute shared tasks. Thus, properties of mobile cloudlet extracted from these two trace files can characterize the real mobile cloudlet system.

4.4.2 Cloudlet Size

According to the definition of mobile cloudlet, we analyze the size of $\mathcal{C}_\tau$ by calculating the average number of encountered nodes over time τ in traces Exp2 and Exp3. As shown in Figures 4.10 and 4.11, size of $\mathcal{C}_\tau$ increases as τ increases. Using curve fitting, we can see that negative exponential distributions approximately fit the data in Figures 4.10 and 4.11. In other words, the size of $\mathcal{C}_\tau$ is a negative exponential function of τ and the number of nodes n in the network (e.g., the maximum cloudlet size is 11 in trace Exp2).

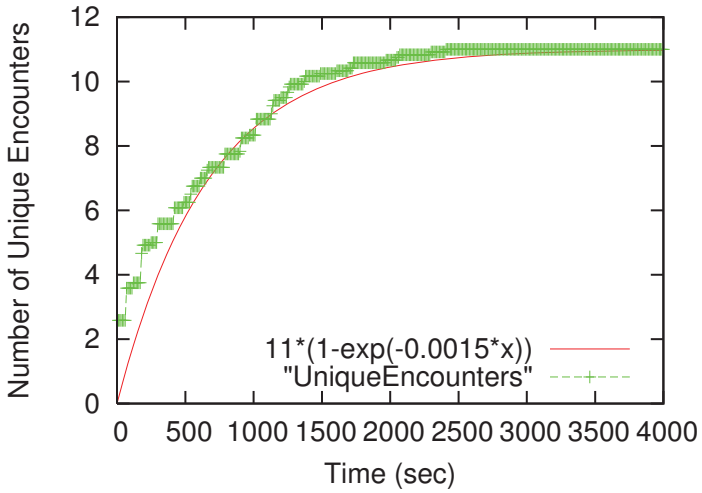

Figure 4.10: In the trace of Exp2, sizes of mobile cloudlet $\mathcal{C}_\tau$ follow negative exponential growth with τ.

Besides the network size n and time τ, the exponential growth rate is vital for cloudlet size. A large growth rate means that an initiator frequently meets resource-rich devices, thus likely acquires a large pool of potential computing resources. On the contrary, a small growth rate means that an initiator seldom encounters new nodes and can only acquires computing resources from a small portion of nodes in the network, which may lead to poor computing performance. We will further study the increase rate and how it affects the computing performance of mobile cloudlet through mathematical analysis.

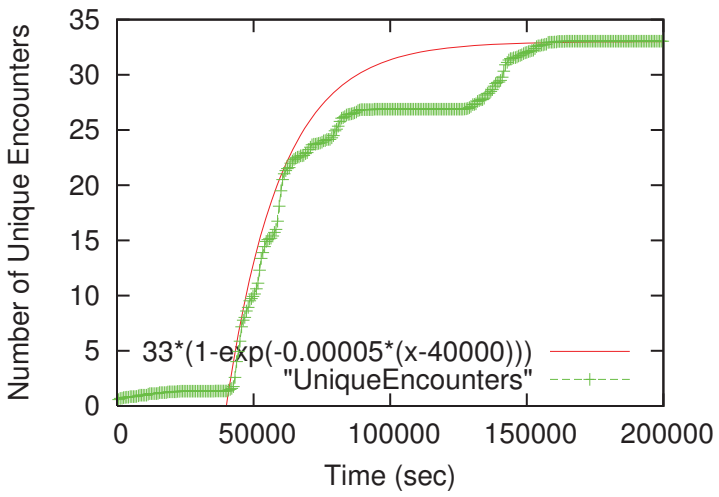

Figure 4.11: In the trace of Exp3, sizes of mobile cloudlet $\mathcal{C}_\tau$ follow negative exponential growth with τ.

4.4.3 Lifetime

Based on Definition 12, node v_j's lifetime in a mobile cloudlet for initiator v_i is from v_i and v_j's first contact to their last contact within time τ. Figures 4.12 and 4.13 show the average $LT(\tau)$ in traces Exp2 and Exp3, respectively. $LT(\tau)$ increases as τ increases. Lifetime increases slowly when $80000 < \tau < 132000$ (about 14-hour period) in trace Exp2 and when $90000 < \tau < 129000$ (about 11-hour period) in trace Exp3. This is probability because users have little contact during nights, which is also observed in cloudlet node's reachable time in Figures 4.14 and 4.15.

From Figures 4.12 and 4.13, it is difficult to determine the lifetime when τ is small because of the randomness of inter-contact time T_I. But, it agrees with our intuition that a node's lifetime increases linearly with rate 1 when τ is large as shown in Figures 4.12 and 4.13. This implies that for delay tolerant application (i.e., large τ), the optimal computing performance of mobile cloudlet—achieved by exploiting the cloudlet nodes' whole lifetime for computing—is hardly influenced by intermittent connections between an initiator and cloudlet nodes.

4.4.4 Reachable Time

The reachable time $RT(\tau)$ of a cloudlet node is its total contact duration with the initiator within time τ. Figures 4.14 and 4.15 show the average reachable time in traces Exp2 and Exp3, which are piecewise linear functions. The piecewise linearity is due to different mobility patterns of users at different times (daytime and night time). For instance, in trace Exp2, students are working in the lab during $50000 < \tau < 80000$ (about 8-hour period), producing long contact

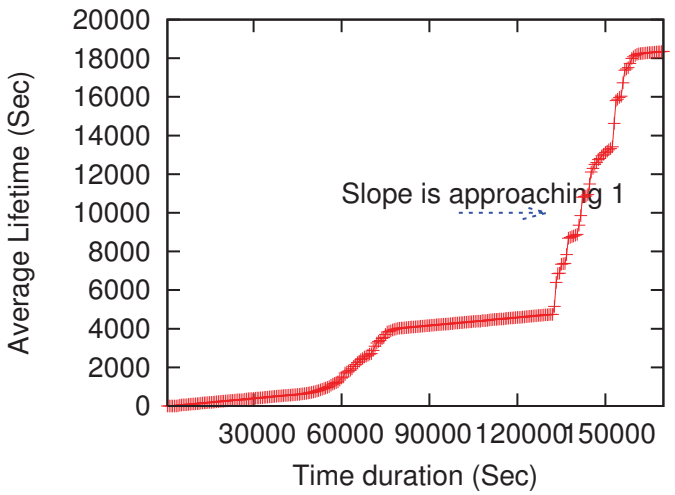

Figure 4.12: In the trace of Exp2, average lifetimes of cloudlet nodes increase approximately linearly with τ when time τ is large.

time and short inter-contact time. This leads to high growth rate of reachable time, as shown in Figure 4.14. On the other hand, they have short contacts and long inter-contacts during off time, which lead to small growth rate of reachable time. Examining Figures 4.14 and 4.15 more closely, we discover that the increase rate depends on the average contact time and inter-contact time between two nodes over the corresponding period of time. For instance, the slope of segment $\tau \in [44400, 85200]$ in Figure 4.15 is 0.01882, which is very close to average T_C/(average T_C + average T_I)=0.01877 during $44400 < \tau < 85200$.

Remark 27 *In traces Exp2 and Exp3, cloudlet nodes' reachable times increase linearly with τ and the increase rates are approximately average T_C/(average T_C + average T_I), which are varying according to users' mobility patterns. The increase rate indicates the connection likelihood between an initiator and a cloudlet node. If an initiator can connect to devices with high likelihood, it could receive omnipresent and reliable mobile cloudlet computing service.*

4.5 Theoretical Analysis of Mobile Cloutlet Properties

In this section, we mathematically analyze the properties of mobile cloudlet. Our analysis not only confirms our previous observations but also enables us to investigate the computing performance of a mobile cloudlet, such as computing capacity and speed, which determine when a mobile cloudlet is competent for executing mobile applications.

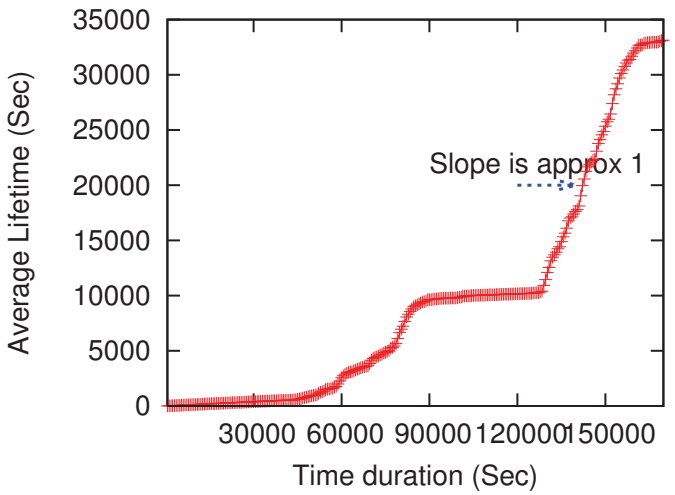

Figure 4.13: In the trace of Exp3, average lifetimes of cloudlet nodes increase approximately linearly with τ when time τ is large.

4.5.1 Cloudlet Size Analysis

Denote by $\mathcal{N}_{\mathcal{C}_\tau}(t)$ $(0 \leq t \leq \tau)$ the number of devices that an initiator encounters within in time t. Then, $\mathcal{N}_{\mathcal{C}_\tau}(0) = N^*$, where N^* is the number of the initiator's neighbors at $t = 0$, and cloudlet size $N_{\mathcal{C}_\tau} = \mathcal{N}_{\mathcal{C}_\tau}(\tau)$. We can study $\mathcal{N}_{\mathcal{C}_\tau}(t)$ by superposing multiple 0-1 processes that are sequences of on (contact) and off (inter-contact) times between two nodes.

Over a period of time t, a cloudlet node and the initiator will be in contact (on) and inter-contact (off) states alternately. If this process begins at the origin of a contact or an inter-contact, it can be modeled as the conventional alternating renewal process. However, in practice, the choice of time origin does not always coincide with the beginning of a contact or an inter-contact. Thus, we adopt the following modified alternating renewal process.

Definition 14 *Define a stochastic process* $\{\chi(t), 0 \leq t < \infty\}$ *with values in an abstract space* $X = A + B$. *The process* $\{\chi(t)\}$ *assumes the states* A *and* B *alternately. Denote by* $\xi^1, \eta^1, \xi^2, \eta^2, \cdots$ *the successive sojourn times spent in states* A *and* B, *respectively, where* $\xi^2, \xi^3, \cdots$ *are i.i.d.,* $\eta^1, \eta^2, \cdots$ *are i.i.d., while* ξ^1 *has a different distribution. Define* $S_0 = 0$, $S_1 = \xi^1$, *and* $S_n = \xi^1 + \eta^1 + \xi^2 + \eta^2 + \cdots + \eta^{n-1} + \xi^n$ *for* $n \geq 2$, *the process* $N(t) = \max_{n \geq 0}\{n | S_n \leq t\}$ *is called* modified alternating renewal process. *This process is also called* equilibrium alternating renewal process, *if* ξ^1 *has the PDF* $[1 - F_\xi(x)]/E(\xi)$, *where* $F_\xi(x)$ *and* $E(\xi)$ *are the cumulative distribution function (CDF) and expectation of* ξ^i *for all* $i > 1$.

Let A and B represent contact and inter-contact states. When two node are in contact at $t = 0$, $\xi^1 = \widetilde{T_C^1}$ (the residual time of T_C^1), $\xi^i = T_C^i$ $(\forall i > 1)$, and $\eta^i = T_I^i$ $(\forall i \geq 1)$; when two nodes

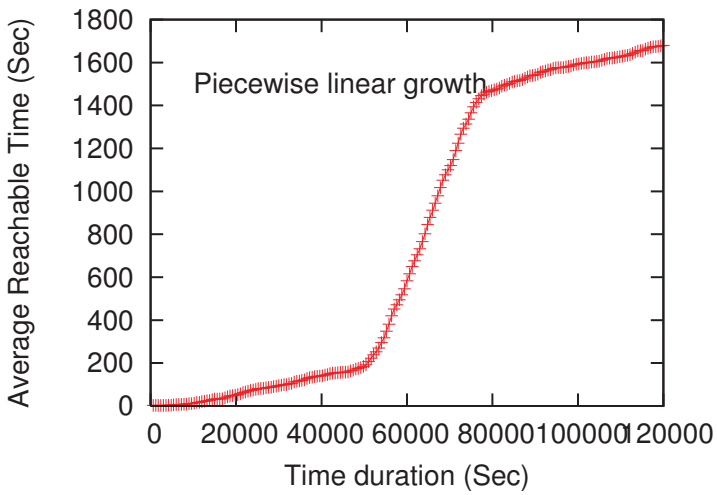

Figure 4.14: In the trace of Exp2, average reachable times of cloudlet nodes are piecewise linear functions of time τ with slope depending on contact and inter-contact time.

are *not* in contact at $t = 0$, $\xi^1 = \widetilde{T_I^1}$ (the residual time of T_I^1), $\xi^i = T_I^i$ $(\forall i > 1)$, and $\eta^i = T_C^i$ $(\forall i \geq 1)$. It is reasonable to assume that the process of contact and inter-contact between nodes has been running for a long time before it is first observed. Then ξ^1 will have the PDF of $[1-F_\xi(x)]/E(\xi)$ [106]. Therefore, this process can be seen as an equilibrium alternating renewal process.

Theorem 12 *The expectation of cloudlet size is*

$$E(N_{\mathcal{C}_\tau}) = (n-1)\left[1 - \left(1 - \frac{\pi r^2}{n/\lambda}\right) e^{-\lambda_I \tau}\right]. \tag{4.13}$$

Proof : Assume at $t = 0$, there are N^* nodes in the initiator's transmission range, $\mathcal{N}_{\mathcal{C}_\tau}(t) - \mathcal{N}_{\mathcal{C}_\tau}(0)$ $(t > 0)$ can be seen as superposition of $n - N^* - 1$ number of 0-1 processes $1_{\{N_i(t)>0\}}$, where $N_i(t)$ has the same distribution as $N(t)$.

$$\mathcal{N}_{\mathcal{C}_\tau}(t) = N^* + \sum_{i=1}^{n-N^*-1} 1_{\{N_i(t)>0\}} = n - 1 - \sum_{i=1}^{n-N^*-1} 1_{\{\xi_i^1 > t\}}.$$

It is worth noting that N^* is a random variable depending on initial node distribution in the network. Rigorously, $P(\mathcal{N}_{\mathcal{C}_\tau}(t) = k) = E(P(\mathcal{N}_{\mathcal{C}_\tau}(t) = k|N^*))$ and $E(\mathcal{N}_{\mathcal{C}_\tau}(t)) = E(E(\mathcal{N}_{\mathcal{C}_\tau}(t)|N^*))$. Therefore, $\mathcal{N}_{\mathcal{C}_\tau}(t)$ is determined by the initial node distribution and the residual inter-contact time between two nodes. In homogeneous network, N^* satisfies $P(N^* = m) = \binom{n-1}{m}\left(\frac{\pi r^2}{n/\lambda}\right)^m \times \left(1 - \frac{\pi r^2}{n/\lambda}\right)^{n-1-m}$. Then, $\mathcal{N}_{\mathcal{C}_\tau}(t)$ has a binomial distribution with parameters $n-1$ and $\frac{\pi r^2}{n/\lambda} +$

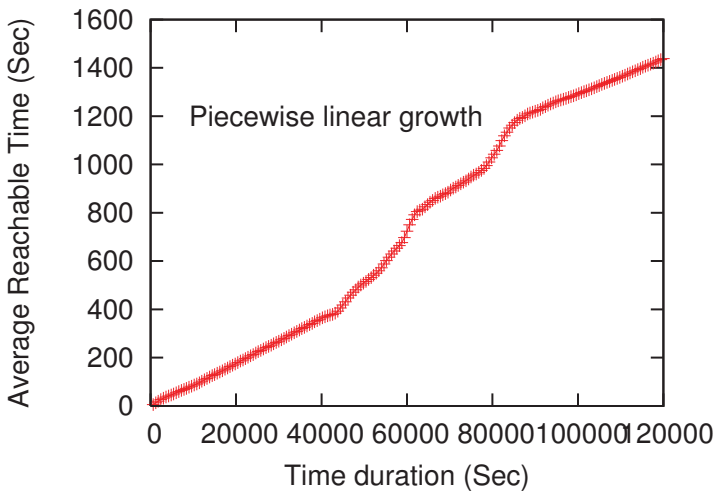

Figure 4.15: In the trace of Exp3, average reachable times of cloudlet nodes are piecewise linear functions of time τ with slope depending on contact and inter-contact time.

$F_{\widetilde{T_I}}(t)\left(1-\frac{\pi r^2}{n/\lambda}\right)$, where $F_{\widetilde{T_I}}(t) = P(\widetilde{T_I} \leq t)$. Thus,

$$E(\mathcal{N}_{\mathcal{C}_\tau}(t)) = (n-1)\left(\frac{\pi r^2}{n/\lambda} + \left(1-\frac{\pi r^2}{n/\lambda}\right)F_{\widetilde{T_I}}(t)\right). \tag{4.14}$$

In equilibrium alternating renewal process, the density function of $\widetilde{T_I}$ is $\lambda_I[1-F_{T_I}(x)]$, where $F_{T_I}(x)$ is the CDF of T_I and $\lambda_I^{-1} = \int_0^\infty x * F_{T_I}(dx) = \int_0^\infty (1-F_{T_I}(x))dx$. When the inter-contact time T_I follows exponential distribution with parameter λ_I, $\widetilde{T_I}$ is identically distributed with T_I. By $E(N_{\mathcal{C}_\tau}) = E(\mathcal{N}_{\mathcal{C}_\tau}(\tau))$, we prove Eq. (4.13).

Remark 28 *Theorem 12 shows that the expected cloudlet size follows negative exponential growth with τ, which is consistent with our observation in Figures 4.10 and 4.11. Moreover, Theorem 12 gives the negative exponential growth rate as $\lambda_I = 1/E(T_I)$. The smaller $E(T_I)$ is, the more nodes an initiator encounters within time τ, and vice versa. This implies that the more frequently nodes meet one another and the larger the network size n is, the more devices are in the mobile cloudlet to provide computing resources and the better computing performance is likely to be achieved, and vice versa.*

4.5.2 Lifetime Analysis

We study a cloudlet node's lifetime based on the modified alternating renewal process in Definition 14, and deduce the following theorem.

Theorem 13 *The expected lifetime of a cloudlet node is approximately* $\tau - \frac{1}{\lambda_I}(1 - \frac{\pi r^2}{n/\lambda} + \frac{\lambda_C}{\lambda_I + \lambda_C})$ *when* τ *is large.*

Proof : In equilibrium alternating renewal process in Definition 14, $\chi(t) = 1$ when two nodes are in contact at time t; $\chi(t) = 0$, otherwise.

(i) When $\chi(0) = \chi(\tau) = 1$, the node's entrance time is 0 and its exit time is τ. Clearly, its lifetime $LT(\tau) = \tau$.

(ii) When $\chi(0) = 0, \chi(\tau) = 1$, the node's entrance time is ξ^1 and its exit time is τ. Then, $LT(\tau) = \tau - \xi^1 \cdot 1_{\{\xi^1 < \tau\}}$, where $\xi^1 = \widetilde{T_I}$ is the forward recurrence time of T_I.

(iii) When $\chi(0) = 1, \chi(\tau) = 0$, the node's entrance time is 0 and its exit time is $S_{N(\tau)}$. Thus, $LT(\tau) = S_{N(\tau)} = \tau - \xi^{N(\tau)} \cdot 1_{\{\xi^{N(\tau)} < \tau\}}$, where $\xi^{N(\tau)} = \widehat{T_I}$ is the backward recurrence time of T_I.

(iv) When $\chi(0) = 0, \chi(\tau) = 0$, if $N(\tau) = 0$, the node's lifetime is 0; if $N(\tau) > 0$, the node's entrance time is ξ^1 and its exit time is $S_{N(\tau)} + \eta^{N(\tau)}$. Hence, lifetime equals $[\tau - (\xi^1 + \xi^{N(\tau)+1}) \cdot 1_{\{\xi^1 + \xi^{N(\tau)+1} < t\}})] \cdot 1_{\{\xi^1 < t\}}$, where $\xi^1 = \widetilde{T_I}$ and $\xi^{N(\tau)+1} = \widehat{T_I}$.

Denote $\pi_{ij}(t)$ as the equilibrium probability, given that $\chi(0) = i$ and $\chi(t) = j$ $(i, j = 0, 1)$. Let p_0 and p_1 denote $P(\chi(0) = 0)$ and $P(\chi(0) = 1)$, respectively. Because T_I and T_C are exponential random variables with parameters λ_I and λ_C, respectively, $\widetilde{T_I}$ and $\widehat{T_I}$ have the same distribution as T_I and $\widetilde{T_I} + \widehat{T_I}$ follows Erlang-2 distribution Erlang(2,λ_I). Thus,

$$\begin{aligned} E(LT(\tau)) &= \tau^2 \lambda_I e^{-\lambda_I \tau} \pi_{00}(\tau) p_0 \\ &+ \tau[1 + (\pi_{01}(\tau)p_0 + \pi_{10}(\tau)p_1 + \pi_{00}(\tau)p_0)e^{-\lambda_I \tau}] \\ &- \frac{1}{\lambda_I}(1 - e^{-\lambda_I \tau})(\pi_{01}(\tau)p_0 + \pi_{10}(\tau)p_1 + 2\pi_{00}(\tau)p_0), \end{aligned} \tag{4.15}$$

where $p_1 = \frac{\pi r^2}{n/\lambda}$ and $p_0 = 1 - p_1$. The equilibrium probability $\pi_{ij}(\tau)$ can be derived based on Cox's Renewal Theory (Chapter 7.4) [106]: $\pi_{00}(\tau) = \beta + \gamma e^{-\beta\tau/\lambda_C}$, $\pi_{01}(\tau) = \gamma - \gamma e^{-\beta\tau/\lambda_C}$, $\pi_{10}(\tau) = \beta - \beta e^{-\beta\tau/\lambda_C}$, and $\pi_{11}(\tau) = \gamma + \beta e^{-\beta\tau/\lambda_C}$, where $\beta = \frac{\lambda_C}{\lambda_I + \lambda_C}$ and $\gamma = \frac{\lambda_I}{\lambda_I + \lambda_C}$. When τ is large, $e^{-\lambda_I \tau}$ and $e^{-(\lambda_I + \lambda_C)\tau}$ approach 0,

$$E(LT(\tau)) \approx \tau - \frac{1}{\lambda_I}(1 - \frac{\pi r^2}{n/\lambda} + \frac{\lambda_C}{\lambda_I + \lambda_C}), \tag{4.16}$$

i.e., the expected lifetime grows linearly with time τ. To better understand $E(LT(\tau))$, we have numerical analysis of $E(LT(\tau))$ to show how $E(LT(\tau))$ changes with τ in Figure 4.16. We set $p_0 = 0.9$, $p_1 = 0.1$, $\lambda_I = 0.0001$, and $\lambda_C = 0.01$. Parameters λ_I and λ_C are set to be approximately equal to $1/E(T_I)$ and $1/E(T_C)$ in trace Exp3. Figure 4.16 shows that when $\tau > 4 \times 10^4$, $E(LT(\tau))$ grows linearly with slope 1, which is consistent with Eq. (4.16) and Figures 4.13. When τ is small $(0 < \tau < 1000)$, the close-up figure shows that $E(LT(\tau))$ is

mainly influenced by τ^2.

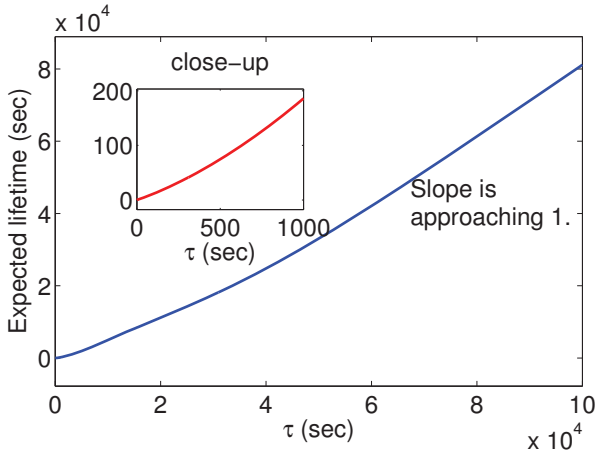

Figure 4.16: Expected lifetime of a cloudlet node grows linearly with slope 1 when τ is large.

Remark 29 *When τ is small, $E(LT(\tau))$ exhibits* quadratic growth. *As τ increases, $E(LT(\tau))$ tends to grow* linearly *with slope 1 and the gap between τ and $E(LT(\tau))$ is a constant $\frac{1}{\lambda_I}(1-\frac{\pi r^2}{n/\lambda}+\frac{\lambda_C}{\lambda_I+\lambda_C})$. This indicates that for application with long delay tolerance τ, intermittent connectivity between an initiator and cloudlet nodes has a small constant negative effect on the optimal computing performance achieved when cloudlet nodes compute task for the initiator throughout their lifetimes.*

4.5.3 Reachable Time Analysis

In order to study the reachable time $RT(\tau)$, i.e., total contact duration between two nodes within time τ, we rewrite Definition 14 as follows.

Definition 15 *Let $\{(I_n, Z_n\}_{n=0}^{\infty}$ be a bivariate stochastic process on a probability space (Ω, χ, P) such that $Z_0 = 0$ and $I_n = 1 - I_{n-1}$ for all $n \geq 1$, where $p_0 = P(I_0 = 0)$ and $p_1 = P(I_0 = 1)$ satisfying $p_0 + p_1 = 1$. Assume that, conditional on I_{n-1}, the $Z_n - Z_{n-1}$ for all $n \geq 1$ are mutually independent. Let the conditional distribution of Z_1 be $F^1_{\widetilde{T_C}}$ ($F^1_{\widetilde{T_I}}$) if $I_0 = 1$ (0), and the distributions of $Z_n - Z_{n-1}$ conditioned on I_{n-1} be F_{T_C} (F_{T_I}) if $I_{n-1} = 1$ (0), for all $n \geq 2$. Distributions of $F^1_{\widetilde{T_C}}$ and $F^1_{\widetilde{T_I}}$ have density $\lambda_C[1 - F_{T_C}]$ and $\lambda_I[1 - F_{T_I}]$, respectively, where $\lambda_C^{-1} = \int_0^{\infty} F_{T_C}(dx)$ and $\lambda_I^{-1} = \int_0^{\infty} F_{T_I}(dx)$. Then, the point process characterized by $\{(I_n, Z_n)\}_{n=0}^{\infty}$, in which I_n is the point type and Z_n is the waiting time until the nth event, is an* equilibrium alternating renewal process.

The total contact duration between two nodes within time τ depends on not only distribution of contact time but also the number of contacts, which can be represented as follows.

Definition 16 *Let $\{(I_n, Z_n\}_{n=0}^{\infty}$ be an equilibrium renewal alternating process. For all $\tau > 0$, let $K(\tau) = \sum_{n=1}^{\infty} 1_{Z_{2n-1} \leq \tau}$, and $L(\tau) = \sum_{n=1}^{\infty} 1_{Z_{2n} \leq \tau}$. Then, $K(\tau)$ $(L(\tau))$ is the odd (even) counting random variable of the process at time τ.*

When $I_0 = 1$, the number of contacts is $K(\tau)$ (or $K(\tau) + 1$) if $K(\tau) = L(\tau)$ (if $K(\tau) \neq L(\tau)$). When $I_0 = 0$, the number of contacts is $L(\tau)$ (or $L(\tau) + 1$) if $K(\tau) = L(\tau)$ (if $K(\tau) \neq L(\tau)$). To find out $RT(\tau)$, we define the sojourn time of the equilibrium renewal alternating process as follows.

Definition 17 *Let $\{(I_n, Z_n\}_{n=0}^{\infty}$ be an equilibrium renewal alternating process with odd (even) counting random variable $K(\tau)$ $(L(\tau))$. Let $S_0 = T_0 = 0$, $S_n = \sum_{i=1}^{n}(Z_{2i-1} - Z_{2i-2})$, and $T_n = \sum_{i=1}^{n}(Z_{2i} - Z_{2i-1})$, for all $n \geq 1$. Then S_n (T_n) is called the nth sum of the odd (even) states of the process, $n \geq 0$. For all $\tau > 0$, let*

$$\alpha_0(\tau) = T_{L(\tau)} \cdot 1_{\{K(\tau)=L(\tau)\}} + [\tau - S_{K(\tau)}] \cdot 1_{\{K(\tau)\neq L(\tau)\}},$$

$$\alpha_1(\tau) = S_{K(\tau)} \cdot 1_{\{K(\tau)\neq L(\tau)\}} + [\tau - T_{L(\tau)}] \cdot 1_{\{K(\tau)=L(\tau)\}}.$$

Then $\alpha_0(\tau)$ $(\alpha_1(\tau)$ $)$ is called the sojourn time of the even (odd) states of the process. Clearly, $\alpha_0(\tau) + \alpha_1(\tau) = \tau$. The sojourn time in the on state of the process (i.e., total contact duration) during $(0, \tau]$ is given by

$$RT(\tau) = \alpha_0(\tau) \cdot 1_{\{I_0=0\}} + \alpha_1(\tau) \cdot 1_{\{I_0=1\}}. \tag{4.17}$$

Lemma 6 *For an equilibrium alternating renewal process, the total contact duration (i.e., $RT(\tau)$) satisfies*

$$E(RT(\tau)) = E(\alpha_0(\tau))p_0 + E(\alpha_1(\tau))p_1, \tag{4.18}$$

where $E(\alpha_0(\tau))$ and $E(\alpha_1(\tau))$ are inverse Laplace transform of Eqs. (4.19) and (4.20), respectively, p_1 and p_0 represent the probabilities that two nodes are originally in contact and out of contact, respectively.

Proof : In [110], M.H. Rossiter derived the sojourn time distribution in on state for a two-state system by applying Laplace transform and double Laplace transform. For an equilibrium renewal alternating process, the Laplace transform of the expected sojourn time in on state conditioning on $I_0 = 0$

$$L(E[\alpha_0(\tau)]; s) = \frac{[1 - F_{T_C}(s)]F_{\widetilde{T_I}}(s)}{s^2[1 - F_{T_C}(s)F_{T_I}(s)]}, \tag{4.19}$$

while conditioning on $I_0 = 1$

$$L(E[\alpha_1(\tau)]; s) = \frac{1 - F_{\widetilde{T_C}}(s) - (F_{T_C} - F_{\widetilde{T_C}})(s)F_{T_I}(s)}{s^2[1 - F_{T_C}(s)F_{T_I}(s)]}, \tag{4.20}$$

where $L\{ \cdot ; s\}$ represents the Laplace transform, and $F_X(s)$ is the Laplace transform of random variable X, i.e., $F_X(s) = \int_0^\infty e^{-sx}F(dx)$.

We have $E(\alpha_0(\tau))$ and $E(\alpha_1(\tau))$ by taking inverse Laplace transform of Eqs. (4.19) and (4.20). Substituting $E(\alpha_0(\tau))$ and $E(\alpha_1(\tau))$ in the expectation of Eq. (4.17) completes our proof.

Theorem 14 *In homogeneous network with uniform node distribution, if T_C and T_I follow exponential distributions with parameters λ_C and λ_I, respectively, the expected reachable time of a cloudlet node*

$$E(RT(\tau)) = \frac{\lambda_I \tau}{\lambda_I + \lambda_C} + \frac{\lambda_C p_1 - \lambda_I p_0}{(\lambda_I + \lambda_C)^2}(1 - e^{-(\lambda_I + \lambda_C)\tau}), \tag{4.21}$$

where $p_1 = \frac{\pi r^2}{n/\lambda}$ and $p_0 = 1 - p_1$.

Proof : For exponential random variable T_C (T_I), its forward recurrence time $F_{\widetilde{T_C}}$ ($F_{\widetilde{T_I}}$) also has exponential distribution with parameter λ_C (λ_I) because of the memoryless property of exponential random variable. Then, $F_{T_C}(s) = F_{\widetilde{T_C}}(s) = \frac{\lambda_C}{s+\lambda_C}$ and $F_{T_I}(s) = F_{\widetilde{T_I}}(s) = \frac{\lambda_I}{s+\lambda_I}$.

Based on results in Eqs. (4.19) and (4.20) from [110],

$$L(E[\alpha_1(\tau)]; s) = \frac{1 - \frac{\lambda_C}{s+\lambda_C}}{s^2[1 - \frac{\lambda_C}{s+\lambda_C}\frac{\lambda_I}{s+\lambda_I}]},$$

and $L(E[\alpha_0(\tau)]; s) = L(E[\alpha_1(\tau)]; s)\frac{\lambda_I}{s+\lambda_I}$. Performing inverse Laplace transform, we have

$$E[\alpha_0(\tau)] = \frac{\lambda_I \tau}{\lambda_I + \lambda_C} + \frac{\lambda_I}{(\lambda_I + \lambda_C)^2}(e^{-(\lambda_I + \lambda_C)\tau} - 1), \tag{4.22}$$

$$E[\alpha_1(\tau)] = \frac{\lambda_I \tau}{\lambda_I + \lambda_C} + \frac{\lambda_C}{(\lambda_I + \lambda_C)^2}(1 - e^{-(\lambda_I + \lambda_C)\tau}). \tag{4.23}$$

In our homogeneous network model, $p_0 = \frac{\pi r^2}{n/\lambda}$ and $p_1 = 1 - p_0$. Substituting them into Eq. (4.18) completes our proof.

Numerical results of Theorem 14 are shown in Figure 4.17. The parameter settings are the same as those in Figure 4.16. Theorem 14 and Figure 4.17 show that when T_C and T_I are exponentially distributed, the reachable time of a cloudlet node grows linearly with slope $\frac{\lambda_I}{\lambda_I+\lambda_C}$, i.e., $\frac{E(T_C)}{E(T_I)+E(T_C)}$ as $\lambda_I = \frac{1}{E(T_I)}$ and $\lambda_C = \frac{1}{E(T_C)}$. The linear growth in Figure 4.17

is different from the piecewise linear growth in Figures 4.14 and 4.15 because depending on people's schedules, T_I and T_C follow different distributions during different time in traces Exp2 and Exp3.

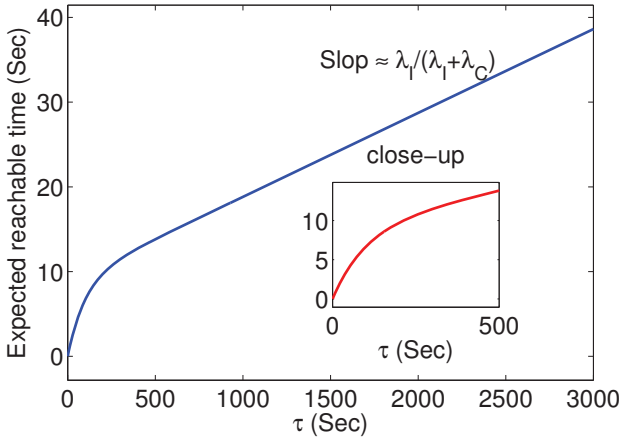

Figure 4.17: Expected reachable time of a cloudlet node grows linearly with slope approximately $\lambda_I/(\lambda_I + \lambda_C)$ when τ is large.

Remark 30 *The increase rate of $E(RT(\tau))$ shows that the mean reachable time within time τ is mainly determined by the ratio $\frac{E(T_C)}{E(T_I)+E(T_C)}$ when τ is large. Nodes that meet an initiator frequently and have long contact time have high connection likelihood and can provide reliable computing services while still support mobility of the initiator. The initiator can estimate the ratio $E(T_C)/(E(T_I) + E(T_C))$ based on its contact histories and use it as an indicator for whether an encountered node is suitable for providing mobile application services.*

4.6 Computing Capacity and Speed of Mobile Cloudlet

The amount of computation that the cloudlet nodes can provide for the initiator not only depends on the computing capabilities of cloudlet nodes and how the task is partitioned for parallel processing but also depends on the cloudlet node's lifetime and reachable time. Evaluating the computing capability of mobile processor and designing application partition schemes [111] are beyond the scope of this study. In this chapter, we simply assume that computing speed of each device is a constant V, and the initiator can partition the task into M subtasks that can be computed on cloudlet nodes in parallel. The data sizes of each subtask before and after

computing are D_d^i and D_r^i $(1 \leq i \leq M)$, respectively.

Assume a total of B Hz spectrum is shared by all nodes and each node has a fixed transmission power P. The noise N—including ambient and interference noise—is constant everywhere in the network. We characterize the wireless link using a pass loss model with attenuation exponent $\alpha \geq 2$. The capacity of a wireless link is $B\log_2(1+\frac{P}{N}d^{-\alpha})$, where d is the Euclidean distance between the sender and the receiver. Assume that advanced error control coding is used such that the available link bandwidth is equal to its capacity. For a task contains D_d^i $(1 \leq i \leq M)$ bits, the transmission time of dispatching a task is

$$0 < t_d = \frac{D_d^i}{B\log_2(1+\frac{P}{N}d^{-\alpha})} \leq \frac{D_d}{B\log_2(1+\frac{P}{N}r^{-\alpha})}, \tag{4.24}$$

where $D_d \triangleq \sum_{1\leq i\leq M} D_d^i$. The transmission time of retrieving a task is

$$0 < t_r = \frac{D_r^i}{B\log_2(1+\frac{P}{N}d^{-\alpha})} \leq \frac{D_r}{B\log_2(1+\frac{P}{N}r^{-\alpha})}, \tag{4.25}$$

where $D_r \triangleq \sum_{1\leq i\leq M} D_r^i$. Then the total transmission time is

$$0 < t_d + t_r \leq \frac{D_d + D_r}{B\log_2(1+\frac{P}{N}r^{-\alpha})} \triangleq \rho. \tag{4.26}$$

If tasks are computed on cloudlet nodes during their whole lifetime, the optimal mobile cloudlet capacity is achieved. If an initiator only employs cloudlet nodes when they are in contact, i.e., the computing times equal to reachable times minus transmission times, we have a lower bound on the computing capacity of a mobile cloudlet. Based on this methodology and results in previous section, we have the following theorem on the computing capacity of a mobile cloudlet.

Theorem 15 *A mobile cloudlet $\mathcal{C}_\tau$'s expected computing capacity is upper bounded by $C_{\mathcal{C}_\tau}^u$ and lower bounded by $C_{\mathcal{C}_\tau}^l$ in Eqs. (4.27) and (4.30), respectively.*

Proof : During a cloudlet node's lifetime, the maximum time a cloudlet node can use for computing is its lifetime minus the transmission time of the task. Hence, the computing capacity of a mobile cloudlet $\mathcal{C}_\tau$ satisfies

$$C_{\mathcal{C}_\tau} \leq \sum_{i=1}^{N_{\mathcal{C}_\tau}} (LT^i(\tau) - t_r - t_d)V.$$

Hence,

$$E(C_{\mathcal{C}_\tau}) < C_{\mathcal{C}_\tau}^u \triangleq E(N_{\mathcal{C}_\tau})E(LT(\tau))V, \tag{4.27}$$

where $E(N_{\mathcal{C}_\tau})$ and $E(LT(\tau))$ can be found in Eqs. (4.13) and (4.15), respectively.

To employ cloudlet nodes in contact, a cloudlet node's total computing time over time τ is

$$\begin{aligned} CT(\tau) =& RT(\tau) - (p_0 + p_1\pi_{10}) \sum_{i=1}^{N(\tau)} \overline{CT^i} \\ & - p_1\pi_{11} P(T_C < \tau) \sum_{i=1}^{N(\tau)+1} \overline{CT^i}, \end{aligned} \tag{4.28}$$

where $\overline{CT^i} \triangleq T_C^i \cdot 1_{T_C^i<\rho} + \rho \cdot 1_{T_C^i \geq \rho}$, $\pi_{i,j}$ $(i,j = 0,1)$ are the equilibrium probabilities and $N(\tau)$ is number of renewals within time τ in the equilibrium alternating renewal process.

Denote $N^0(\tau) = N(\tau)|\{I_0 = 0\}$ and $N^1(\tau) = N(\tau)|\{I_0 = 1\}$. According to the renewal equation for modified renewal process,

$$E(N^0(\tau)) = F_{\widetilde{T_I}}(\tau) + \int_0^\tau E(N^0(\tau)) dF_{T_I+T_C}(s),$$

$$E(N^1(\tau)) = F_{\widetilde{T_C}}(\tau) + \int_0^\tau E(N^1(\tau)) dF_{T_I+T_C}(s).$$

Taking the Laplace transform on both sides,

$$L_{E(N^0(\tau))}(s) = L_{\widetilde{T_I}}(s)/s + L_{N^0(\tau)}(s) L_{T_I+T_C}(s),$$

$$L_{E(N^1(\tau))}(s) = L_{\widetilde{T_C}}(s)/s + L_{N^1(\tau)}(s) L_{T_I+T_C}(s),$$

where $L_{\widetilde{T_I}}(s) = \lambda_I/(s+\lambda_I)$ and $L_{\widetilde{T_C}}(s) = \lambda_C/(s+\lambda_C)$. T_I+T_C has density function $\frac{\lambda_I\lambda_C}{\lambda_C-\lambda_I}(e^{-\lambda_I t} - e^{-\lambda_C t})$, which gives the Laplace transform $L_{T_I+T_C}(s) = \lambda_I\lambda_C/(s+\lambda_I)(s+\lambda_C)$. Thus, $L_{E(N^0(\tau))}(s) = \lambda_I(s+\lambda_C)/[s^2(s+\lambda_I+\lambda_C)]$ and $L_{E(N^1(\tau))}(s) = \lambda_C(s+\lambda_I)/[s^2(s+\lambda_I+\lambda_C)]$. Taking the inverse Laplace transform, we then have

$$E(N^0(\tau)) = \gamma\lambda_C\tau + \gamma^2(1 - e^{-(\lambda_I+\lambda_C)\tau}),$$

$$E(N^1(\tau)) = \beta\lambda_I\tau + \beta^2(1 - e^{-(\lambda_I+\lambda_C)\tau}),$$

where $\beta = \frac{\lambda_C}{\lambda_I+\lambda_C}$ and $\gamma = \frac{\lambda_I}{\lambda_I+\lambda_C}$. Subsequently,

$$\begin{aligned} E(N(\tau)) &= E(N(\tau)|I_0 = 0)p_0 + E(N(\tau)|I_0 = 1)p_1, \\ &= \frac{\lambda_I\lambda_C\tau}{\lambda_I+\lambda_C} + \frac{p_0\lambda_I^2 + p_1\lambda_C^2}{(\lambda_I+\lambda_C)^2}(1 - e^{-(\lambda_I+\lambda_C)\tau}), \end{aligned} \tag{4.29}$$

where $p_1 = \frac{\pi r^2}{n/\lambda}$ and $p_0 = 1 - p_1$.

The computing capacity of a mobile cloudlet $\mathcal{C}_\tau$ satisfies

$$C_{\mathcal{C}_\tau} \geq \sum_{j=1}^{N_{\mathcal{C}_\tau}} CT^j(\tau)V.$$

Therefore, $E(C_{\mathcal{C}_\tau})$ is lower bounded by

$$C^l_{\mathcal{C}_\tau} \triangleq E(N_{\mathcal{C}_\tau})V \left\{E(RT(\tau)) - \left[(1 - p_1\pi_{11}e^{-\lambda_C\tau})E(N(\tau)) + p_1\pi_{11}(1 - e^{-\lambda_C\tau})\right] E(\overline{CT})\right\}, \tag{4.30}$$

where $E(\overline{CT}) = E(T_C \cdot 1_{T_C<\rho}) + \rho P(T_C \geq \rho) = \frac{1-e^{-\lambda_C\rho}}{\lambda_C}$.

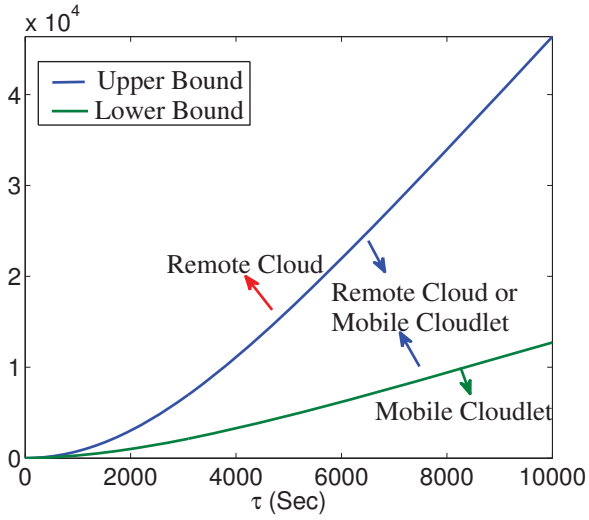

Figure 4.18: Bounds on computing capacity of mobile cloudlet where $\lambda_I = 0.0002$, $\lambda_C = 0.001$, $p_0 = 0.9$, $p_1 = 0.1$, $n = 10$, $V = 1$, $\rho = 0.1$.

Figure 4.18 shows the numerical results of this theorem by setting $\rho = 0.1$ second (typical packet transmission time in mobile wireless networks), $n = 10$, $E(T_I) \approx 83$ minutes, and $E(T_C) \approx 17$ minutes. This parameter setting reflects a scenario with 10 students in a team or 10 colleagues at a conference working on the same project. They meet for about 17 minutes between classes or conference sessions to compute a common task by sharing computing resources on their mobile devices. Two curves in Figure 4.18 divide MCC into three categories: MCC relying on i) remote cloud, ii) remote cloud or mobile cloudlet, iii) mobile cloudlet. When this group of users need to execute a task with computational demand C and delay requirement τ, if

$C \geq C^u_{\mathcal{C}_\tau}$, they should offload the task to remote cloud; if $C^l_{\mathcal{C}_\tau} < C < C^u_{\mathcal{C}_\tau}$, they can use remote cloud or mobile cloudlet; if $C \leq C^l_{\mathcal{C}_\tau}$, they can simply share the task computation during their contacts with each other for mobile cloudlet computing.

The average computing speed of a mobile cloudlet is $E(C_{C_\tau})/\tau$. When $\tau \to \infty$, we have the long-term computing speed $CS = \lim_{\tau\to\infty} \frac{E(C_{C_\tau})}{\tau}$.

Theorem 16 *The long-term computing speed of a mobile cloudlet is upper bounded by* $CS^u = (n-1)V$ *and lower bounded by* $CS^l = \frac{(n-1)V\lambda_I e^{-\lambda_C \rho}}{\lambda_C+\lambda_I}$.

Proof : Based on Theorem 15, we have the upper bound of CS,

$$CS \leq \lim_{\tau\to\infty} \frac{C^u_{\mathcal{C}_\tau}}{\tau} = (n-1)V \triangleq CS^u. \tag{4.31}$$

Similarly, the lower bound of CS is $\lim_{\tau\to\infty} \frac{C^l_{\mathcal{C}_\tau}}{\tau}$. As in modified renewal process,

$$\lim_{\tau\to\infty} \frac{E(N(\tau))}{\tau} = \frac{1}{E(T_C + T_I)} = \frac{\lambda_C\lambda_I}{\lambda_C+\lambda_I}.$$

Accordingly, we have

$$CS \geq \lim_{\tau\to\infty} \frac{C^l_{\mathcal{C}_\tau}}{\tau} = \frac{\lambda_I e^{-\lambda_C\rho}}{\lambda_C+\lambda_I}(n-1)V \triangleq CS^l. \tag{4.32}$$

Remark 31 *The bounds on long-term computing speed can also be used by an initiator mobile device to decide where to offload its task for computing service. Suppose the initiator has a task with computational demand* C *and delay requirement* τ*, if* $C/\tau \geq (n-1)V$*, the initiator needs to offload its task to a remote cloud; if* $C/\tau \leq \frac{\lambda_I e^{-\lambda_C\rho}}{\lambda_C+\lambda_I}(n-1)V$*, the initiator can distribute its task to nearby devices in a mobile cloudlet; otherwise, the initiator can choose either remote cloud or mobile cloudlet based on other constraints, such as battery life and quality of wireless communication.*

4.7 Summary

In this chapter, we study the cloudlet computing for mobile applications, in which mobile users offload tasks to nearby resource-rich devices for instant service access and saving on roaming charges. If a cloudlet locates at a community site that a mobile user visits, its access probability for this mobile user is $\mu_{T_C}/(\mu_{T_I} + \mu_{T_C})$ determined by the mobile user's mean connection time and inter-connection time with the cloudlet. Moreover, cloudlet access probability affects the cloudlet computing performance, such as probability of successful task execution, average number of tasks executed over time t, and the limiting task execution speed. In summary,

a mobile user's mobility pattern has significant impact on its cloudlet access probability and cloudlet computing performance, which can be measured by $\mu_{T_C}/(\mu_{T_I} + \mu_{T_C})$. In the domain of mobile cloudlet, we examine the properties and computing capacity of mobile cloudlet in order to determine whether mobile cloudlet can be used for mobile applications. The negative exponential growth of cloudlet size shows that the number of resource-rich devices an initiator can connect to for computing service is determined by the number of nodes in the network and how frequently they meet. When the task is delay-tolerant, the intermittent connection has little negative effect on optimal performance of a mobile cloudlet. Furthermore, $E(T_C)/(E(T_C)+E(T_I))$ implies the connection likelihood of a cloudlet node to an initiator, thus can be used by the initiator to choose reliable cloudlet node. Based on cloudlet properties, we have also derived upper and lower bounds on the computing capacity and long-term computing speed of a mobile cloudlet. An initiator can use these bounds to decide whether to upload a task to remote clouds or utilize nearby mobile cloudlet.

Chapter 5

Assessing Content Sharing through D2D Communications

Owing to the development of wireless technologies and smart devices, mobile users can conveniently access the enormous contents on the Internet, which is generating large amount of mobile data traffic and quickly leading to overloaded cellular network. In order to offload mobile data traffic, device-to-device (D2D) communication is explored for content delivery. However, adoption of this approach remains elusive due to various challenges of D2D communications. In this chapter, aiming to lift the fog of opportunistic content delivery, we explore the potential for content delivery through D2D communications. We first demonstrate through a study of a YouTube traffic dataset that practical opportunities exist for peer-assisted content delivery because of the temporal locality of user requests and the stochastic dominance of device contact time over content transmission time. Then, we derive the number of copies of a content in the network under different content caching policies. Finally, we evaluate the content fetching and sharing probabilities and show that D2D communication can indeed reduce network load especially for popular content delivery.

5.1 Motivation and Related Work

With the emergence of mobile devices and the development of wireless technologies, users can browse websites, listen to music, and watch videos while on the move. As a result of this, mobile data traffic has been growing at a phenomenal pace. Cisco [31] reported that global mobile data traffic grew 70 percent and mobile video traffic exceeded 50 percent in 2012. Such tremendous increase in mobile data traffic is predicted to overload cellular networks.

In order to mitigate the pressure on cellular networks, mobile data offloading is introduced to deliver contents through complementary network technologies. WiFi and femtocells are the

preferred offloading technologies [30]. However, the coverage of WiFi hotspots and femtocells may be constrained by their limited deployment, and mobile users may not be able to fetch contents through them. Thus, opportunistic D2D communication is also explored for content delivery. When two nearby users request the same content around the same time, one user can download the content from content servers through infrastructure networks while the other can fetch the content from the first user through D2D communications.

There are many challenges in D2D communications, such as the battery and storage constraints of mobile devices, the unpredictable device connectivity and user cooperation, and the complexity of interference handling and transmission scheduling. Nevertheless, peer-assisted content delivery is an attractive solution to mobile data offloading. Nowadays, the content popularity distribution is more skewed. In other words, a small number of websites or videos are extremely popular, drawing many users to access them. Moreover, IEEE 802.15 TG8 (Peer Aware Communication) [112] is formed to provide a global standard for scalable, low power, and reliable wireless D2D communications for emerging mobile services. These factors present promising potential for delivery of contents (especially popular contents) through D2D communications.

Existing studies [5–8] have investigated mobile data offloading through D2D communications. Han et al. [5] propose to exploit opportunistic communications to facilitate information dissemination in Mobile Social Networks (MoSoNets). The service providers can first deliver the information to only a small fraction of target-users; the target-users then propagate the information to other users through their contacts, thus offloading mobile data traffic. Similarly, Peng et al. [6] propose to let a set of socially important users download the content through cellular link and forward the content to their acquaintances when they are in contact. Cai et al. [7] exploit the double opportunities for content propagation in wireless networks, where the content could propagates to the users directly from the central controller or by exchange with other mobiles in a peer-to-peer manner. These works have not explored content popularity distribution in the system. Besides offloading cellular network, D2D communications can also alleviate the burden of content servers. Ciullo et al. [8] show that users' cooperation can dramatically reduce the bandwidth requirement of servers for perfect video-on-demand delivery. However, this work does not consider the dynamics of device connectivity due to user mobility. There still lacks a comprehensive understanding of content sharing through D2D communications in the existing literature.

In this chapter, we aim to demystify the opportunities that emerge from opportunistic D2D communications for content delivery by jointly considering content popularity distribution, content caching at mobiles, and user mobility. Our contributions are three-fold.

1. We first analyze a video request dataset to determine whether there exist practical op-

portunities for content delivery through D2D communications. Our trace analysis demonstrates that the skewed content popularity distribution, the clustered user request pattern, and the dominated mobile device contact time over content transmission time together offer practical opportunities for successful content sharing through D2D communications.

2. Based on models of user request pattern and content caching policy, we derive the number of users $M(t)$ that have a content in the network. When users cache a content for a constant time CT, expectation of $M(t)$ equals $n(1-e^{\frac{-pt^*}{\mu}})$, where $t^* = \min\{t, CT\}$, p is the content request rate, and μ is the mean inter-arrival time of a user's content requests. When users cache the most K recently requested contents, $E(M(t))$ goes to $n[1-(1-p)^K]$ at steady state. These results show that content popularity and caching time (size) determine the availability of a content on devices in the vicinity.

3. We obtain the possibility p_p of successfully fetching a content from peers, which is determined by factor $q\beta_\tau p_c$, where content caching probability $p_c = M(t)/n$, β_τ depends on nodes' inter-contact time and delay tolerance τ, and q is the success rate of content transmission. Numerical results reveal that content popularity and delay tolerance of content request have significant impact on p_p. Finally, we study the number of users that will fetch a content from a providing node, which is determined by ratio $\frac{p}{\mu\lambda_C}$ as well as the number of encountered nodes within content caching time. Our analysis validates that D2D communications can indeed reduce network load especially for delivering popular contents.

The rest of this chapter is organized as follows. Trace analysis of video requests in Section 5.2 demonstrates the existence of practical opportunities for content sharing through D2D communications. Theoretical analysis of content distribution and peer-assisted content fetching is performed in Sections 5.4 and 5.5, respectively. Section 5.7 concludes this chapter.

5.2 Do Opportunities Exist for Content Sharing through D2D Communications?

In this section, we analyze the characteristics of network traffic using a YouTube traffic dataset. Zink et al. [113] collected six traces of YouTube traffic in a large university campus network. These traces contain information about client requests for YouTube video clips, including video ID, the time that videos are requested, YouTube server IP, and client IP. The data were gathered between May, 2007 and March, 2008, spanning a period of 10 months. Using this dataset, we study content popularity distribution, request inter-arrival time, and content transmission time. Rather than serving as a comprehensively study on characteristics of mobile data traffic, our

trace analysis is meant to validate the existence of practical opportunity for content sharing through D2D communications.

5.2.1 Content Popularity

One of the design criteria for content delivery systems is the content popularity, which is represented by the number of user requests. Obtaining and analyzing the popularity of a file enables network designers to decide which contents to cache. Content popularity also influences the number of copies of a content in a network, thus affects the potential benefits of content sharing through D2D communications.

We study content popularity through a trace from Zink's dataset: youtube.parsed.012908.dat, which records YouTube traffic of two weeks. Figure 5.1 shows the distribution of video requests during two weeks on a log-log scale. The x-axis shows videos reverse sorted by the number of downloads. We get a reasonably good fit for a Zipf distribution, which means that a few videos are very popular receiving many user downloads. This finding has also been observed in other studies, such as [113–115], thus it is reasonable to assume Zipf content popularity in our analysis. Such skewed distribution of content popularity implies that users can likely share popular contents with nearby peers.

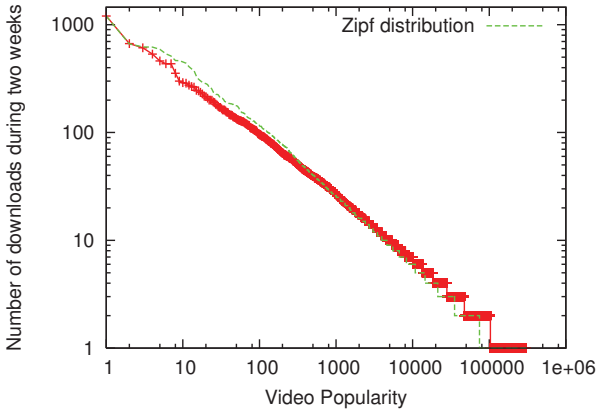

Figure 5.1: The videos are ranked according to the number of downloads (ranging within $[1, 1454]$) over two weeks.

5.2.2 Content Server Distribution

In Zink's dataset, users download YouTube videos from cloud servers. The number of servers that store a content indicates how likely the network and cloud servers can be overloaded.

In Figure 5.2, we plot the number of content servers that provide the downloads of a video. An interesting finding is that users fetch a video at most from four servers. No matter how many clients request a video, the video is most likely fetched from one or two content servers. Furthermore, when the number of clients that request a video increases, the likelihood that the video is downloaded from one server is decreasing to 0 while from two servers is increasing to 1. Fetching popular videos only from one or two servers may lead to overloaded servers. This reveals a need to offload server and network loads, especially for delivery of popular contents, by content sharing through D2D communications.

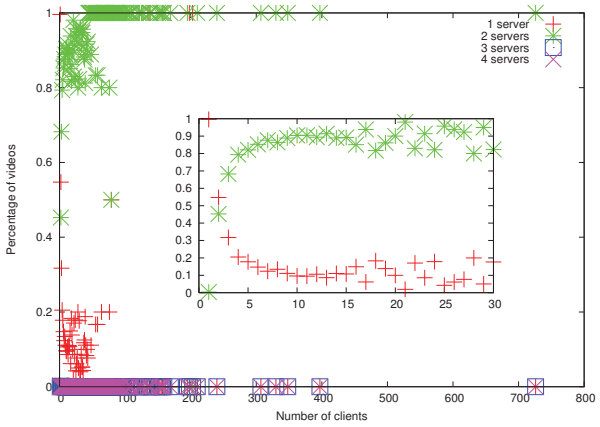

Figure 5.2: The percentage of videos that are downloaded from k ($k = 1, 2, 3, 4$) content servers when each video is requested by x clients.

5.2.3 Spatial and Temporal Locality of User Request Pattern

Characteristics of content requests are important not only because it represents users' request behaviors, but also because it allows us to learn users' potential of serving peers' content requests. For instance, when a client requests a content, it can download the content and cache the content. If other nearby users request the same content, they can fetch the content directly from the client that cached the content. Because mobile devices have limited cache size and can only cache contents for a limited period of time, content sharing through D2D communications becomes possible only if nearby users request same contents around the same time and the same geographic region. The spatial and temporal locality of content request are the key for content sharing among mobile users.

However, there is no trace that contains temporal and spatial information of mobile users' content requests due to privacy issue. In order to explore the spatial and temporal locality

of content request, we integrate content request traces with user mobility traces. We choose to integrate the content request dataset with a campus mobility dataset. The Content request dataset [113] contains information about client requests for YouTube video clips, including video ID, the time that the video is requested, YouTube server IP, content server IP, and client IP. On the other hand, campus phone dataset [116] gathers smartphones' various environmental data (e.g., detected wireless signal strengths, Bluetooth proximity) and communication patterns (SMS, e-mail, phone, and Facebook). But the content of said communications are not recorded. We believe that because both traces are collected in campus scenario, mobile users who produce such mobility trace (i.e., campus phone dataset) are likely generate content requests patterns found in the content request dataset.

Two major challenges lie in the integration of the content request dataset and the mobility dataset: user linkage and time synchronization. We choose 80 users from mobility dataset that on average update information at least once an hour over around a month period of time. We further randomly choose 80 clients from YouTube traffic dataset and randomly link one client with a user. Finally, we adjust the time stamps to synchronize these two traces by constructing a new time stamp for the new integrated trace. We denote T_s^1 and T_s^2 as the first location update in Phone dataset and content request in YouTube dataset. Suppose a client requests a content at time t_2 in the content request trace, the time stamp in the constructed new trace will be $t = t_2 - T_s^2$, and the location of the client is its corresponding user's location at $t_1 = t + T_s^1$ in the Phone dataset. By linking users and synchronizing times in these two traces, we produce a new integrated trace with information of userID, time, requested content, and location. Note that the locations in the Phone DataSet is in the form of GPS coordinates. We convert the locations into UTM coordinates.

Temporal Locality of Content Request

We choose the top 10 popular videos in the YouTube dataset and extract the inter-arrival time of requests directed to the same video. The red curve in Figure 5.3 shows that the inter-arrival time of requests for a video is smaller than 30 minutes with probability 0.8, and smaller than one and a half hour with probability 0.95. Such clustered requests directed to the same contents indicate a high likelihood that users request the same popular contents around the same time. The temporal locality in content request pattern shows great promise of using D2D communications to deliver popular contents.

Besides fetching a content from a neighbor, a user may also obtain a content from its own cache if a user repetitively requests the same content. We analyze the number of times that a video is requested by the same client and the inter-arrival time of users' repetitive requests, shown in Figures 5.4 and 5.3, respectively. Figure 5.4 reveals that about 22% of videos are

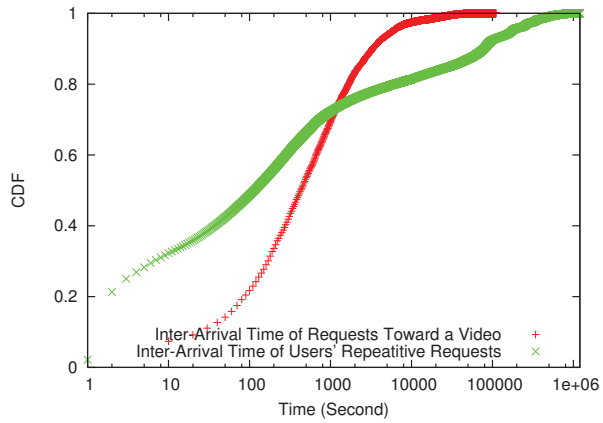

Figure 5.3: The time intervals between a client's two requests for the same videos and the time intervals of two requests directed to the same videos.

requested by the same user for at least two times. Furthermore, the green curve in Figure 5.3 shows that when a client requests a video multiple times, the time interval between two consecutive requests of the same video is less than 727 seconds (12 minutes) with 70% likelihood, less than 6160 seconds (102 minutes) with 80% likelihood, and less than 21.7 hours with 90% likelihood. Figures 5.3 and 5.4 together suggest that users could fetch contents from their own caches as users tend to request a video multiple times within a short period of time.

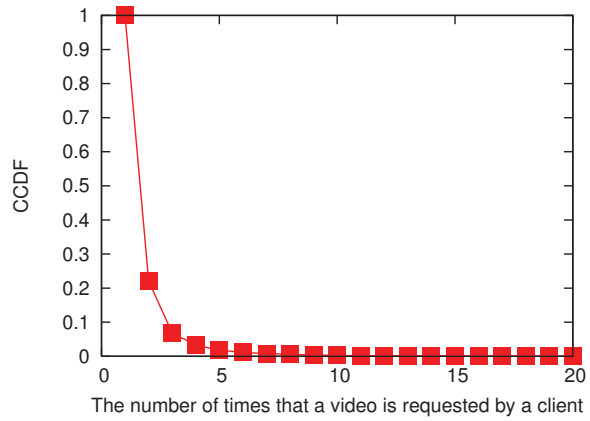

Figure 5.4: The number of times that a video is requested by the same client.

The short time interval between two requests of the same content ensures that content request of users can partly be served by neighbors or local caches. In other words, users can cache downloaded contents for a short period of time in order to facilitate their own recurrent

content requests as well as neighbors' requests of the same contents.

In addition, we plot the content request times of the 15 most popular contents from the integrated trace in Fig. 5.5. This figure shows that most contents' popularity last a very short time, i.e., most of user requests that a content receives are clustered within a few hours. A few content's popularity lasts a longer time (around 2 to 3 days).

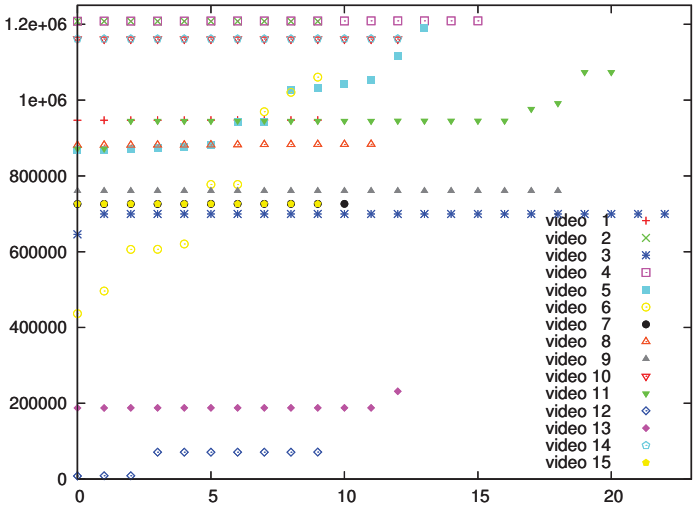

Figure 5.5: Request times of 15 videos.

Both the inter-arrival time of user requests in the YouTube traffic dataset and the user request time in the integrated trace indicate temporal locality in users' requests toward a content. Therefore, caching contents for a short period of time would lead to opportunities for content sharing through D2D communications.

Spatial Locality of Content Request

Furthermore, we choose 6 contents from 15 popular contents that received at least 10 requests. Fig. 5.6 shows that most of the contents are requested at certain areas, such as videos "9-rj3sHpbNY" and "BRX6N4cxFUs". However, the most popular video "SaH2M9-l4gY" is requested by users located at different places. This means that requests of global popular videos are requested at different locations in the network, while requests of regional popular videos are clustered at certain geographic areas.

On one hand, popular contents are widespread in the network, which is promising for content sharing through D2D communications. On the other hand, unpopular contents have spatial

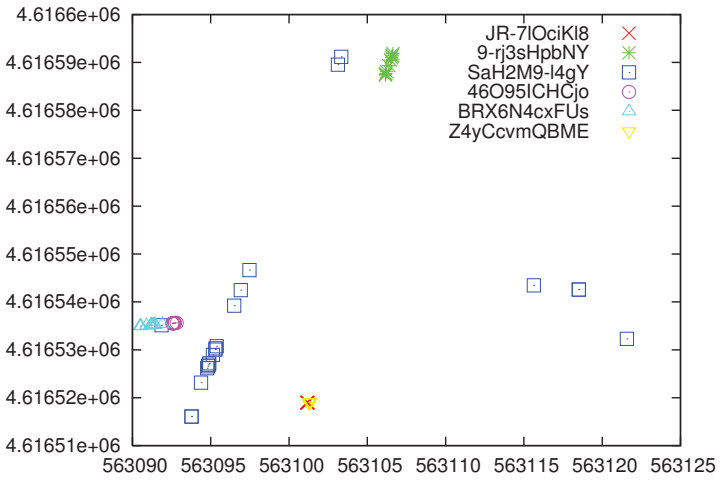

Figure 5.6: Locations where a video is requested.

locality in users' requests, which also produce opportunities for content sharing through D2D communications. Overall, the temporal and spatial locality of content requests make content sharing through D2D communications practical and promising.

5.2.4 Content Transmission Time Vs. Device Contact Time

When a user u requests a content o and its neighbor v has content o in cache, node u can successfully fetch the content from v only if their contact is longer than the transmission time of content o. In order to identify whether content sharing through D2D communications can be successful, we compare content transmission time with device contact duration.

Distributions of contact time between mobile devices are extracted from the *Cambridge* dataset [109]. The dataset includes one trace that records contacts of 12 graduate students from the System Research Group at University of Cambridge over 5 days in 2005 and and another trace that collects contacts of 41 students when they attend the student workshop at the IEEE Infocom Conference in 2005. We compare contact time in the Cambridge dataset with video transmission time in the YouTube dataset. The CCDF (complementary CDF) of the device contact time and the video transmission time are plotted in Figure 5.7.

Figure 5.7 shows that a large percentage of videos take a short transmission time. For example, transmissions of 60% of videos take less than 60 seconds and transmissions of 90% of videos need less than 160 seconds. Moreover, the CCDF of contact time has a heavier tail than the CCDF of video transmission time. When t is larger than 60 seconds, device contact time stochastically dominates video transmission time. More specifically, 56% (42%) of mobile devices' contacts last longer than 60 seconds and 40% (28%) of contacts last longer than 160

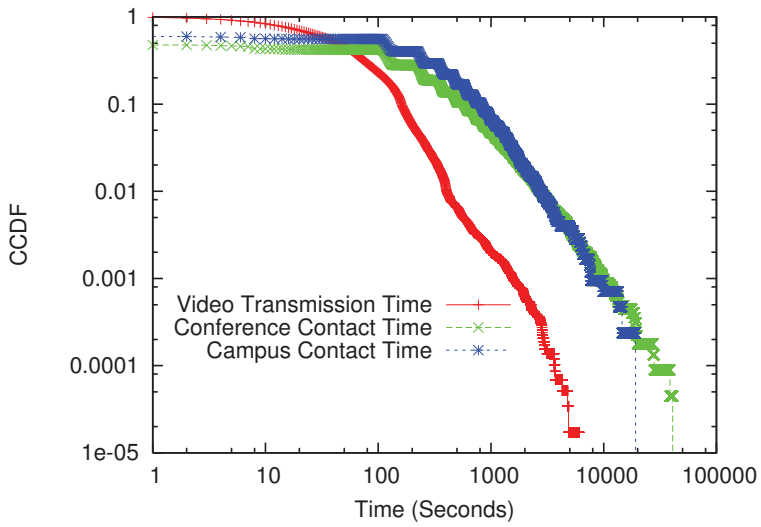

Figure 5.7: CCDF of video transmission time and mobile devices' contact time

seconds in campus (conference) scenario. There is a relatively high probability that most videos can be successfully transmitted during devices' contacts.

Remark 32 *Our trace analysis demonstrates that the skewed content popularity distribution, the temporal and spatial locality of user request pattern, and the dominated mobile device contact time over content transmission time together offer practical opportunities for successful content sharing through D2D communications.*

5.3 Models and Definitions

In this section, we present the network and mobility models, traffic model, and content caching model, which are essential to our analysis of content availability on mobile devices in the network and evaluation of the performance of content sharing through D2D communications.

5.3.1 Network and Mobility Model

We consider a content delivery network, shown in Figure 5.8. A network comprises n nodes moving over a region $\mathcal{S}$ with node density λ. The network also contains a number of infrastructure nodes (i.e., cellular base stations and WiFi access points) that can provide ubiquitous connectivity to the cloud servers. User can download a content either from content servers in the cloud through infrastructure wireless networks or from close-by mobile devices through D2D communications.

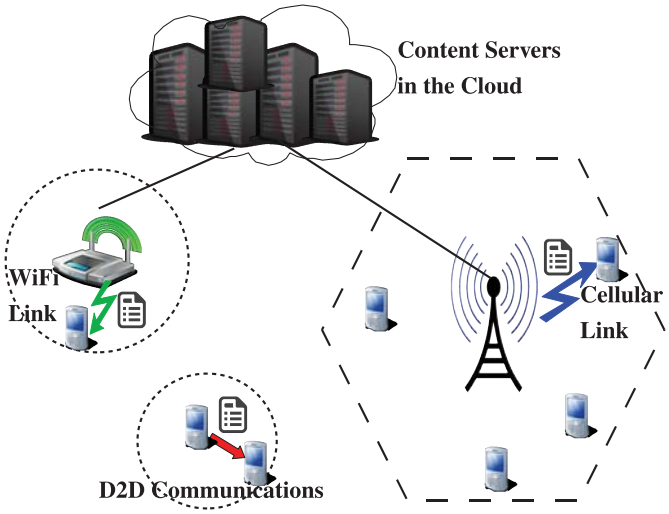

Figure 5.8: Content Delivery Network

We assume that each node's transmission range is r. We adopt the commonly used protocol model for the simplicity of our analysis. Transmission between node u and node v is feasible only if the distance between u and v is smaller than or equal to r. Denote by $\mathcal{X}_t = \{X_1(t), \ldots, X_n(t)\}$ the positions of users at time t. Nodes are moving according to mobility process $\mathcal{M}$. We assume that $\mathcal{M}$ is stationary and ergodic that a node's location $X_i(\cdot)$ has uniform stationary distribution in the network area. Mobility processes of nodes are independent and identically distributed (i.i.d.).

Let $X_u(t)$ and $X_v(t)$ denote the locations of users u and v at time t, we call that one *contact event* between users u and v occurs during $[t_0, t_1)$ if $||X_u(t_0^-) - X_v(t_0^-)|| > r$ and $||X_u(t) - X_v(t)|| \leq r$ for all $t \in [t_0, t_1)$, and $||X_u(t_1) - X_v(t_1)|| > r$. The *inter-contact time* is the time between the end and the start of two consecutive contact events between the same pair of users. Obtaining complete knowledge of contact process can be extremely difficult. Thus, we assume that contact time T_C and inter-contact time T_I follow exponential distributions with parameters λ_C and λ_I, respectively, which has been shown to be a good approximation and used by other existing studies (such as [109, 117]).

5.3.2 Traffic and Content Transmission Model

We denote the set of content items by $\mathcal{O}$ and the total number of content items by O. We use a Zipf's law for the content popularity distribution, which is observed in traffic measurements shown in Figure 5.1 and widely adopted in performance evaluation studies [8, 118]. This law implies that, having sorted the contents in decreasing order of popularity, a request is directed

to content o with probability

$$p_o = \frac{R(O)}{o^\gamma}, 1 \leq o \leq O, \tag{5.1}$$

where $R(O) \triangleq (\sum_{i=1}^{O} i^{-\gamma})^{-1}$ is a normalization constant, and γ is the Zipf's law exponent.

We denote the inter-arrival time of two consecutive requests from a user by random variable T. The content request process of a user can be modeled by a renewal process. Let $\{N(t), t \geq 0\}$ be a renewal process with inter-arrival times T_i, $i \geq 1$, and the time of the i-th renewal by $S_i = T_1 + \cdots + T_i$. Then, the random variable $N(t)$ $(t \geq 0)$ given by $N(t) = \sup\{i : S_i \leq t\}$ represents the number of requests of a user by time t.

We assume that all contents are stored in cloud servers, while the set of contents stored at each node is dynamically changing. When a node u requests a content o, it first tries to fetch the content from its own cache or a neighbor; if u could not get content o after delay τ $(\tau \geq 0)$, node u will download the content from cloud servers through infrastructure-based wireless networks. We restrict content sharing through D2D communications to occur over a single hop because multihop D2D communication is unreliable due to node mobility.

We denote ϵ as a content o's transmission time, which is determined by content's data size and network bandwidth. When user u has a content o in its cache, transmission of content o from node u to node v is feasible only if the connection time between u and v is equal to or larger than ϵ. Denote by q the success rate of a content transmission, which takes account into transmission time, transmission errors, and collisions, we have $q \leq P(T_C \geq \epsilon) = e^{-\lambda_C \epsilon}$.

5.3.3 Content Caching Model

The set of contents stored at each node is dynamically changing according to their caching policies as well as request patterns. Papers [119,120] give comprehensive surveys on web caching and replacement algorithms, which can be classified into *recency-based policy*, *frequency-based policy*, *size-based policy*, *function-based policy*, and *randomized policy*. In mobile cache-based content delivery network, how long a content is cached at nodes is the most important factor that influences the content delivery performance. Therefore, we define the following generic caching policy based on content caching time.

Definition 18 (Generic Caching (GC) Policy) *When a user requests a content o, it will cache this item for a period of time CT, which is a random variable dependent on caching and replacement algorithms, content popularity and size, user request pattern, and user's cache size.*

In order to facilitate analysis on content caching on mobile devices in a network, we further specify the following two cases.

Definition 19 (Constant Time Caching (CTC) Policy) *When a user downloads a content o, the user will cache this item for a constant period of time CT. After storing a content*

for CT time, the user will delete the content.

We assume that nodes have limited storage capacity, which is realistic as mobile devices are limited in memory size. Let K be the storage capacity of each node, measured in number of (equal-size) contents, we give a recency-based caching policy.

Definition 20 (Most Recent Caching (MRC) Policy) *When a user downloads a content o, the user will cache this item. If the cache is full, the user will delete the oldest cached item.*

Under the most recent caching policy, a user only stores the most K recently requested contents. In other words, the caching time CT of a content is a random variable, which equals the sum of K request inter-arrival time T.

Remark 33 *In CTC policy, users cache a content for a constant period of time. This means that users have heterogenous caching capabilities. More specifically, users that request contents frequently have a large cache size, while users that request content occasionally have a small cache size. Such dependency of users' request behaviors on their mobile device capabilities is generally consistent with the reality. On the other hand, in MRC policy, users have a limited cache size. Users that request contents frequently can only store contents for a short period of time, while users that request contents occasionally can store contents for a long time. Both CTC and MRC policy reflect mobile caching in real mobile device systems.*

5.4 How Many Copies of a Content are there in the Network?

When a user requests a content o at time t, how likely the user can find content o in nearby nodes' caches is determined by the number of nodes $M(t)$ that store content o (i.e., *content distribution*). Hence, we study the dynamics of $M(t)$ under different caching policies.

5.4.1 Generic Content Request Pattern and Caching Policy

We study the asymptotic behavior of $M(t)$ in a network under generic content request pattern and content caching policy. Using renewal theory to analyze the content request process of a user, we have the following lemma.

Lemma 7 *When inter-arrival time T of a user's content requests follows general distribution with expectation μ, the expected number of times that a content o is requested by a user within time $[t, t+\Delta t]$ is asymptotically equal to $\frac{p_o \Delta t}{\mu}$ for a fixed Δt as t goes to infinity, where p_o is the request rate of content o shown in Eq. (5.1).*

Proof : Denote by $N_o(t)$ the number of times that content o is requested by a user within time t. Let R_o^i represent the event that the i-th content request is directed to content o, then $N_o(t) = \sum_{i=1}^{N(t)} 1_{R_o^i}$. According to Blackwell's theorem on renewal process, as $t \to \infty$, for any fixed Δt,

$$E(N(t+\Delta t)) - E(N(t)) \to \frac{\Delta t}{\mu}. \tag{5.2}$$

As $P(R_o^i) = p_o$, $E(N_o(t+\Delta t) - N_o(t)) \to \frac{p_o \Delta t}{\mu}$.

In the following, we examine content distribution and delivery for a content o. Hence, we shall omit the content index o for simplicity in the rest of paper unless specifically specified. Based on the above lemma, we derive an upper bound on the number of nodes M that have a content o at time t as $t \to \infty$.

Theorem 17 *Suppose that users have zero-delay tolerance with content requests and GC policy is used, the expectation of M (denoted by $\overline{M}$) is asymptotically upper bounded by $\frac{npE(CT)}{\mu}$ as $t \to \infty$, where CT is a random content caching time.*

Proof : A node has a content o in its cache at time t only if it requested content o within time interval $[t - CT, t]$, i.e., the number of requests directed to o within time interval CT is equal to or lager than 1. Upon Markov's inequality, the probability p_c that a user has a content o in its own cache at time t is asymptotically upper bounded by $E(N(t)) - E(N(t - CT))$ as $t \to \infty$. Based on Lemma 1, we have that conditioning on random variable $CT = \Delta t$, $E(N(t)) - E(N(t - \Delta t))$ is asymptotically equal to $\frac{p\Delta t}{\mu}$. Then, $E[E(N(t)) - E(N(t - CT))|CT]$ is asymptotically equal to $\frac{pE(CT)}{\mu}$ as $t \to \infty$. Subsequently, p_c is asymptotically upper bounded by $\frac{pE(CT)}{\mu}$. As the average number of nodes $\overline{M}$ that have a content o equals np_c, we complete our proof.

Remark 34 *The number of copies of a content in a network is proportional to content request rate and content caching time. When nodes use CTC policy, $\overline{M}$ is asymptotically upper bounded by $\frac{npCT}{\mu}$ as $t \to \infty$; when nodes cache the most K recently requested contents, $\overline{M}$ is asymptotically upper bounded by npK as $t \to \infty$.*

5.4.2 Constant Time Caching (CTC) Policy

In order to facilitate the non-asymptotic analysis of content distribution, we assume that T has exponential distribution function $F_T(x) = 1 - e^{-x/\mu}$ in the rest of paper, which has been used in other studies [121]. Subsequently, the inter-arrival time of a user's two consecutive requests for a content o also follows an exponential distribution with parameter p/μ. Hence, we have the following theorem on the expectation of $M(t)$ (denoted by $\overline{M}(t)$) under CTC policy.

Theorem 18 *Assume that users' content requests have zero-delay tolerance (i.e., $\tau = 0$), under CTC policy, the average number of nodes with content o in their cache at time t satisfies $\overline{M}(t) = n(1 - e^{-\frac{pt^*}{\mu}})$, where $t^* = \min\{t, CT\}$ and p is the request rate of content o.*

Proof : $M(t)$ varies over time as nodes download and delete content o. We now build a differential equation which describes how $M(t)$ varies over time. Consider the time interval $[t, t + \Delta t]$, and let $\Delta M(t) = M(t + \Delta t) - M(t)$. Denote by $\Delta I(t)$ as the number of users without content o at time t that will get content o within time $(t, t + \Delta t]$. Similarly, denote by $\Delta D(t)$ as the number of users with content o at time t that will delete content o from their caches within time $(t, t + \Delta t]$. Hence, $\Delta M(t) = \Delta I(t) - \Delta D(t)$.

We assume that a user's requests of content o follow Poisson process with rate p/μ. Thus, the probability that a user requests content o within time interval Δt is given by $1 - e^{-\frac{p\Delta t}{\mu}}$. Then, the expectation of $\Delta I(t)$ is $[n - M(t)](1 - e^{-\frac{p\Delta t}{\mu}})$, which can be approximated by $\frac{p\Delta t}{\mu}[n - M(t)]$.

Under CTC policy, users cache content o for a constant period of time CT. For $t + \Delta t < CT$, $\Delta D(t) = 0$; for $t \geq CT$, the probability that a node deletes content o from its cache within $(t, t + \Delta t]$ is the joint probability of at least one request directed to content o within $[t - CT, t - CT + \Delta t]$ while no request for content o within $[t - CT + \Delta t, t + \Delta t]$. Then, when $t \geq CT$, the expectation of $\Delta D(t)$ is given by $n(1 - e^{-\frac{p\Delta t}{\mu}})e^{-\frac{pCT}{\mu}}$, which can approximated by $\frac{np\Delta t}{\mu}e^{-\frac{pCT}{\mu}}$.

Finally, dividing $\Delta M(t)$, $\Delta I(t)$, and $\Delta D(t)$ by Δt and letting this time interval go to zero, we get

$$\frac{d\overline{M}(t)}{dt} = \begin{cases} \frac{p}{\mu}[n - \overline{M}(t)] & 0 \leq t < CT \\ \frac{p}{\mu}[n - \overline{M}(t)] - \frac{np}{\mu}e^{-\frac{pCT}{\mu}} & t \geq CT \end{cases} \tag{5.3}$$

Let $\overline{M}(0) = 0$, we solve the differential equation for $\overline{M}(t)$:

$$\overline{M}(t) = \begin{cases} n(1 - e^{-\frac{pt}{\mu}}) & 0 \leq t < CT, \\ n(1 - e^{-\frac{pCT}{\mu}}) & t \geq CT. \end{cases} \tag{5.4}$$

Remark 35 *The probability $p_c(t)$ that a user has a content in its cache at time t is indicated by $\frac{\overline{M}(t)}{n}$, i.e., $1 - e^{-t^*p/\mu}$. When $t \geq CT$, the network is at steady state and the expected number of nodes that have a content is $n(1 - e^{-pCT/\mu})$, which is upper bounded by $npCT/\mu$ as shown in Theorem 17.*

Corollary 3 *Suppose that the delay tolerance $0 < \tau < CT$, under CTC policy, the average number of nodes that have a content o at steady state is lower bounded by $n(1 - e^{-p(CT-\tau)/\mu})$ and upper bounded by $n(1 - e^{-p(CT+\tau)/\mu})$.*

Proof : At time t, nodes that request content o within in time $[t-CT, t-\tau]$ definitely have content o in their cache and only nodes that request content o within time $[t-CT-\tau, t]$ may have content o in their cache. Hence, $\overline{M} \geq n(1-e^{-p(CT-\tau)/\mu})$ and $\overline{M} \leq n(1-e^{-p(CT+\tau)/\mu})$.

It is reasonable to assume $\tau < CT$ because users can cache a content for a few days while delay tolerance can only be a few minutes or hours. Under CTC policy, delay tolerance introduces more dynamics into the content distribution in a network. When $\tau = 0$, the lower and upper bounds of $\overline{M}$ in Corollary 3 conform to $\overline{M}(t)$ at steady state in Theorem 18.

5.4.3 Most Recent Caching (MRC) Policy

We continue to analyze $\overline{M}(t)$ under MRC policy.

Theorem 19 *Assume that $\tau = 0$ and users apply MRC policy, the average number of nodes that have a content o in their cache at time t satisfies*

$$\overline{M}(t) = e^{-\frac{p}{\mu}t} \int^{t} \frac{np}{\mu}[1-(1-p)^K g(x)] e^{\frac{p}{\mu}x} dx, \tag{5.5}$$

where $g(x) = \sum_{i=K}^{\infty} \frac{(\frac{x}{\mu})^i}{i!} e^{-\frac{x}{\mu}}$.

Proof : Similar to proof of Theorem 18, the expectation of $\Delta I(t)$ is approximated by $\frac{p\Delta t}{\mu}[n - M(t)]$. At time $t+\Delta t$, when a user's cache has K or less than K contents, $\Delta D(t) = 0$. A user will delete content o from its cache within time interval $(t, t+\Delta t)$ if and only if the following four conditions are satisfied: 1) up to time $t+\Delta t$, the user has more than K requests, 2) there is at least one request within time $(t, t+\Delta t)$, 3) its K recent requests are not directed to content o, and 4) its $K+1$th recent request is directed to content o. Conditioning on the event that a user generates j requests within $(t, t+\Delta t)$, i.e., $N(\Delta t) = j$, the probability $r_{t,\Delta t}$ that this user will remove o from its cache within time interval $(t, t+\Delta t]$ is given by $P(N(t+\Delta t) \geq K+1)p(1-p)^K$.

$$\begin{aligned} r_{t,\Delta t} &= \sum_{i=K+1}^{\infty} \sum_{j=1}^{i} P(N(t+\Delta t) = i) p(1-p)^K P(N(\Delta t) = j) \\ &\approx p(1-p)^K (1-e^{-\Delta t/\mu}) \sum_{i=K}^{\infty} \frac{(\frac{t}{\mu})^i}{i!} e^{-\frac{t}{\mu}}. \end{aligned} \tag{5.6}$$

Subsequently, the average number of nodes that will delete content o within time $(t, t+\Delta t)$ is $nr_{t,\Delta t}$. Hence, we have a first-order linear differential equation for $\overline{M}(t)$

$$\frac{d\overline{M}(t)}{dt} + \frac{p}{\mu}\overline{M}(t) = \frac{np}{\mu}\left[1-(1-p)^K g(t)\right]. \tag{5.7}$$

where $g(t) = \sum_{i=K}^{\infty} \frac{(\frac{t}{\mu})^i}{i!} e^{-\frac{t}{\mu}}$. Then, solving the differential equation for $\overline{M}(t)$ completes our proof.

Theorem 20 *Let $K \ll \infty$ and t is large enough, $\overline{M} = n[1-(1-p)^K]$ at the equilibrium state of the network.*

Proof : As nodes have finite cache size K, every node's cache is full almost surely (a.s.) when t is large. Hence, the probability that a node has generated more than K requests within time t is almost surely 1, i.e., $P(N(t) \geq K) \rightarrow 1$ a.s.. Then, the differential equation for $M(t)$ can be simplified as:

$$\frac{d\overline{M}(t)}{dt} + \frac{p}{\mu}\overline{M}(t) = \frac{np}{\mu}\left[1-(1-p)^K\right]. \tag{5.8}$$

When the network is in equilibrium state, i.e., $\frac{d\overline{M}(t)}{dt} = 0$, we obtain that $\overline{M}(t) = n[1-(1-p)^K]$.

Remark 36 *When users only cache the most K recently requested contents, the probability that a user has a content o in its cache is $1-(1-p)^K$ at steady state. Content distribution is only determined by a content's request probability p and user's cache size K. According to Bernoulli inequality, $(1-p)^K \geq 1-pK$ for $K \geq 0$ and $p \leq 1$. Then, we have $n[1-(1-p)^K] \leq npK$, which is consistent with Theorem 17.*

Corollary 4 *Suppose that delay tolerance $\tau > 0$ and users adopt MRC policy, the average number of nodes that have a content o at steady state is $n\left[1-(1-p)^K\right]$.*

Proof : Suppose that a user generate i $(0 \leq i < \infty)$ requests within time $[t-\tau, t]$ (denoted as event A_i), among which j $(0 \leq j \leq i)$ requests are satisfied by time t (denoted as event B_j). When $j \leq K$, a user has a content o in its cache if at least one of the j received contents is content o or at least one of $K-j$ requests by time $t-\tau$ is directed to content o. Conditioning on B_j $(j \leq K)$, the probability that a user has content o equals $[1-(1-p)^j]+(1-p)^j[1-(1-p)^{K-j}]$, i.e., $1-(1-p)^K$. When $j > K$, a user has a content o in its cache if among j received contents, at least one of the most K recently received contents is content o. Considering both cases, we have the probability p_c that a user has a content o equals $[1-(1-p)^K]$. Hence, we complete our proof.

Remark 37 *In contrast to content distribution under CTC policy, delay tolerance of content requests does not affect the average number of copies of a content under MRC policy.*

Content request rate and content caching time (cache size) determine content distribution, which in turn affects content sharing performance as we will show in the next section.

5.5 How Likely Can a User Get/Share a Content through D2D Communications?

We evaluate the performance of content sharing through D2D communications in two complimentary aspects under both CTC and MRC policies: i) when a user requests a content o, how likely the user can fetch o from its neighbors (i.e., peer fetching probability); ii) when a user caches content o, how many nodes the user would be able to share the content with.

5.5.1 Peer Fetching Probability

Suppose that a user requests a content at time t with $\tau = 0$, the prospect of fetching the content from peers is the probability that at least one neighbor has the content in cache and stays in connection for successful content transmission.

Theorem 21 *The probability that a user can successfully fetch a content from its neighbor satisfies that*

$$p_p = [1 - p_c][1 - (1 - q\alpha p_c)^{n-1}], \tag{5.9}$$

where $p_c = 1 - e^{-pCT/\mu}$ *under CTC policy while* $p_c = 1 - (1-p)^K$ *under MRC policy,* $\alpha = \pi r^2 \lambda / n$*, and* q *is the success rate of content transmission.*

Proof : We assume that a node requests a content o at time t. Each of the nodes in the network falls in a disc of radius r around the requesting node with probability $\alpha = \pi r^2 \lambda / n$. Hence, the number of neighbors X of the requesting node follows $Binomial(n-1, \alpha)$.

Results in Section 5.4 demonstrate that a node has a content o with probability p_c at steady state. Accordingly, the number of neighbors X_c that have a content o satisfies $X_c = \sum_{i=1}^{X} 1_{\{C_i\}}$. Denote by C_i the probability that a neighbor v_i has content o and I the indicator function $1_{\{C_i\}}$, which has Bernoulli distribution with probability p_c. The probability generating function of I is $G_I(z) = p_c z + 1 - p_c$ for $|z| \leq 1$. Similarly, $G_X(z) = (\alpha z + 1 - \alpha)^{n-1}$ for $|z| \leq 1$. As X_c is a random sum of indicator random variables, its probability generating function

$$G_{X_c}(z) = E[(p_c z + 1 - p_c)^X] = G_X[p_c z + 1 - p_c] = [\alpha p_c z + 1 - \alpha p_c]^{n-1}, \tag{5.10}$$

for $|z| \leq 1$. The inversion of $G_{X_c}(z)$ shows that $X_c \sim Binomial(n-1, \alpha p_c)$.

Then, the probability that a requesting node can successfully fetch a content from a neighbor is

$$p_p = [1 - p_c] \sum_{i=0}^{n-1} [1 - (1-q)^i] P(X_c = i) = [1 - p_c][1 - (1 - q\alpha p_c)^{n-1}]. \tag{5.11}$$

Clearly, the probabilities that a user fetches a content from its own cache, neighbors, and cloud servers are p_c, p_p, and $1 - p_c - p_p$, respectively. We have their numerical results in Fig. 5.9 in order to illustrate how peer and server fetching probabilities vary over content popularity. We set the parameters for the numerical analysis according to our trace analysis. Specifically, $n = 100$, $\alpha = 0.08$, $q = 0.25$, $\lambda_C = 0.003$, $\lambda_I = 0.00001$, $\mu = 10800$ (i.e., 3 hours), $CT = 24$ hours, $K = 10$. Suppose that there are 10000 contents in the network and γ for Zipf law is 1, content request rate $p = 1/o \ln 10000$, where o varies from 5 to 100.

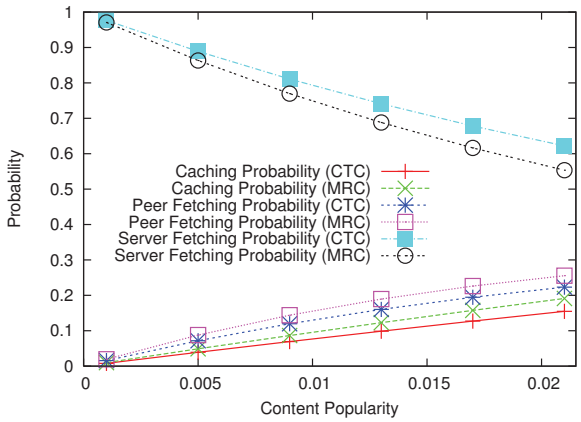

Figure 5.9: Caching probability p_c, peer fetching probability p_p, and server fetching probability under CTC and MRC policies.

Fig. 5.9 shows that content fetching from neighbors through D2D communications can indeed reduce network and server load. In Fig. 5.9, peer fetching probability p_p under MRC policy is higher than p_p under CTC policy because the average caching time of MRC policy (i.e., 30 hours) is set to be longer than CTC policy (i.e., 24 hours). Besides content caching time, content popularity p has a significant impact on p_p. For the top 5 popular contents, server fetching probability can be reduced by 40% even when users have zero delay tolerance. On the other hand, for less popular contents, p_p is very small.

Remark 38 *Although D2D communication is not as reliable as infrastructure-based wireless networks, contents can still be successfully fetched from close-by devices through D2D communication. Fig. 5.9 further illustrates that D2D communication is especially beneficial for popular content fetching.*

If a user allows τ delay tolerance in its content requests, it is more likely to fetch contents from its neighbors as it has opportunity to meet more nodes. We derive the upper and lower bounds on the peer fetching probability over time τ.

Theorem 22 *The probability p_p^τ that a user can successfully fetch a content from nearby users within time τ satisfies*

$$p_p^\tau \geq (1-p_c)\left[1-(1-\beta_\tau \breve{p}_c q)^{n-1}\right], \tag{5.12}$$

$$p_p^\tau \leq (1-p_c)\left[1-(1-\beta_\tau \hat{p}_c^\tau q)^{n-1}\right], \tag{5.13}$$

where $\beta_\tau = 1-(1-\alpha)e^{-\lambda_I \tau}$, $\breve{p}_c = 1-e^{-p(CT-\tau)/\mu}$ and $\hat{p}_c^\tau = 1-e^{-p(CT+2\tau)/\mu}$ under CTC policy, $\breve{p}_c = 1-(1-p)^K$ and $\hat{p}_c^\tau = 1-(1-p)^K e^{-\frac{2\tau p}{\mu}}$ under MRC policy.

Proof : Let a user u request a content o at time t, and the delay tolerance is τ. Nodes that user u encounters within time $[t, t+\tau]$ are candidates to provide content o for u. A user v encounters user u within time τ if their inter-contact time is smaller than τ. Accordingly,

$$\beta_\tau \triangleq \alpha + (1-\alpha)P(T_I \leq \tau) = 1-(1-\alpha)e^{-\lambda_I \tau}. \tag{5.14}$$

Suppose that node u meets node v at least once within time $[t, t+\tau]$, there exists a time point $t_0 \in [t, t+\tau]$ when nodes u and v are in contact. The probability that requesting node u can successfully obtain content o from node v at time t_0 is $p_c q$. Corollary 3 shows that $p_c \geq \breve{p}_c \triangleq 1-e^{-p(CT-\tau)/\mu}$ under CTC policy; Corollary 4 reveals that $p_c = \breve{p}_c \triangleq 1-(1-p)^K$ under MRC policy. Subsequently, the probability that a user u can obtain content o from another user v is lower bounded by $\beta_\tau \breve{p}_c q$. Hence, we prove Eq. (5.12).

Content fetching from peers is possible only if a peer has the content at a time within delay tolerance. Denote by $\hat{p}_c^\tau$ the probability that there exists at least one time point $t_0 \in [t, t+\tau]$ when a node v has content o. Apparently, $\hat{p}_c^\tau$ equals 1 minus the probability that node v does not have the content throughout $[t, t+\tau]$. When users apply CTC policy, if user v does not request content o during time $[t-CT-\tau, t+\tau]$, it would not have content o during time $[t, t+\tau]$. Then,

$$\hat{p}_c^\tau = 1-e^{-p(CT+2\tau)/\mu}. \tag{5.15}$$

When users apply MRC policy, if the most K recent requests before time $t-\tau$ and all i requests within time $[t, t+\tau]$ are not directed to content o, a user will not have content o during time $[t, t+\tau]$. Thus,

$$\hat{p}_c^\tau = 1-\sum_{i=0}^{\infty} \frac{(\frac{2\tau}{\mu})^i}{i!} e^{-\frac{2\tau}{\mu}} (1-p)^{K+i} = 1-(1-p)^K e^{-\frac{2\tau p}{\mu}}.$$

Accordingly, the probability that a user u can obtain content o from another user v is upper bounded by $\beta_\tau \hat{p}_c^\tau q$. Therefore, we prove Eq. (5.13).

Numerical results of Theorem 22 are shown in Fig. 5.10 by fixing content request rate as $p = 0.01$ and letting τ vary from 0 to 3 hours. As delay tolerance increases, the probability

of fetching a content from neighbors increases while that from cloud server decreases, approximately linearly.

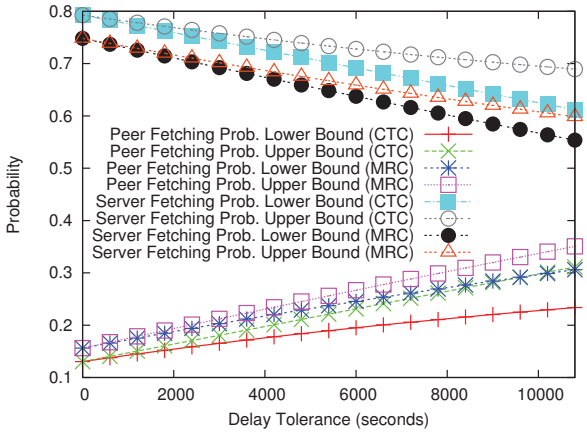

Figure 5.10: Upper and lower bounds on p_p^τ, and server fetching probability under CTC and MRC policies.

Remark 39 *When $\tau = 0$, upper and lower bounds of p_p^τ converges to p_p in Theorem 21. When $\tau > 0$, peer-assisted content fetching probability increases mainly because the number of encountered nodes within τ increases.*

5.5.2 Content Sharing Capacity

When a user u caches a content o for CT period of time, user u can share this content with encountered nodes if they request content o during their contacts with u and their contact duration is long enough for successful content transmission. Because nodes' contact duration T_C is a random variable, we derive the number of requests within a random interval.

Lemma 8 *Assume random time T_C is exponentially distributed with parameter λ_C and request arrivals at a user form a Poisson process with rate $1/\mu$, the number of times that a content o is requested by a user within T_C has Geometric distribution with mean $\frac{p}{\mu\lambda_C}$.*

Proof : The random variable $N(t)$ is defined as the number of requests directed to a content o in a fixed time interval $(0, t)$. We denote by $f(x)$ the probability density function (p.d.f.) of the request inter-arrival time, $p_i(t) = P(N(t) = i)$, and $G(z,t) = \sum_{i=0}^{\infty} z^i p_i(t)$. It is known that these functions' Laplace transform with respect to t are: $p_i^*(s) = [f^*(s)]^i(1 - f^*(s))/s$, and $G^*(z,s) = (1 - f^*(s))/[s(1 - zf^*(s))]$, where $f^*(s) = \int_0^\infty e^{-st} f(t)dt$.

Suppose that random variable T_C is distributed independently of request inter-arrival time and has p.d.f. $f_{T_C}(t)$. Let $N(T_C)$ be the number of requests that are directed to content o in random interval $(0, T_C)$. Then the probability generating function of $N(T_C)$ is given by

$$G(z) = \sum_{i=0}^{\infty} p_i z^i = \int_0^{\infty} G(z,t) f_{T_C}(t) dt. \tag{5.16}$$

When $f_{T_C}(t) = \lambda_C e^{-\lambda_C t}$,

$$G(z) = \lambda_C [G^*(z,s)]_{s=\lambda_C} = \frac{1 - f^*(\lambda_C)}{(1 - z f^*(\lambda_C)}, \tag{5.17}$$

so that $N(T_C) \sim Geometry(1 - f^*(\lambda_C))$ with

$$P(N(T_C) = k) = [f^*(\lambda_C)]^k (1 - f^*(\lambda_C)), k = 0, 1, \ldots.$$

Taking for $f(x)$ the special exponential form $f(t) = p/\mu e^{-pt/\mu}$, $f^*(s) = \frac{p/\mu}{s+p/\mu}$ and $N(T_C)$ has geometric distribution with mean $\frac{p}{\mu\lambda_C}$, i.e., $N(T_C) \sim Geometry(\frac{\mu\lambda_C}{p+\mu\lambda_C})$.

A contact between a providing node u and another node v is effective for sharing a content o if node v will request and successfully fetch content o from u. We denote by $p_s^{T_C}$ the probability that a providing node u shares a content o with a user v within their contact duration T_C, and derive the following theorem on $p_s^{T_C}$.

Theorem 23 *Content sharing probability $p_s^{T_C}$ satisfies that*

$$\frac{qp\mu\lambda_C(1-p_c)}{(p+\mu\lambda_C)^2} \leq p_s^{T_C} \leq \frac{p}{p+\mu\lambda_C}, \tag{5.18}$$

Proof : We assume that node u has a content o in cache during its contact with another user v. User v will fetch content o from node u if v requests content o only once within T_C and at that time v does not have o in its cache. Then, $p_s^{T_C} \geq (1-p_c)P(N^o(T_C) = 1)q$. On the other hand, if user v generates no request directed to o (i.e., $N^o(T_C) = 0$), user v will not need to fetch content o from node u. Accordingly, $p_s^{T_C} \leq 1 - P(N^o(T_C) = 0)$. Upon results on $N^o(T_C)$ in Lemma 8, we obtain the lower and upper bounds for $p_s^{T_C}$.

Remark 40 *Ratio $\frac{p}{\mu\lambda_C}$ represents a node's request rate toward a content o within a contact time, which in turn determines the probability that a providing node shares content o with another node during their contact time.*

Lemma 9 *The expected number of nodes $E(N_E(CT))$ that a providing node encounters within content caching time CT is $(n-1)\beta_{CT}$, where $\beta_{CT} = 1 - (1-\alpha)e^{-\lambda_I CT}$ under CTC policy,*

and $\beta_{CT} = 1 - (1-\alpha)\left(\frac{1}{1+\lambda_I \mu}\right)^K$ *under MRC policy.*

Proof : Clearly, $E(N_E(CT)) = (n-1)\beta_{CT}$, where β_{CT} is the probability that a node meets a providing node within time CT. More specifically, $\beta_{CT} = \alpha + (1-\alpha)P(T_I \leq CT)$. For constant caching time CT, $\beta_{CT} = 1 - (1-\alpha)e^{-\lambda_I CT}$. When CT is a $Erlang(K, 1/\mu)$ random variable,

$$P(T_I \leq CT) = \int_0^\infty (1 - e^{-\lambda_I x}) \frac{(\frac{1}{\mu})^K x^{K-1} e^{-x/\mu}}{(K-1)!} dx = 1 - \left(\frac{1}{1+\lambda_I \mu}\right)^K. \tag{5.19}$$

Hence, we complete our proof.

Based on Theorem 23 and Lemma 9, we can deduce the number of nodes that will fetch a content o from a providing node u during node u's caching time of content o.

Theorem 24 *The expected number of nodes that can fetch a content o from a providing node u with o in cache is upper bounded by* $\left[1 - (\mu\lambda_C/(p+\mu\lambda_C))^{E(N_C(CT))}\right] E(N_E(CT))$*, where* $E(N_C(CT))$ *is shown in Eqs. (5.20) and (5.22) for CTC and MRC policies, respectively.*

Proof : A node v fetches a content o from a provider node u if and only if it successfully fetches a content o during at least one contact with node u within node u's content caching time CT. Hence, we need to derive the number of times two nodes meet during time CT, denote by $E(N_C(CT))$. As two nodes alternates between contact and inter-contact states, their contact process can be modeled by alternating renewal process.

i) When users apply CTC policy, i.e., CT is constant, by renewal theory, $E(N_C(CT))$ is the inverse Laplace transform of $L_{T_C+T_I}(s)/[s(1-L_{T_C+T_I}(s))]$, where $L_{T_C+T_I}(s)$ is the Laplace transform of random variable $T_I + T_C$. As T_C and T_I are exponential random variables with rates λ_C and λ_I, respectively, $L_{T_I+T_C}(s) = \frac{\lambda_C \lambda_I}{(s+\lambda_C)(s+\lambda_I)}$. Accordingly,

$$L(E(N_C(CT)), s) = \frac{\lambda_C \lambda_I}{s^2(s+\lambda_C+\lambda_I)}.$$

Performing inverse Laplace transform, we have

$$E(N_C(CT)) = \frac{\lambda_C \lambda_I CT}{\lambda_C + \lambda_I} + \frac{\lambda_C \lambda_I}{(\lambda_C+\lambda_I)^2}(1 - e^{-(\lambda_C+\lambda_I)CT}). \tag{5.20}$$

ii) When users apply MRC policy, content caching time $CT = \sum_{i=1}^K T_i$, where T_i is the content inter-arrival time. As T_i is an exponential random variable with mean μ, content caching time follows Erlang distribution with mean $K\mu$, i.e.,

$$f_{CT}(t) = \frac{(1/\mu)^K t^{K-1} e^{-t/\mu}}{(K-1)!}. \tag{5.21}$$

The probability generating function of $N_C(t)$ can be represented by $G(z,t) = \sum_{i=0}^{\infty} z^i P(N_C(t) = i)$. Hence, the probability generating function of $N_C(CT)$ is $G(z) = \int_0^{\infty} G(z,t) f_{CT}(t) dt$. Apparently,

$$E[N_C(CT)] = \left[\frac{dG(z)}{dz}\right]_{z=1}, \tag{5.22}$$

where for Erlang distributed CT [122],

$$G(z) = \frac{(1/\mu)^K}{(K-1)!}\left(\frac{\partial}{\partial s}\right)^{K-1}\left[\frac{1 - f_{T_C+T_I}(s)}{s(1 - z f_{T_C+T_I}(s))}\right]_{s=1/\mu}.$$

Subsequently, a user v fetches content o from a providing node u within time CT with probability

$$E\left[1 - \left(1 - p_s^{T_C}\right)^{N_C(CT)}\right] \leq 1 - \left(\frac{\mu\lambda_C}{p + \mu\lambda_C}\right)^{E(N_C(CT))}.$$

Multiplying this probability upper bound with $E(N_E(CT))$ gives an upper bound on the expected number of nodes that can fetch a content from a providing node.

Remark 41 *The content sharing capacity is proportional to the number of nodes that a providing node encounters during its content caching time, and is also determined by content sharing probability during a contact time and the number of contacts within content caching time.*

5.6 Applications

Based on our theoretical analysis, we show that D2D communications can indeed enable content sharing especially for popular contents and delay tolerant contents. Therefore, we identify several promising applications of content sharing through D2D communications.

Special Events Social networking: Special events, such as sports games, attract lots of people in a small area. A large number of users try to access the Internet through the same cellular base station or a few WiFi access points will cause a high pressure on wireless networks. Fortunately, people participating group activities are likely demand related or same contents. For example, during a football game, football related pictures and videos become highly popular. Moreover, high user density makes D2D communication possible, and users are likely to stay in connection for a sufficiently long period of time in order to successfully transmit contents. Users at such special events can leverage the D2D communication opportunities to share contents.

Delay Tolerant Subscription Based Services: Many users subscript services to download newspaper, podcasts, shows to their mobile devices. For instance, RSSRadio provides video/audio podcasts downloading or streaming to mobile devices based on mobile users? subscriptions. Many of current social applications are based on pub/sub abstractions, such as

Instagram, Pinterest, and YouTube channels. Many users subscribe to popular publishers on Instagram, Pinterest, and YouTube. When there is status updates, blogs, pictures, or videos are published, subscribers can download these contents on their mobile devices. Mobile user scan avoid cellular traffic by waiting until wifi access or delivery from neighboring user is available, i.e., contents are delay tolerant. Moreover, there is no limitations on content sharing privacy as these applications fall into the social network category, . Therefore, contents can be shared opportunistically among subscribers.

5.7 Summary and Future Work

In this chapter, we explored the potential for content delivery through D2D communications. We first demonstrated through trace analysis that practical opportunities exist for peer-assisted content delivery because of the temporal locality of user requests and the stochastic dominance of device contact time over content transmission time. Then, we derived the number of copies of a content in the network, and proved that $M(t)$ is an exponential function and a power-law function of content popularity under CTC and MRC policies, respectively. Finally, we evaluated the content fetching and sharing probabilities, and showed that D2D communication can indeed reduce network load especially for popular content delivery. Numerical results exhibit that content request rate, content caching time, and delay tolerance of user request have a dramatic impact on the performance of peer-assisted content sharing. In the future, we will conduct experiments of real mobile wireless networks in order to evaluate the performance of content sharing through D2D communications in real mobile device systems.

Chapter 6

Conclusion and Future Directions

In this report, we have presented our research results on the mobility and traffic correlation in D2D communication networks. Next, we summarize our research results and discuss the possbile future directions.

6.1 Conclusion

Our study has focused on the analysis and understanding of mobility correlation and traffic flow in D2D communication networks. In Chapter 2, we measured mobility correlation in spatial and temporal locality domains such as to detect group structures, evaluate group stability and evolution, assist topology control and data forwarding. As message dissemination is essential for many applications of D2D communication networks, we examined the performance of message dissemination in VANET with intermittent connectivity and high vehicle mobility in Chapter 3. The temporal and spatial limits of message dissemination can provide guidelines on message dissemination algorithm design. In Chapter 4, we studied the D2D communication in the new paradigm of mobile cloud computing. We investigated the computing capacity of mobile cloudlet, evaluating the feasibility of mobile cloudlet to support mobile applications.

In Chapter 2, we characterized mobility correlation in spatial and temporal locality domains such as to detect group structures. We defined a metric, namely Dual-Locality Ratio (DLR), to quantify the mobility correlation between a pair of users, taking account of both similarities in spatial locality (i.e., location and speed) and in temporal locality (i.e., mobility pattern). Simulations and real traces showed that DLR can effectively identify meaningful and stable node groups compared with existing community detection algorithms. Moreover, we showed that DLR has implications on link lifetime and group stability. In order to show the application of DLR, we utilize DLR to evaluate group stability, provide conditions for group evolutions (e.g., node switching and group merging), assist data forwarding and topology control.

In Chapter 3, we studied the performance of information propagation in geocast applications of VANETs. In order to overcome the challenge of dynamic destinations in geocast, we modeled active message mobility that focuses on where the message is broadcasted rather than who is spreading the message. Using message mobility, we analyzed the dissemination distance and hitting time, specifically the farthest distance that the dissemination reaches by time t and the first time that the dissemination reaches location distance d from the source, respectively. Simulation results of four dissemination algorithms are well bounded by our analytic bounds for the dissemination distance and hitting time. Furthermore, two VANET applications are presented to show how our results can provide guidance to design of message dissemination mechanisms in order to satisfy application requirements.

In Chapter 4, we investigated the performance of cloudlet computing for mobile applications. Mobile users can either access remote cloud through cloudlet infrastructures at community sites or offload computational task to encountered mobile devices. In the first scenario, we showed that a mobile user's visiting pattern to the community sites with cloudlets has significant impact on cloudlet computing performance, such as cloudlet access probability, task success rate, and average task execution speed. In the second scenario, we examine the properties and computing capacity of mobile cloudlet in order to determine whether and when mobile cloudlet is able to support mobile applications. Properties of mobile cloudlet, including cloudlet size, node's reachable time and lifetime, are extracted from real traces and analyzed mathematically. Based on the properties of mobile cloudlet, we further derived upper and lower bounds on the computing capacity and long-term computing speed of a mobile cloudlet, which a mobile user can user to decide whether to upload a task to remote cloud or utilize nearby mobile cloudlet. In both cases, mean contact time T_C over mean inter-meeting time T_I represents the impact of user mobility on cloudlet computing performance.

In Chapter 5, we demystified the opportunities of content delivery through D2D communications. Using both trace and theoretical analysis, we found that i) content sharing through D2D communications is feasible because of the temporal locality in user request pattern and the stochastic dominance of device contact time over content transmission time; ii) the number of copies of a content is an exponential or a power law function of content popularity under constant time caching and recency-based caching, respectively; iii) peer-assisted content delivery can greatly reduce the network load for popular contents and delay-tolerance requests, while it achieves little success for less popular contents.

6.2 Future Directions

The work in this dissertation focuses on understanding the mobility and traffic correlation in emerging network systems, such as vehicular ad hoc networks, mobile cloud computing,

and content delivery networks. Our current work looks at the mobility and traffic correlation mainly from an analysis perspective. A joint study from the analysis perspective as well as the experiment and design perspective can offer more comprehensive understanding of the performance limits of a D2D communication network. Although we have showed that some mobile applications can be served by mobile cloudlet, a framework needs to be developed for mobile devices to share computational resources efficiently and reliably. It is important to identify mobile applications that can be served by mobile cloudlet, and design computation partition and offloading strategies that are adaptive to application requirements and contexts of mobile devices.

In addition, as mobile devices are sensing richer contexts and mobile applications are evolving toward context-awareness, information and computation sharing through D2D communications not only needs to consider mobility but also needs to adapt to the behaviors of the applications according to user contexts, such as user activities, preferences, and friendship relationships. Therefore, it is highly desirable to study the correlation between context and information and computation sharing. This will not only provide guidelines to design tailored network services which fit to the current context of the users, but also save network resources, computing resources and battery of mobile devices.

REFERENCES

[1] K. Doppler, M. Rinne, C. Wijting, C.B. Ribeiro, and K. Hugl. Device-to-device communication as an underlay to LTE-advanced networks. *IEEE Communications Magazine*, 47(12):42–49, 2009.

[2] Bezalel Gavish and Suresh Sridhar. The impact of mobility on cellular network configuration. *Wireless Networks*, 7(2):173–185, 2001.

[3] I.F. Akyildiz, Jiang Xie, and S. Mohanty. A survey of mobility management in next-generation all-IP-based wireless systems. *IEEE Wireless Communications*, 11(4):16–28, 2004.

[4] Vincent Lenders, Jörg Wagner, and Martin May. Analyzing the impact of mobility in ad hoc networks. In *Proc. of the 2nd International Workshop on Multi-hop Ad Hoc Networks: From Theory to Reality*, REALMAN '06, 2006.

[5] Bo Han, Pan Hui, V.S.A. Kumar, M.V. Marathe, Jianhua Shao, and A. Srinivasan. Mobile data offloading through opportunistic communications and social participation. *IEEE Trans. on Mobile Computing*, 11(5):821–834, 2012.

[6] Wei Peng, Feng Li, Xukai Zou, and Jie Wu. The virtue of patience: Offloading topical cellular content through opportunistic links. In *IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, 2013.

[7] Han Cai, I. Koprulu, and N.B. Shroff. Exploiting double opportunities for deadline based content propagation in wireless networks. In *Proc. of IEEE INFOCOM*, pages 764–772, 2013.

[8] D. Ciullo, V. Martina, M. Garetto, and E. Leonardi. How much can large-scale video-on-demand benefit from users' cooperation? In *Proc. of IEEE INFOCOM*, 2013.

[9] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. In *Wireless Communication and Mobile Computing Special Issue on Mobile Ad Hoc Networking: Research, Trends and Applications*, 2002.

[10] F. Bai and A. Helmy. *Wireless Ad Hoc and Sensor Networks Chapter 1: A Survey of Mobility Modeling and Analysis in Wireless Adhoc Networks*. Kluwer Academic Publishers, 2004.

[11] Mirco Musolesi and Cecilia Mascolo. Mobility models for systems evaluation a survey. In *Middleware for Network Eccentric and Mobile Applications*, 2009.

[12] F. Bai, N. Sadagopan, and A. Helmy. IMPORTANT: A framework to systematically analyze the impact of mobility on performance of routing protocols for ad hoc networks. In *Proc. of IEEE Information Communications Conference (INFOCOM)*, 2003.

[13] Wenrui Zhao, Mostafa Ammar, and Ellen Zegura. A message ferrying approach for data delivery in sparse mobile ad hoc networks. In *Proc. of ACM MobiHoc*, 2004.

[14] Thrasyvoulos Spyropoulos. Spray and Focus: Efficient mobility-assisted routing for heterogeneous and correlated mobility. In *Proc. of IEEE PerCom Workshop on Intermittently Connected Mobile Ad Hoc Networks*, 2007.

[15] M. Balazinska and P. Castro. Characterizing mobility and network usage in a corporate wireless local-area network. In *Proc. of ACM MobiSys*, 2003.

[16] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. In *Computer Networks*, 2008.

[17] IEEE Standard 802.11p. The IEEE Working Group for WLAN Standards. Available: `http://www.ieee802.org/11/`.

[18] CAR to CAR communication consortium. Available: `http://www.car-to-car.org/`.

[19] Rong Zheng. Information dissemination in power-constrained wireless networks. In *Proc. of IEEE INFOCOM*, 2006.

[20] Yi Xu and Wenye Wang. The speed of information propagation in large wireless networks. In *Proc. of IEEE INFOCOM*, 2008.

[21] P. Jacquet, B. Mans, and G. Rodolakis. Information propagation speed in mobile and delay tolerant networks. In *Proc. of IEEE INFOCOM*, 2009.

[22] Zhenning Kong and Edmund M. Yeh. Connectivity and latency in large-scale wireless networks with unreliable links. In *Proc. of IEEE INFOCOM*, 2008.

[23] Hoang T. Dinh, Chonho Lee, Dusit Niyato, and Ping Wang. A survey of mobile cloud computing: Architecture, applications, and approaches. *Wireless Communications and Mobile Computing*, 2011.

[24] Niroshinie Fernando, Seng W. Loke, and Wenny Rahayu. Mobile cloud computing: A survey. *Future Generation Computer Systems*, 29(1):84–106, 2013.

[25] Emmanouil Koukoumidis, Dimitrios Lymberopoulos, Karin Strauss, Jie Liu, and Doug Burger. Pocket cloudlets. In *The Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2011.

[26] Eugene E. Marinelli. Hyrax: Cloud computing on mobile devices using MapReduce. Master's thesis, Carnegie Mellon University, 2009.

[27] Gonzalo Huerta-Canepa and Dongman Lee. A virtual cloud computing provider for mobile devices. In *ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond. MCS'10*, 2010.

[28] N. Fernando, S.W. Loke, and W. Rahayu. Dynamic mobile cloud computing: Ad hoc and opportunistic job sharing. In *Proc. of IEEE UCC*, 2011.

[29] Cong Shi, Vasileios Lakafosis, Mostafa H. Ammar, and Ellen W. Zegura. Serendipity: Enabling remote computing among intermittently connected mobile devices. In *Proc. of the ACM MobiHoc*, 2012.

[30] A. Aijaz, H. Aghvami, and M. Amani. A survey on mobile data offloading: Technical and business perspectives. *IEEE Wireless Communications*, 20(2):104–112, 2013.

[31] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017. Technical report, Cisco, Feburary 2013.

[32] V. Naumov, R. Baumann, and T. Gross. An evaluation of inter-vehicle ad hoc networks based on realistic vehicular traces. In *Proc. of ACM MobiHoc*, 2006.

[33] P. Hui and J. Crowcroft. Human mobility models and opportunistic communications system design. *Philosophical Transactions of Royal Society A*, (366):2005–2016, 2008.

[34] D. Ciullo, V. Martina, M. Garetto, and E. Leonardi. Impact of correlated mobility on delay-throughput performance in mobile ad-hoc networks. In *Proc. of IEEE INFOCOM*, 2010.

[35] J.-L. Huang and M.-S. Chen. On the effect of group mobility to data replication in ad hoc networks. *IEEE Trans. on Mobile Computing*, 2006.

[36] X. Hong, M. Gerla, G. Pei, and C. Chiang. A group mobility model for ad hoc wireless networks. In *Proc. of ACM MSWiM*, 1999.

[37] A. Nandan, S. Tewari, S. Das, M. Gerla, and L. Kleinrock. AdTorrent: Delivering location cognizant advertisements to car networks. In *Proc. of IEEE/IFIP WONS*, 2006.

[38] Simon Heimlicher and Kave Salamatian. Globs in the primordial soup the emergence of connected crowds in mobile wireless networks. In *Proc. of ACM MobiHoc*, 2010.

[39] K. Wang and B. Li. Efficient and guaranteed service coverage in partitionable mobile ad-hoc networks. In *Proc. of IEEE INFOCOM*, 2002.

[40] B. Gu and X. Hong. Mobility identification and clustering in sparse mobile networks. In *Proc. of IEEE Military Communications Conference (MILCOM)*, 2009.

[41] L. Backstom, D. Huttenlocher, X. Lan, and J. Kleinberg. Group formation in large social networks: Membership, growth, and evolution. In *Knowledge Discovery and Data Mining*, 2006.

[42] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying group evolution. *Nature*, 2007.

[43] P. Hui, E. Yoneki, S.-Y. Chan, , and J. Crowcroft. Distributed community detection in delay tolerant networks. In *Proc. of ACM Sigcomm Workshop MobiArch*, 2007.

[44] A. Ephremides, J.E. Wieselthier, and D.J. Baker. A design concept for reliable mobile radio networks with frequency hopping signaling. 75(1):56–73, 1987.

[45] M. Piorkowski, N. S.-Djukic, and M. Grossglauser. A parsimonious model of mobile partitioned networks with clustering. In *The First International Conference on COMmunication Systems and NETworkS*, 2009.

[46] J. G. Jetcheva, Y-C. Hu, S. PalChaudhuri, A. K. Saha, and D. B. Johnson. Design and evaluation of a metropolitan area multitier wireless adhoc network architecture. In *The 5th IEEE Workshop on Mobile Computing Systems & Applications (WMCSA)*, 2003.

[47] K. H. Wang and B. Li. Group mobility and partition prediction in wireless ad-hoc networks. In *Proc. of IEEE ICC*, 2002.

[48] J. M. Ng and Y. Zhang. Impact of group mobility on ad hoc networks routing protocols. In *The 8th International Conference on Advanced Communication Technology (ICACT)*, 2006.

[49] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, fifth edition, 2002.

[50] D. Lelescu, U. Kozat, R. Jain, and M. Balakrishnan. Model T++: An empirical joint space-time registration model. In *Proc. of ACM MobiHoc*, 2006.

[51] M. C. Gonzalezl, C. A. Hidalgo1, and A.-L. Barabási. Understanding individual human mobility patterns. *Letters to Nature*, 2008.

[52] I. Rhee, M. Shin, S. Hong, and etc. On the Levy-walk nature of human mobility. In *Proc. of IEEE INFOCOM*, 2008.

[53] C. Song, N. Blumm Z. Qu, and A.-L. Barabási. Limits of predictability in human mobility. *Science magazine*, 2010.

[54] T. A. Alhmiedat and Sh.-H. Yang. A survey: Localization and tracking mobile targets through wireless sensors network. Proc. of PGNet, 2007.

[55] L. Song, D. Kotz, and R. Jain. Evaluating location predictors with extensive Wi-Fi mobility data. In *Proc. of IEEE INFOCOM*, 2004.

[56] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *Journal of mathematical models and methods in applied sciences*, 2007.

[57] Ivan Stojmenović. Handbook of wireless networks and mobile computing. chapter Location updates for efficient routing in ad hoc networks, pages 451–471. John Wiley & Sons, Inc., 2002.

[58] András Varga. OMNeT++ discrete event simulation system. http://www.omnetpp.org/.

[59] A. Mei and J. Stefa. SWIM: A simple model to generate small mobile worlds. In *Proc. of IEEE INFOCOM*, 2009.

[60] H. Cai and D. Y. Eun. Crossing over the bounded domain: From exponential to power-law inter-meeting time in MANET. In *Proc. of ACM MobiCom*, 2007.

[61] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 2008.

[62] L. Danon, J. Duch, A. Díaz-Guilera, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005.

[63] S. Fortunato. Community detection in graphs. *Physics Reports*, 2010.

[64] Q. Wang, X. Wang, and X. Lin. Mobility increases the connectivity of k-hop clustered wireless networks. In *Proc. of ACM MobiCom*, 2009.

[65] D. B. West. *Introduction to graph theory*. Prentice-Hall, second edition, 2000.

[66] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. In *National Academy of Sciences*, 2004.

[67] J.B. Kenney. Dedicated Short-Range Communications (DSRC) Standards in the United States. *Proc. of the IEEE*, 99(7):1162–1182, 2011.

[68] CarTALK2000, 2007. http://www.cartalk2000.net.

[69] FleetNet, 2007. http://www.fleetnet.de.

[70] S. Panichpapiboon and W. Pattara-Atikom. A review of information dissemination protocols for vehicular ad hoc networks. *IEEE Communications Surveys Tutorials*, 14(3):784–798, 2012.

[71] F. De Pellegrini, D. Miorandi, I. Carreras, and I. Chlamtac. A graph-based model for disconnected ad hoc networks. In *Proc. of IEEE INFOCOM*, 2007.

[72] P. Jacquet, S. Malik, B. Mans, and A. Silva. On the throughput-delay trade-off in georouting networks. In *Proc. of IEEE INFOCOM*, 2012.

[73] H. Wu, R. Fujimoto, and G. Riley. Analytical models for information propagation in vehicle-to-vehicle networks. In *Proc. of IEEE Vehicular Technology Conference (VTC-Fall)*, 2004.

[74] ManWo Ng and S. T. Waller. A static network level model for the information propagation in vehicular ad hoc networks. *Transportation Research Part C: Emerging Technologies*, pages 393–407, June 2010.

[75] E. Baccelli, P. Jacquet, B. Mans, and G. Rodolakis. Information propagation speed in bidirectional vehicular delay tolerant networks. In *Proc. of IEEE INFOCOM*, 2011.

[76] E. Baccelli, P. Jacquet, B. Mans, and G. Rodolakis. Highway vehicular delay tolerant networks: Information propagation speed properties. *IEEE Trans. on Information Theory*, 58(3):1743–1756, 2012.

[77] R. Fracchia and M. Meo. Analysis and design of warning delivery service in intervehicular networks. *IEEE Trans. on Mobile Computing*, (7):823–845, July 2008.

[78] Yujin Li, Wenye Wang, and A. Duel-Hallen. The latency of gaining α-reliability for message dissemination in vehicle-to-vehicle networks. In *Proc. of IEEE GLOBECOM*, 2012.

[79] Giovanni Resta, Paolo Santi, and Janos Simon. Analysis of multi-hop emergency message propagation in vehicular ad hoc networks. In *Proc. of ACM MobiHoc*, 2007.

[80] Yujin Li and Wenye Wang. Geo-dissemination in vehicular ad hoc networks. In *Proc. of IEEE International Conference on Communications (ICC)*, pages 302–306, 2012.

[81] Sze-Yao Ni, Yu-Chee Tseng, Yuh-Shyan Chen, and Jang-Ping Sheu. The broadcast storm problem in mobile ad hoc network. In *Proc. of ACM MobiCom*, 1999.

[82] Chi-Kin Chau and Prithwish Basu. Analysis of latency of stateless opportunistic forwarding in intermittently connected networks. In *IEEE/ACM Trans. on Networking*, 2011.

[83] Min-Te Sun, Wu-Chi Feng, Ten-Hwang Lai, K. Yamada, H. Okada, and K. Fujimura. GPS-based message broadcast for adaptive inter-vehicle communications. In *Proc. of IEEE Vehicular Technology Conference (VTC-Fall)*, 2000.

[84] M. Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.

[85] F. Xue and P.R. Kumar. The number of neighbors needed for connectivity of wireless networks. *Wireless Networks*, 2004.

[86] Lei Sun and Wenye Wang. On the dissemination latency of cognitive radio networks under general node mobility. In *Proc. of IEEE ICC*, 2011.

[87] M.J. Neely and E. Modiano. Capacity and delay tradeoffs for ad-hoc mobile networks. *IEEE Trans. on Information Theory*, 2005.

[88] Han Cai and Do Young Eun. Toward stochastic anatomy of inter-meeting time distribution under general mobility models. In *Proc. of ACM MobiHoc*, 2008.

[89] T. Aven. Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *Journal of Applied Probability*, 22(3), 1985.

[90] P. Hui, A. Chaintreau, R. Gass, J. Scott, J. Crowcroft, and C. Diot. Pocket switched networking: Challenges, feasibility, and implementation issues. In *Proc. of the Second IFIP on Autonomic Communications*, 2005.

[91] S. Dolev, E. Schiller, and J. L. Welch. Random walk for self-stabilizing group communication in ad hoc networks. In *IEEE Trans. on Mobile Computing*, 2006.

[92] C. Gkantsidis, M. Mihail, and A. Saberi. Random walks in peer-to-peer networks: Algorithms and evaluation. In *Perform. Eval.*, 2006.

[93] I. Mabrouki, X. Lagrange, and G. Froc. Random walk based routing protocol for wireless sensor networks. In *Proc. of ValueTools*, 2007.

[94] Stuart Kurkowski, Tracy Camp, and Michael Colagrosso. MANET simulation studies: The incredibles. *ACM SIGMOBILE Mobile Computing and Communications Review*, 9:50–61, 2005.

[95] D. Jiang and L. Delgrossi. IEEE 802.11p: Towards an international standard for wireless access in vehicular environments. In *Proc. of IEEE Vehicular Technology Conference (VTC-Spring)*, 2008.

[96] J. Ni and S. Chandler. Connectivity properties of a random radio network. In *Proc. of the IEEE Communications*, 1994.

[97] Vehicle safety communication project: Task 3 final report, identify vehicle safety applications enabled by DSRC. Technical report, National Highway Traffic Safety Administration, Department of Transportation, US, 2005.

[98] Qi Zhang, Lu Cheng, and Raouf Boutaba. Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1:7–18, May 2010.

[99] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies. The case for VM-based cloudlets in mobile computing. *IEEE Pervasive Computing*, 8(4):14–23, 2009.

[100] Byung-Gon Chun, Sunghwan Ihm, Petros Maniatis, Mayur Naik, and Ashwin Patti. CloneCloud: Elastic execution between mobile device and cloud. In *Proc. of the 6th conference on Computer systems*, pages 301–314, 2011.

[101] Abdul Nasir Khan, M.L. Mat Kiah, Samee U. Khan, and Sajjad A. Madani. Towards secure mobile cloud computing: A survey. *Future Generation Computer Systems*, 2012.

[102] Tim Verbelen, Pieter Simoens, Filip De Turck, and Bart Dhoedt. Cloudlets: Bringing the cloud to the mobile user. In *Proc. of the 3rd ACM workshop on Mobile cloud computing and services*, pages 29–36, 2012.

[103] D. Fesehaye, Yunlong Gao, K. Nahrstedt, and Guijun Wang. Impact of cloudlets on interactive mobile cloud applications. In *The IEEE 16th International Enterprise Distributed Object Computing Conference (EDOC)*, 2012.

[104] Ming Zhao, Yujin Li, and Wenye Wang. Modeling and analytical study of link properties in multihop wireless networks. *IEEE Trans. on Communications*, 60(2):445–455, 2012.

[105] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovic. Power law and exponential decay of inter contact times between mobile devices. In *Proc. of ACM MobiCom*, 2007.

[106] D.R. Cox. *Renewal Theory*. Methuen & Co, 1962.

[107] Richard E. Barlow and Larry C. Hunter. Reliability of analysis of a one-unit system. *Operations Research*, 9, 1961.

[108] I. Gruber and Hui Li. Link expiration times in mobile ad hoc networks. In *Proceedings of 27th Annual IEEE Conference on Local Computer Networks (LCN)*, pages 743–750, 2002.

[109] A. Chaintreau, P. Hui, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Trans. on Mobile Computing*, 6(6):606–620, 2007.

[110] M.H. Rossiter. The sojourn time distribution of an alternating renewal process. *Australian Journal of Statistics*, 31(1):143–152, 1989.

[111] Eduardo Cuervo, Aruna Balasubramanian, Dae ki Cho, Alec Wolman, Stefan Saroiu, Ranveer Chandra, and Paramvir Bahl. MAUI: Making smartphones last longer with code offload. In *Proc. of the 8th international conference on Mobile systems, applications, and services (MobiSys)*, 2010.

[112] IEEE 802.15 WPAN Task Group 8 (TG8) Peer Aware Communications. http://www.ieee802.org/15/pub/TG8.html, 2014.

[113] Michael Zink, Kyoungwon Suh, Yu Gu, and Jim Kurose. Characteristics of YouTube network traffic at a campus network — measurements, models, and implications. *Computer Networks*, 53(4):501–514, 2009.

[114] M. Zubair Shafiq, Lusheng Ji, Alex X. Liu, and Jia Wang. Characterizing and modeling Internet traffic dynamics of cellular devices. In *Proc. of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, 2011.

[115] Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. Youtube traffic characterization: A view from the edge. In *Proc. of the ACM SIGCOMM conference on Internet measurement*, 2007.

[116] Shu Liu and Aaron D. Striegel. Exploring the potential in practice for opportunistic networks amongst smart mobile devices. In *Proc. of ACM MobiCom*, 2013.

[117] Yujin Li and Wenye Wang. Can mobile cloudlets support mobile applications? In *Proc. of IEEE INFOCOM*, 2014.

[118] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the world's largest user generated content video system. In *ACM Internet Measurement Conference*, 2007.

[119] K.-Y. Wong. Web cache replacement policies: A pragmatic approach. *IEEE Network*, 20(1):28–34, Jan 2006.

[120] Waleed Ali, Siti Mariyam Shamsuddin, and Abdul Samad Ismail. A survey of web caching and prefetching. *Int. J. Advance. Soft Comput. Appl.*, 3(1):18–44, March 2011.

[121] V. Pacifici and G. Dan. Content-peering dynamics of autonomous caches in a content-centric network. In *Proc. of IEEE INFOCOM*, 2013.

[122] D. R. Cox. On the number of renewals in a random interval. *Biometrika*, 47(3/4):449–452, 1960.