Large scale data collection and storage using smart vehicles : An information-centric approach

Junaid Khan

Université Paris-EST

Ecole Doctorale MSTIC
Mathématiques et Sciences et Technologies de l'Information et de la Communication

Ph.D Thesis

Mention: Informatique

defended on 04 November 2016 by:

Junaid Ahmed Khan

Large scale data collection and storage using smart vehicles : An information-centric approach

---

Advisor:

Yacine Ghamri-Doudane

Jury:

M. André-Luc Beylot	IRIT, Université de Toulouse	Rapporteur
M. Jean-Loup Guillaume	Université de La Rochelle	Examinateur
M. Marcelo Dias de Amorim	Université Pierre et Marie Curie	Rapporteur
M. Rami Langar	Université Paris-Est MLV	Examinateur
M. Sidi-Mohammed Senouci	Université de Bourgogne	Examinateur
M. Yacine Ghamri-Doudane	Université de La Rochelle	Directeur de thése

Laboratoire d'Informatique Gaspard-Monge (LIGM)
5, bd Descartes
77454 Marne La Vallée Cedex 02

Université Paris-Est
école Doctorale MSTIC
Mathématiques, Sciences et Technologies de l'information et de la communication
5, bd Descartes
77454 Marne La Vallée Cedex 02

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Prof. Yacine Ghamri-Doudane for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. André-Luc Beylot, Prof. Jean-Loup Guillaume, Prof. Marcelo Dias de Amorim, Prof. Rami Langar and Prof. Sidi-Mohammed Senouci, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also goes to Prof. Jean-Marc Ogier, President of the University of La Rochelle and Prof. Cedric Westphal of Basking School of Engineering, UC Santa Cruz, both of whom provided me an opportunity to join their lab, and who gave access to the laboratory and research facilities. Without their precious support it would not be possible to conduct this research.

I thank my fellow lab-mates for the stimulating discussions and for all the fun we have had in the last three years. Also I thank my friends for their moral support during stressful and difficult moments.

Last but not the least, I would like to thank my family, specially my parents, Mr. Pervez Ahmed Khan and Ms Shaheen Pervez, for supporting me spiritually throughout writing this thesis and my life in general.

Résumé

De nos jours, le nombre de dispositifs ne cesse d'augmenter ce qui induit une forte demande des applications en données multimédia. Cependant gérer des données massives générées et consommées par les utilisateurs mobiles dans une zone urbaine reste une problématique de taille pour les réseaux cellulaires existants qui sont à la fois onéreux et limités en terme de bande passante mais aussi due à la nature des échanges de données centrées-connexion. D'autre part, l'avancée technologique en matière de véhicules autonomes permet de constituer une infrastructure prometteuse capable de prendre en charge le traitement, la sauvegarde, et la communication de certaines de ces données. En effet, Il est maintenant possible de recruter des véhicules intelligents pour des fins de collecte, de stockage, et de partage de données hétérogènes en provenance d'un réseau routier dans le but de répondre aux demandes des citoyens via des applications. Dites de villes intelligentes, nous tirons profit de l'évolution récente dans le domaine de l'Information Centric Networking (ICN) afin d'introduire deux nouvelles approches de collecte et de stockage efficaces de contenus par les véhicules, au plus proche de l'utilisateur mobile en zone urbaine. Ainsi nous remédions aux problèmes liés à la bande passante et au coût mentionnées plus haut. Ces deux approaches sont respectivement nommées VISIT et SAVING. VISIT est une plate-forme qui définit de nouvelles mesures de centralité basées sur l'intérêt social des citoyens et ayant pour but d'identifier et de sélectionner l'ensemble approprié des meilleurs véhicules candidats pour la collecte des données urbaines. SAVING est un système de stockage de données sociales, qui présente une solution de mise en cache des données d'une façon collaborative entre un ensemble de véhicules parmi d'autres désignés et recrutés selon une stratégie formalisée en utilisant la théorie des jeux basée sur les réseaux complexes. Nous avons testé ces deux approaches VISIT et SAVING, sur des données simulées dans un environment contenant 2986 véhicules svivant des traces de mobilité réalistes en zone urbaine Les résultats obtenues demontres clairement que les deux approaches permettent non seulement une collecte et un stockage efficaces mais sont aussi scalables.

Publications

Journals:

- Junaid Ahmed Khan, Yacine Ghamri-Doudane, Dmitri Botvich, **"Autonomous Identification and Optimal Selection of Popular Smart Vehicles for Urban Sensing - An Information-centric Approach"** *IEEE Transactions on Vehicular Technology, 2016.*
- Junaid Ahmed Khan and Yacine Ghamri-Doudane, **"SAVING: Socially Aware Vehicular Information-centric Networking"** , *IEEE Communications Magazine August 2016.*

Conferences:

- Junaid Ahmed Khan and Yacine Ghamri-Doudane, **"STRIVE: Socially-aware Three-tier Routing in Information-centric Vehicular Environment "**, *IEEE Globecom 2016*
- Junaid Ahmed Khan, Yacine Ghamri-Doudane, Dmitri Botvich, **"InfoRank: Information-Centric Autonomous Identification of Popular Smart Vehicles"**, *IEEE Vehicular Technology Conference (IEEE VTC Fall) 2015.*
- Junaid Ahmed Khan, and Yacine Ghamri-Doudane, **"CarRank: An Information-Centric Identification of Important Smart Vehicles for Urban Sensing"**, *14th IEEE International Symposium on Network Computing and Applications, IEEE NCA 2015.*
- Junaid Ahmed Khan, Yacine Ghamri-Doudane, Dmitri Botvich, **"GRank - An Information-Centric Autonomous and Distributed Ranking of Popular Smart Vehicles"**, *IEEE Globecom 2015.*
- Junaid Ahmed Khan, Yacine Ghamri-Doudane, Ali El Masri, **"Vers une approche centrée information (ICN) pour identifier les véhicules importants dans les VANETs"**, *IEEE CFIP NOTERE 2015.*

Contents

Acknowledgments **3**

Abstract **5**

Publications **7**

1 Introduction **15**
1.1 Motivation . . . 15
1.2 Problem Statement . . . 16
1.3 Contribution . . . 16
1.3.1 Vehicular Information-centric Socially Inspired Telematics (VISIT) 17
1.3.2 Socially Aware Vehicular Information-centric NetworkinG (SAVING) . . . 17
1.4 Thesis Organization . . . 18

2 State of the Art **19**
2.1 Background: Information Centric Networking . . . 19
2.2 Data Collection using Vehicles . . . 20
2.2.1 ICN meets Urban Data Collection . . . 20
2.2.2 Social Network meets Vehicular Network . . . 21
2.2.3 Discussion . . . 22
2.3 Data Storage in a Vehicular Network . . . 22
2.3.1 Recruitment of Vehicles as Content Caches . . . 22
2.3.2 Content Cache Management . . . 23
2.3.3 Social Networks meet Caching . . . 24
2.3.4 Discussion . . . 24
2.4 Conclusion . . . 25

3 VISIT for Data Collection: Novel Centrality Metrics to Identify Eligible Candidates **27**
3.1 Introduction . . . 27
3.2 Context and Motivation . . . 28
3.3 InfoRank: An Information Importance Based Centrality Scheme . . . 29
3.3.1 Network Model . . . 29
3.3.2 User Interests Satisfaction . . . 31
3.3.3 Information Validity Scope . . . 32

3.3.4 InfoRank Computation . . . 33
3.4 CarRank: A Vehicle Centrality Algorithm . . . 35
3.4.1 Information Importance . . . 36
3.4.2 Spatio-temporal Availability . . . 36
3.4.3 Neighborhood Importance . . . 37
3.4.4 CarRank Computation . . . 37
3.5 Performance Evaluation . . . 39
3.5.1 Simulation Scenario . . . 39
3.5.2 Results: Individual Vehicle Ranking . . . 42
3.5.2.1 Cumulative Satisfied Interests . . . 43
3.5.2.2 Temporal behavior analysis of the top nodes . . . 43
3.5.2.3 Throughput . . . 49
3.5.2.4 ICN Evaluation - In-Network Caching . . . 51
3.5.3 Discussion . . . 53
3.6 Conclusions . . . 53

4 VISIT for Data Collection: An Optimal Vehicles Selection Scheme 55
4.1 Introduction . . . 55
4.2 Context and Motivation . . . 56
4.3 Recruitment of Optimal Vehicles for Efficient Road Sensing (ROVERS) 56
4.3.1 Problem Formulation . . . 57
4.3.1.1 Spatio-temporal Coverage . . . 57
4.3.1.2 Vehicle Centrality . . . 58
4.3.1.3 Dedicated Budget . . . 58
4.3.2 Algorithm: Optimized Set of Vehicles selection . . . 60
4.4 Performance Evaluation . . . 60
4.4.1 Results: Best set of vehicles selection . . . 61
4.4.1.1 Cumulative Satisfied Interests . . . 61
4.4.1.2 Throughput . . . 62
4.5 Conclusions . . . 64

5 SAVING as Data Storage: A Social Choice Game for Urban Cache Recruitment 65
5.1 Introduction . . . 65
5.2 Context and Motivation . . . 66
5.3 Autonomous Information Hub Identification - GRank . . . 67
5.3.1 Information Global Centrality . . . 68
5.3.2 GRank Computation . . . 72
5.3.3 Summary . . . 72
5.4 Information Hubs Recruitment . . . 73
5.4.1 Game Formulation . . . 73
5.4.2 Vehicle Utility as a Social Norm . . . 75
5.4.3 Social Welfare Optimization . . . 76
5.4.4 Algorithm: Vehicle Selection as Content Caches . . . 77
5.5 Performance Evaluation . . . 78
5.5.1 Simulation Scenario . . . 78

5.5.2 Results: Individual Vehicles Ranking . . . 79
5.5.2.1 Cumulative Satisfied Interests . . . 79
5.5.2.2 Temporal Network Behavior Analysis . . . 81
5.5.2.3 Aggregated Per Node Throughput . . . 82
5.5.2.4 In-network Cache Hit-rate . . . 83
5.5.3 Results: Selected Set of Vehicles . . . 84
5.5.3.1 Success Rate . . . 84
5.5.3.2 Aggregated Throughputs . . . 84
5.5.3.3 In-network Cache Hit-Rate . . . 87
5.6 Conclusions . . . 88

6 SAVING as Data Storage: A Collaborative Caching Game at Mobile Fogs 89
6.1 Introduction . . . 89
6.2 Context and Motivation . . . 90
6.3 Who should cache What? . . . 91
6.3.1 Network Model . . . 92
6.3.2 Spatio-temporal Content Profile . . . 92
6.3.3 Node Eligibility . . . 96
6.3.3.1 Local Centrality . . . 96
6.3.3.2 Global Centrality . . . 97
6.4 Distributed Fogs Formation as Coalition Game . . . 98
6.4.1 Solution: The Core . . . 100
6.4.2 Algorithm: Optimal Selection among Coalitions . . . 101
6.5 Performance Evaluation . . . 101
6.5.1 Simulation Scenario . . . 101
6.5.2 Simulation Results . . . 104
6.5.2.1 Cache Hits . . . 104
6.5.2.2 Offload Benefit . . . 104
6.5.2.3 Spatio-temporal coalitions . . . 105
6.5.3 Summary of Findings . . . 106
6.6 Conclusions . . . 106

7 Conclusions and Future Work 107
7.1 Conclusions . . . 107
7.2 Future Work . . . 108

List of Figures 109

List of Tables 111

Bibliography 113

Abstract

The growth in the number of mobile devices today result in an increasing demand for large amount of rich multimedia content to support numerous applications. It is however challenging for the current cellular networks to deal with such increasing demand, both in terms of cost and bandwidth that are necessary to handle the "massive" content generated and consumed by mobile users in an urban environment. This is partly due to the connection-centric nature of current mobile systems. The technological advancement in modern vehicles allow us to harness their computing, caching and communication capabilities to supplement infrastructure network. It is now possible to recruit smart vehicles to collect, store and share heterogeneous data on urban streets in order to provide citizens with different services. Therefore, we leverage the recent shift towards Information Centric Networking (ICN) to introduce two novel schemes, VISIT and SAVING. These schemes aim the efficient collection and storage of content at vehicles, closer to the urban mobile user, to reduce bandwidth demand and cost. VISIT is a platform which defines novel centrality metrics, based on the social interest of urban users, to identify and select the appropriate set of best candidate vehicles to perform urban data collection. SAVING is a social-aware data storage system which exploits complex networks to present game-theoretic solutions for finding and recruiting the vehicles, which are adequate to perform collaborative content caching in an urban environment. VISIT and SAVING are simulated for about 2986 vehicles with realistic urban mobility traces. Comparison results with other schemes in the literature suggest that both are not only efficient but also scalable data collection and storage systems.

Chapter 1

Introduction

1.1 Motivation

Mobile devices, such as smart phones, tablets and sensor-equipped vehicles today entail a constant generation and consumption of massive internet traffic to provide diverse LBS (Location Based Services) applications [37]. For example, vehicles are equipped with a lots of electronic components including sensors, cameras and communication devices to facilitate towards our utmost travel comfort and safety. Such "Smart Vehicles" can be considered as an instance of the Internet of Things (IoT) aimed to harvest and share different sensory and multimedia data on urban streets supporting various Intelligent Transportation System (ITS) applications such as efficient traffic management, infotainment and urban environment sensing. However, cellular network operators are currently struggling to cope with the monetary and bandwidth requirements of such large scale ubiquitous data collection and storage.

One promising solution is to utilize smart vehicles with their relatively high processing, storing and communicating capabilities to offload network data [45], [19]. Such vehicles can supplement the infrastructure network by forming an Internet of Vehicles (IoVs) to facilitate citizens lifestyle with different location-aware services. The issue is, first, the current networks are still unable to bear such high bandwidth consumption, imagine each vehicle simultaneously generates a huge amount of heterogeneous data in the tera bit scale per day. Second, the cost associated with the content demand including multimedia data makes it unsupportable given the numerous unlimited tariff plans offered by network operators.

Information-Centric Networking (ICN) [2] [3] is recently proposed as an alternative networking architecture for replacing the IP-based Internet. An ICN user broadcasts an interest for a content by its name, any corresponding host in the network replies back with the desired content. ICN aims to decouple the service from the host, thus removing content association to any physical location. It also allows intermediate nodes to cache content while forwarding and responding to user interests. Thus, we strongly believe that the named-data networking concept introduced by the information-centric networking paradigm can be helpful to cater with the mobility and intermittent connectivity challenge in vehicular network.

1.2 Problem Statement

Assuming a large number of vehicles on urban roads, it is challenging to find the right set of vehicles available at the right time and place for efficient data collection, storage and sharing through low-cost inter-vehicle communications. Therefore, we target the problem towards efficient large scale data collection and storage using vehicles in an urban environment with the following questions:

- What are the metrics that classify a particular information as well as vehicle "eligibility" for data collection based on its commute and social importance in an urban environment?
- Which set of vehicles can be selected (out of the best candidate vehicles) as appropriate candidates for urban data collection to optimally achieve city-wide spatio-temporal coverage with the least redundancy under a given budget?
- How to classify a particular vehicle as an eligible candidate for content storage at different urban locations? Once such eligible candidate vehicles are identified, how to optimally recruit the best set of vehicles for content availability requirements under a given budget? Last but not the least, how to ensure fairness among recruited vehicles compensating for their participating cost and resources while dealing with vehicle individual-rationality?
- After the recruitment of the vehicles as caches, there is a need to carefully address questions regarding "Who should cache what, when and where?" requiring an understanding of the content as well as the vehicle spatio-temporal profile based on the social interests of large number of urban users.

1.3 Contribution

To address the above questions, we propose the concept of socially aware information-centric vehicular networking as a publish-subscribe framework exploiting content-centric networking architecture. To do this, we present novel distributed data collection and caching schemes to complement infrastructure network for urban mobile users in order to maximize content availability. Figure 1.1 represents an example of such urban data collection and caching where the "User Vehicle" is interested for information regarding a location in a particular zone, assuming the city is divided into different urban zones. It forwards the interest to a near by information facilitator vehicle which subsequently facilitate the interest by collecting and caching the desired content. The "Source vehicle" provides the content to nearby information facilitators which are responsible for the content availability in the network.

In order to achieve this by answering precisely the above questions, the major contributions of this thesis are regrouped under two groups and are summarized below: the Vehicular Information-centric Socially Inspired Telematics (VISIT) for data collection, and the Socially-Aware Vehicular Information-centric NetworkinG (SAVING) for data storage.

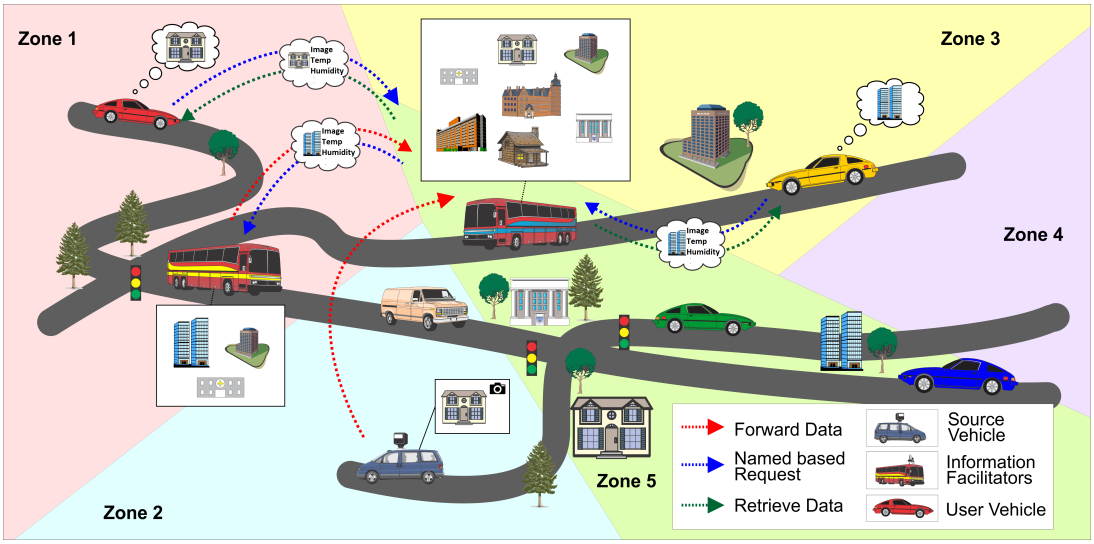


Figure 1.1: System Overview

1.3.1 Vehicular Information-centric Socially Inspired Telematics (VISIT)

VISIT deals with the large scale urban data collection by identifying and selecting eligible vehicles thanks to the following set of complementary contributions:

- First, we propose a novel metric "InfoRank" [20] enabling a vehicle to autonomously (1) rank important location-aware information associated to it based on the satisfaction of users social interests and the information validity, and then (2) rank itself based on the obtained information importance and the vehicle's mobility pattern in the city.

- Then, an innovative vehicle centrality metric, "CarRank", is proposed, where each vehicle can find its importance in the network. This importance is linked to the InfoRank score, the vehicle spatio-temporal availability and the neighborhood topology.

- Finally, an optimal recruitment algorithm to select the minimum set of important vehicles identified by CarRank for urban data collection. This minimum set is selected while maintaining acceptable city-wide coverage with less redundant data at the minimum cost.

1.3.2 Socially Aware Vehicular Information-centric NetworkinG (SAVING)

SAVING is a data storage framework to find and recruit vehicles for collaborative urban caching. This is realized thanks to the following interlinked contributions:

- A content spatio-temporal availability-popularity relation is proposed to decide a content importance in order to be cached at a vehicle.

- A vehicle eligibility metric "GRank" is presented for social aware content caching allowing a vehicle to classify its caching capability proportionally to its connectivity, cost and reachability in the network.
- An incentive driven social welfare game is formulated to fairly select among the best ranked vehicles the eligible candidates to cache content for different urban neighborhoods catering individual rationality.
- An optimization problem is modeled following by a vehicle selection algorithm to ensure the selected vehicles satisfy urban content availability requirements for a given budget.
- In order to even further optimize the caching, a coalition game is proposed to form mobile fogs [50] for resource pooling at spatio-temporally co-located vehicles where service provider selects coalitions as urban caches to maximize content availability.

1.4 Thesis Organization

The remaining of the document is organized as follows. The following chapter highlights the background along a review on the state of the art data collection and storage schemes in the literature. The proposed data collection approach VISIT is presented by first defining novel metrics to classify vehicle eligibility in Chapter 3 followed by an optimized selection of the best set of vehicles to collect data from urban streets in Chapter 4. Chapters 5 and 6 deals with the data storage scheme SAVING where the identification and recruitment of individual urban information hubs is addressed in Chapter 5 while the collaborative distributed caching among vehicles is described in Chapter 6. The Chapter 7.1 concludes our work with a discussion on future insight in Chapter 7.2.

Chapter 2

State of the Art

2.1 Background: Information Centric Networking

We observe a recent shift towards Information-Centric Networking (ICN) [2] [3] [11] as the underlying routing protocol for vehicular networking. ICN is a content-centric networking architecture proposed to replace the current IP based Internet. In ICN, a user broadcasts an interest for a content by its name, any corresponding host in the network replies back with the desired content. ICN aims to decouple the service from the host, thus removing content association to any physical location. It also offers In-Network caching at intermediate nodes while forwarding and responding to subsequent user interests.

Thus, the named-data networking concept introduced by the information-centric networking paradigm is capable to co-exist with the mobility and intermittent connectivity challenge in mobile networks. ICN inherent in-network caching and provider-consumer decoupling maximize content availability by allowing users to retrieve the content cached at “any” nearby source independent of the underlying network connectivity.

We adapt the Content Centric Networking (CCN) instance of the general ICN architecture to the requirement of vehicular environment while advocating the importance of the “information-centric" or “content-centric" networking philosophy. We privilege the user-relevant importance of an information rather than its localization. The ICN paradigm addresses the issue by decoupling the content provider-consumer and support in-network caching at intermediate nodes to improve content availability with minimum delays. Assuming an urban environment where vehicles constantly receive and satisfy interests for different content from the neighboring vehicles using multi-hop interest forwarding in a vehicular network, each ICN-enabled vehicle can maintain three routing parameters:

- Forwarding Information Base - FIB: It resembles a routing table which maps content name components to interfaces. Each vehicle FIB is populated by the routes discovered using name-based interest/content forwarding protocol.
- Pending Interest Table - PIT: It keeps track of all the incoming interests that the vehicle has forwarded but not satisfied yet. Each PIT entry records the content name carried in the interest, together with its incoming and outgoing interface(s).

- Content Store - CS: It is a temporary cache to store content each intermediate vehicle has received while forwarding content. Since a named-data packet is meaningful independent of where it comes from or where it is forwarded to, it can be cached to satisfy future interests.

We consider a publish-subscribe ICN model allowing a mobile node such as a vehicle in our case to subscribe for the following three roles:

1. **Information Provider** An information provider vehicle acts as the content source to publish content. For example, it can subscribe itself to publish sensory information collected from urban streets using the vehicle embedded cameras and sensors.

2. **Information Facilitator** Vehicle responsible to collect, cache in CS and relay using FIB and PIT the content generated by information provider vehicles as well. It also forwards the user interest using PIT for content to "facilitate" efficient content caching and distribution.

3. **Information Consumer** The vehicle subscribed to request different content from the information facilitators/providers with in the vehicular network are considered as information consumers to "pull" content from CS of different nearby vehicles in an information-centric vehicular network.

Having this in mind, the proposed data collection and storage systems consider the importance of the location-aware information associated to vehicles as an information-centric approach instead of relying on physical hosts in the ephemeral vehicular network topology.

2.2 Data Collection using Vehicles

We consider a smart city application of sensory data collection from urban streets using vehicles. Our work focus towards proposing an efficient information-centric data collection system for vehicles inspired from social networks. Therefore, we concentrate here on two aspects of the related work on data collection using vehicles a) How the ICN paradigm recently emerged along discussing different urban sensory data collection schemes proposed in vehicular network and b) How social network analysis motivates the identification and selection of important vehicles in underlying vehicular networks.

2.2.1 ICN meets Urban Data Collection

Urban sensing and vicinity monitoring using vehicles has attracted lots of researchers in the past few years and several schemes are proposed [27] [44], where sensor-equipped vehicles sense and share data in a vehicular network. For example, authors in [8] focus on the collection of multimedia data from urban streets using vehicles present on roads. Another example is CarTel [17] which is a distributed sensor communication system designed to gather, visualize and send data form vehicle-embedded sensors. CarSpeak [25] is another example which allows vehicle to collaborate and access sensory information captured by neighboring vehicles in the same manner as it can access its

own content. All these preliminary approaches proposed architectures and general frameworks to adapt ICN to vehicular network. Optimizing the data collection and storage was not specifically addressed.

Recruitment of such vehicles for urban sensing is being studied only since recently. For example, in [13], participants with high reputation are recruited to perform urban sensing. The idea is to cover an area of interest with a limited budget, however, the coverage metric is confined to particular road sections with the limitations of utilizing the infrastructure network. Moreover, the authors do not provide any metric to classify and identify the important participants. Another approach suggests that the ICN paradigm is as a promising solution to cater the peculiarities of the vehicular environment, characterized by dynamic topologies, unreliable broadcast channels, short-lived and intermittent connectivity [40]. Similarly, the authors in [54] studied the coverage problem for urban vehicular sensing where a metric called as Inter-Cover time is proposed to characterize the coverage opportunities as well as assess the coverage quality. Based on that, a vehicle selection algorithm is proposed to select the minimum number of vehicles to achieve the required coverage quality requirements. In both cases, the authors only considered coverage quality as the metric while ignoring the vehicle inherent abilities such as its availability and connectivity to facilitate data collection in an urban scenario. The above coverage metrics are unable to consider spatio-temporal coverage for the vehicles availability for the data collection and storage at different locations and times.

2.2.2 Social Network meets Vehicular Network

Identification of influential information hubs for publishing/spreading information is required in applications such as social networks. Another interesting application is found in medical sciences to find epidemic disease spreaders [24]. Similarly, Google's PageRank [33] algorithm ranks the importance of a web-page in an Internet search based on the number of web links directed towards it. More generally, Social Network Analysis (SNA) [36] is required to identify important nodes in a social network usually relying on well known network centrality schemes such as Degree, Closeness, Betweenness and Eigenvector centrality.

Degree centrality considers the number of direct (one hop) neighbors of a node, where Closeness centrality is the inverse of the sum of the lengths of the shortest paths from a node to the rest of the nodes in the network. Betweenness centrality is the fraction of all pairs of shortest paths passing through a node, where Eigenvector centrality is the node's influence measure in the network [7]. By tweaking these centrality measures, algorithms such as BubbleRap [16] and ML-SOR[38] are proposed, where nodes with high centrality score are preferred for data dissemination and routing in Opportunistic Social Networks. Another important work [4] suggests VIP delegation to offload network traffic based on opportunistic contacts. The authors consider well known social network attributes (betweenness, closeness, degree centrality and pagerank) to select delegate nodes in an urban area according to two methods; global (network-based) and hood (community-based) selection.

In a recent survey on Vehicular Social Networks (VSNs) [43], different sociability, security and applicability aspects of the VSN paradigm are discussed. The authors

advocate the emergence of social networks in vehicular scenarios due to the emergence of numerous novel crowd-sourcing applications and an increasing number of services based on social networking content sharing. In [48], the authors provide an extensive survey on different social-aware routing protocols for Delay Tolerant Networks (DTNs) and suggest to adapt the social metrics according to application requirements which motivates the need to define a new metric for important vehicle identification. However, It is unfeasible to use centrality-based popularity schemes to identify important information hubs in vehicular networks for multiple reasons discussed below.

2.2.3 Discussion

The reasons previously existing well known network centrality metrics are inapplicable in vehicular networks are as follows. First, the rapid topological changes due to the high mobility of vehicles requires a continuous time varying analysis of the network which is unfeasible by a practical scheme. Indeed, typical schemes assume a static graph topology with respect to time where the temporal network characteristics of vehicular network would be ignored. Second, centrality measures such as Betweenness, Closeness and Eigenvector centrality computation requires network-wide parameters, while in a vehicular network, a vehicle can hardly and with a high cost obtain such information to make run-time decisions. Third, existing schemes consider shortest path metric to compute a node's importance, while the highly dynamic network topologies does not ensure the existence of a stable path between nodes. Therefore, a new vehicle ranking system adapted to vehicular networks and enabling vehicles to decide their relative importance in the network by overcoming the above mentioned constraints need to be thought about.

Unlike the above mentioned applications, we exploit Information Centric Vehicular Networking for a distributed solution for urban data collection. In order to select "important" vehicles, there is a need for a vehicle centrality scheme to classify / rank vehicles as information producers, facilitators and consumers.

2.3 Data Storage in a Vehicular Network

2.3.1 Recruitment of Vehicles as Content Caches

Recruitment of vehicles is studied in [13] where participants with high reputation are recruited to perform urban sensing. The idea is to cover an area of interest with a limited budget, however, the coverage metric is confined to particular road sections with the limitations of utilizing the infrastructure network. The authors in [15] proposed an algorithm to recruit vehicles for crowd-sourcing with the objective to maximize the spatio-temporal coverage. The high quality participants are recruited based on vehicle mobility prediction considering the current and future location of candidate vehicles.

Similarly, [35] proposed recruitment schemes to choose participants for urban data collection based on geographic and temporal availability with consideration of participation habits. However, the authors do not provide any metric to classify and identify participants. In a recent work, [52] the authors presented a self-adaptive behavior-aware recruitment scheme for participatory sensing considering tempo-spatial behavior

and the incurred data quality. The recruitment scheme is modeled as a linear programming optimization problem by combining coverage, data quality, and budget. Another work [55] propose a framework to recruit vehicles with a vehicle selection algorithm for urban sensing while guaranteeing the specific coverage quality requirement.

The above discussed approaches focus is limited to urban sensing where the content is only generated from vehicles. At the same time, none of these approaches targets the issue of content caching for the internet traffic thus limiting their scope. We focus on the problem of content caching in mind with the aim to maximize content availability in an urban environment for any possible content on the internet. To the best of our knowledge based on the latest research [34], there exists no such work focusing on content caching on vehicles.

Moreover, existing efforts are focusing to recruit vehicles while considering only coverage as a requirement under limited budget.This is performed while considering that all vehicles are cooperative, which is not always true. In our case, we target a novel incentive-driven approach towards the autonomous identification of potential content cache vehicles while ensuing fairness by rewarding vehicles according to its natural mobility pattern. At the same time, we tackle the vehicle rational behavior by formulating it as a social welfare game where a vehicle perform content caching in the required locations without perturbing its daily commute.

2.3.2 Content Cache Management

Content Cache management is the attention of several research studies for some time specially with the introduction of ICNs [9] [6]. The inherent in-network caching and mobility support in ICN for a publish-subscribe case is studied in [49]. In [39] distributed cache management decisions are made in order to efficiently place replicas of information in dedicated storage devices attached to nodes of the network using Information Centric Networking (ICN). A recent work [51] pre-fectch content to be cached at edge nodes to support adaptive video streaming in ICN. In [53] the authors presented a mathematical framework based on Markov Chain to evaluate the caching performance for various caching policies.

Similarly [46] address the distribution of the cache capacity across routers under a constrained total storage budget for the network. A suboptimal heuristic method is proposed based on node centrality for the case of dynamic networks with frequent content publishing. The authors found that network topology and content popularity are two important factors that affect where exactly should cache capacity be placed. In another work [41], the authors explored the possibility of probabilistic caching decisions as a content replacement scheme.

In a recent work [28], game theory is exploited for caching popular videos at small cell base stations (SBSs). A Stackelberg game model is presented for the service providers to lease its SBSs to video retailers in order to gain profits as well as reduce the costs for back-haul channel transmissions. The authors aim to maximize the average profit of the service provider as well as individual video retailers by solving a complex optimization problem with (i) uniform and (ii) non uniform pricing schemes. Similarly [31] proposed a game theoretic approach in ICN to stimulate wireless access point owners to jointly lease their unused bandwidth and storage space to a content provider

under partial coverage constraints. An algorithm is provided to determine the optimal allocation of mobile clients to access points that ensures the individual rationality as well as the truthfulness properties by forcing the AP owners to bid the real valuation for the offered resources.

2.3.3 Social Networks meet Caching

An example of Social aware caching approach in Delay Tolerant Networks (DTNs) is presented in [26] where the authors proposed a cooperative caching scheme based on the social relationship among nodes in DTNs. Content is dynamically cached at selective locations in the network such as cluster head nodes, which have the highest social levels, and nodes along the common request forwarding paths. A cache replacement policy based on the content popularity is presented as a function of both the frequency and recency of data access. Similarly, Socially-Aware Caching Strategy (SACS) [5] for Content Centric Networks uses social information and privileges Influential users in the network by pro-actively caching the content they produce. The authors detect the influence of users within a social network by using the Eigenvector and PageRank centrality measures.

In [47], an optimization method is presented to find the optimal cache allocation in CCN. The authors studied factors that affect cache placement, and how they subsequently impact performance (measured by traffic reduction) by proposing a centrality-based heuristic for dynamic networks with frequent content publishing. Results suggested that network operators to make the cache allocation decision based on their network topology and content access patterns. Another important work on social aware cooperative caching[56] highlights the effects of contact duration on caching in DTNs where high centrality nodes are used to cache content. Authors propose to let the high centrality nodes transfer some data items with low priority to the node with lower centrality to make room for the data items with high priority.

2.3.4 Discussion

Existing schemes considers only content popularity along policies such as FIFO, LRU an LFU, which fails to spatio-temporally characterize content. We transform the caching problem in CCN towards a novel approach where, unlike all the above mentioned approaches, (i) we derive a relation between spatio-temporal content availability and popularity to decide it caching decisions as all content are not equally popular. (ii) Our popularity metric is adapted to urban mobile environment while considering vehicle local as well as global reachability. It does not rely on typical centrality metrics which are unsuitable and unstable in dynamic connectivity scenarios, and (iii) The mobile fog formation is a novel approach to ensure maximum content availability with the best candidate vehicles in coalition for different locations and timings, while to our knowledge, there exist no scheme to provide such a scalable content caching model to offload in urban environment.

2.4 Conclusion

The existing literature on data collection lacks stable metrics to identify suitable candidate vehicles in a dynamic environment. Such vehicles should be important and available to satisfy relevant user interests in the network. Additionally, existing approaches do not provide a scalable (city-wide) data collection scheme for a given coverage and budget constraints. We propose VISIT to address this by first introducing novel metrics to identify the important candidate vehicles appropriate considering the user social interests, the vehicle connectivity and its availability to satisfy user interests at different locations and times. Then, we present an optimal selection algorithm to find the set of best vehicles for a scalable data collection under a given spatio-temporal coverage with minimum redundancy along budget constraints. Similarly, data storage on a large scale requires the identification of important hubs to efficiently store and retrieve data. Such information hub vehicles should be accessible from different locations. At the same time, these vehicles should be taking into consideration the limited budget and coverage requirements for different locations. None of the existing work provides such a scalable solution for efficient identification and selection of suitable set of vehicles. Moreover, the content cache management schemes discussed above do not address the content caching decisions at the set of vehicles with respect to user interests. Our approach SAVING not only identify appropriate vehicles for efficient urban content caching based on the vehicle and content reachability with respect to the user interests, but also provides solutions for how to select the best vehicles, and how to optimally manage the content caching between such vehicles under given budgetary and coverage constraints.

Chapter 3

VISIT for Data Collection: Novel Centrality Metrics to Identify Eligible Candidates

3.1 Introduction

We target the data collection problem in an urban environment assuming each vehicle constantly generates and consumes massive amount of heterogeneous data. The high data volume and cost makes it unfeasible to upload such data to the cloud or Internet. Assuming lots of vehicles on urban roads, the first step is to identify relevant vehicles as the best candidate for urban data collection.

We model the problem as important node identification in a dynamic network by allowing nodes in the network find its importance using appropriate node centrality schemes. There exists different centrality schemes in the literature but they are inapplicable in a dynamic vehicular network. To cater the issue, this chapter presents two novel centrality metrics allowing a smart vehicle to first classify different location aware information using a metric, InfoRank, while taking into consideration a content relevance with respect to the users social interest. It then combines InfoRank score with a novel vehicle centrality metric, CarRank, to find its eligibility for efficient data collection and storage with respect to the user social interests, its spatio-temporal availability and its neighborhood connectivity in an urban scenario. Our novel schemes are capable to identify nodes which can satisfy twice the amount of user interests and 4 times more throughput compared to benchmark centrality schemes.

The chapter is organized as follows. The next section discusses the context and motivation for the identification of eligible candidates. Section 3.3 present the information importance metric InfoRank followed by describing the novel vehicle centrality scheme in Section 3.4. Performance evaluation and results to validate the two metrics are discussed in Section 3.5 and the Section 3.6 concludes the chapter.

3.2 Context and Motivation

Vehicles on the road today are constantly generating and consuming a tremendous amount of data that cannot be uploaded to the cloud or Internet due to its large volume. Moreover, most of the generated content is of "local relevance" as the intended users lies within the vehicular network. Relying on the infrastructure network for the collection, storage and distribution of such heterogeneous *Big-Data* from vehicles can thus prove costly and inadequate to its usage. Pre-advertising or broadcasting all the sensing data from each vehicle would result in a massive advertising overhead and a redundant information storm within the network. The major problem is to efficiently locate and collect the user relevant data from the fleet of vehicles with the underlying challenge of intermittent connectivity and mobility in a Vehicular Ad-hoc Network (VANET).

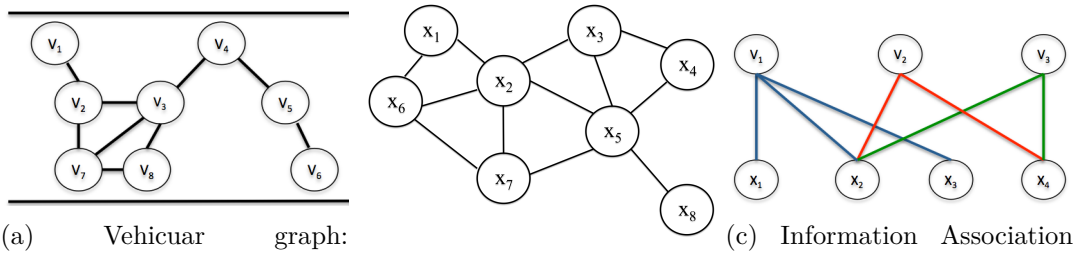
This motivates the need to identify important vehicles to be recruited for distributed data collection based on their daily commute and their social importance with respect to the frequently visited neighborhoods. To identify important nodes, network analysis typically rely on different variants of centrality measures such as Degree, Closeness, Betweenness or Eigenvector centrality. However, such schemes are difficult to use in the sporadic vehicular network topology. Therefore, the challenge is to find the right vehicle available at the right time and place for efficient data collection, storage and distribution through low-cost inter-vehicle communications.

To address this problem, for the first time in vehicular networks, we propose a new concept of finding important vehicles, allowing a smart vehicle to rank itself based on its popularity with respect to the user interests, spatio-temporal availability and its neighborhood in an urban scenario. We envision such vehicles as buses, taxis, commuters available to address user interests in the network. Therefore, the target of this chapter is to introduce innovative vehicle eligibility metrics "InfoRank" and "CarRank" for the identification of Information Facilitator Vehicles (IFVs), responsible for the efficient gathering and publishing of data from urban streets. The vehicle first ranks the information associated to it taking into consideration the relevance to the users interest using the "InfoRank" metric. It then considers the associated location-aware information popularity to find its relative importance in the network using the CarRank algorithm as its *vehicle centrality*.

The major contributions to this chapter can thus be summarized as follows:

- A novel algorithm "InfoRank" [20], enabling a vehicle to autonomously (1) rank important location-aware information associated to it based on the satisfied user interests and the information validity and (2) rank itself based on the information importance and its mobility pattern in the city.
- An innovative vehicle ranking algorithm, "CarRank", is proposed, where each vehicle can find its importance in the network. This importance is linked to the importance of the associated information, vehicle spatio-temporal availability and the neighborhood topology.

The objective here is to show that the proposed algorithms are well suited to help in the efficient identification of the best candidate vehicles in the network using information-centric vehicular networking.



(a) Vehicuar graph: $G(\mathbb{V}(t), \mathbb{E}^v(t))$

(b) Location graph $G(\mathbb{X}, \mathbb{E}^x)$

(c) Information Association Graph $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$

Figure 3.1: Network Model

3.3 InfoRank: An Information Importance Based Centrality Scheme

3.3.1 Network Model

We consider a time varying vehicular network modeled as an undirected vehicular graph $G(\mathbb{V}(t), \mathbb{E}^v(t))$, where $\mathbb{V}(t) = \{v\}$ is a set of vertices v, each representing a vehicle on the road at time t. $\mathbb{E}^v(t) = \{e_{jk}(t) \mid v_j, v_k \in \mathbb{V}, j \neq k\}$ is the set of edges $e_{jk}(t)$ modeling the existence of a direct communication link between vehicles j and k at time t. The number of edges $\mathbb{E}^v(t)$ depends on the transmission range of each vehicle as shown in Figure 3.1a. We assume it as a simple unit disk model bounded by its communication range. The city map is represented by the undirected graph $G(\mathbb{X}, \mathbb{E}^x)$ as in Figure 3.1b, the set of vertices $\mathbb{X} = \{x\}$ represents location-aware content for different urban zones x and the set of edges $\mathbb{E}^x = \{e_{pq} \mid x_p, x_q \in \mathbb{X}, p \neq q\}$ are their respective boundaries that connects different zones through the underlying road network. We define x as a piece of location-aware content cached at a vehicle v which reflect a content associated to a location in an urban environment.

The following section defines the network model enabling the vehicle to compute the respective InfoRank and CarRank score.

Information Association

Information association is defined as a bipartite graph $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$, where $\mathbb{V}$ is the set the vertices in the vehicular graph $G(\mathbb{V}(t), \mathbb{E}^v(t))$ and $\mathbb{X}$ is the set of locations in the city map $G(\mathbb{X}, \mathbb{E}^x)$ as shown in Figure 3.1c. The edge $\mathbb{E} = \{e_{ij} \,|v_i \in \mathbb{V}, x_j \in \mathbb{X}\}$ associates each vehicle to a set of regions $X_v \subset \mathbb{X}$ with respect to the user relevant content.

The information association decouple the dynamics of the vehicular graph by linking it with the stable nature of location graph. Despite the vehicles rapid mobility, $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$ provides a relatively stable location-aware information association. The associated information is classified by clustering the regions using ICN hierarchical naming convention *"/region/road-section/information-type"*. Information type comprises different Intelligent ITS applications (Safety warnings, Road congestion information, Infotainment...) with varying content popularity and priority.

For temporal VANET analysis, we divide the time $T = (\overline{t_1}, \overline{t_2}, ...)$ into a sequence of regular time-slots, where the k^{th} time-slot is $\overline{t_k} = [t_k, t_{k+1})$. Each vehicle finds its

Table 3.1: VISIT I - List of Notations

Notation	Description
$\mathbb{V}$	Set of vehicles
$\mathbb{X}$	Set of locations/regions
$\mathbb{E}^x/\mathbb{E}^v$	Set of edges between locations/vehicles
$\mathbb{E}$	Edge between vehicles and locations
$\bar{t}_k$	Time-slot k for CarRank computation
t_k/t_{k+1}	Current time instant/next time instant
X_v	Set of locations associated to vehicle v
$d(x, x_k)$	Distance from current location x_k to x
I_x^v	Interests satisfaction frequency for x
r_x	Number of successful responds for x in the previous slot
R_x	Total successful responds for x
R_T	Vehicle responds count for all contents
t_x^f	Last successful respond time for x
$\bar{t}_d$	Average interest deadline
n	Total received interests in the previous slot
t_x	Interest validity deadline for content x
τ	Information timeliness
δ	Tuning parameter for information validity
C_x^v	Content x importance for vehicle v
λ	Tune importance based on distance from x
s_x^v	vehicle reliability as content source for x
w_x	Information x weight with respect to vehicle
f_I^v	Information importance function
$p_x^v(t_k, x_k)$	Probability of satisfying interests for location x at current time t_k and position x_k
$R_x^v(t_k, x_k)$	Interests satisfied for content x at current time t_k and position x_k
$I_x^v(t_k; x_k)$	Mutual information shared between the current time and location for content x
$p_x^v(t_k)$	Marginal probability of interest responds at current time
$p_x^v(x_k)$	Marginal probability of interest responds at current location
$f_{T,X}^v$	Vehicle spatio-temporal availability function
k_v	Vehicle degree (number of neighbors)
k_Γ^v	Vehicle average neighbor degree
Γ_v	Set of neighbors for vehicle v
C_Γ^v	Neighbor vehicle centrality
f_Γ^v	Vehicle neighborhood importance function
C_v	Vehicle centrality
$\alpha/\beta/\ \gamma$	Tuning parameters for each function
H_v	Vehicle coverage entropy
θ	Smoothing factor for vehicle centrality

centrality at the time instant t_{k+1} from the known information in the current time-slot, where t_k is the time instant at the beginning of the time-slot $\overline{t_k}$. We will refer to content/information or location/areas/zones interchangeably in the text since content are associated to locations in the urban map.

The information distance $d(x, x_k)$ is the distance between the content location x and the vehicle current position x_k at time instant t_{k+1}, where $x, x_k \in X_v$. Information distance can be computed either as Curve-metric or Euclidean distance between the current vehicle's location and the information's location to assess its importance. We assume each vehicle knows the city map, i.e. the graph $G(\mathbb{X}, \mathbb{E}^x)$, but their cached-content related to this map is limited to the already visited locations or to the content received from neighboring vehicles. This is due to the limited storage capacity at vehicles as well as the limited vehicle coverage scope as the information it possesses is related to either its daily commute or associated neighborhoods.

InfoRank is a centrality measure enabling each vehicle to autonomously find the importance of different locations independent of a centralized database. We do not merely rely on the rapidly changing inter-vehicle contacts since such unstable behavior of the frequency and duration of vehicle contacts do not provide any useful information to decide a location importance in the time evolving vehicular network. InfoRank considers the user interests frequency for content associated to different location as a key metric to classify the importance of different locations with respect to the vehicle as it regularly responds to user interests. It also considers the information validity scope as a metric towards finding the importance for a particular information. Before describing the importance metrics, the following section defines the network model followed by the description of the information importance metrics.

3.3.2 User Interests Satisfaction

We assume vehicles in a distributed VANET encountering each other constantly receiving interests from neighboring vehicles for different location-dependent information. Some of such information can be of more importance to the intended users in the network which can be easily identified by the vehicle by the amount of user interests received for it. We assume that the vehicle is capable of recording the time and position each time it responds as the content provider to a user interest. Therefore, it considers an information as popular if it observes an increase in the number of user interests for a particular location. For this reason, information importance takes into account the vehicle latent ability to satisfy more user interests with its natural mobility pattern.

Definition 1

(Interest Satisfaction Frequency) We define $I_x^v(\bar{t}_k) = \frac{r_x(\overline{t_k})}{R_x}$ as the frequency of user interests satisfied in the time-slot $\overline{t_k}$, where $r_x(\overline{t_k})$ are the number of successful responds in the previous slot and R_x are the overall successful responds for the content $x \in X_v$ associated to the vehicle $v \in \mathbb{V}$.

3.3.3 Information Validity Scope

The importance of each location-aware content is periodically updated based on the interest satisfaction frequency by the vehicle. Interest for each content specifies a temporal scope for the information validity. For instance, road congestion information is only valid during congestion. Therefore, in order to ensure the information importance is not substantially increased after the desired deadline, let t_x^f be the last successful respond time for the content x and the average interest deadline as $\bar{t}_d = \frac{1}{n}\sum_n t_x$ associated with each content, where n are the total number of interests in the previous time-slot and t_x is the deadline of each interest for the content x.

Definition 2

(Information Timeliness) The information timeliness defined as τ where

$$\tau(t_{k+1}) = \begin{cases} 1 & t_{k+1} \leq t_x^f + \bar{t}_d \\ e^{-\delta \bar{t}_d} & t_{k+1} > t_x^f + \bar{t}_d \end{cases}$$

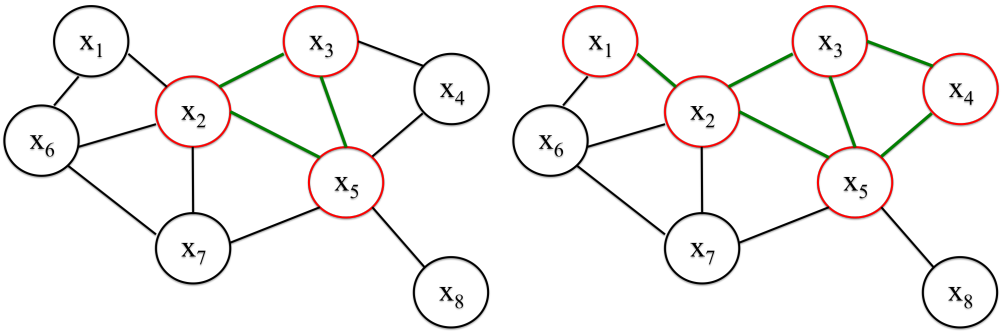
is the measure of the temporal information validity scope where $\delta \in [0, 1]$ is the tuning parameter depending on the application needs. It is used in the information timeliness metric where it tells us how quickly to decay the unnecessary increase in the information importance for the content which is not needed. An example would be the diminishing of the validity of the information regarding a road congestion which is now cleared, where delta defines the rapidness of the decay for such information. This decay can vary between different applications requiring different information diminishing time, thus we left it as a tunable metric.

For each information type, the information timeliness parameter τ considers its validity at the importance computation time instant t_{k+1}. If there are no active interests in the previous slot and the average interest validity deadline has passed, the information importance follows an exponential decay since the information is of less importance in the network. On the other hand, τ is set to unity for the information required to be always available in the network.

The corresponding information importance at the next time instant t_{k+1} is updated as follows:

$$\begin{aligned} C_x^v(t_{k+1}) = C_x^v(t_k) + \tau(t_{k+1}) I_x^v(\overline{t_k})(1 + d(x, x_k))^{-\lambda} \\ + s_x^v(t_{k+1}) \end{aligned} \tag{3.1}$$

The information importance depends on the its value $C_x^v(t_k)$ at the beginning of the time-slot (time instant t_k). If a content is not responded in the previous slot, then $I_x^v(\overline{t_k}) = 0$ avoids unnecessary increase in the information importance. The term $s_x^v \in [0, 1]$ represents the percentage of time the vehicle itself acted as the original source for the content x. It is updated regularly to ensure the content relevant to the vehicle retains its association in case it does not respond in the previous slot. Thus, the interests for a particular content later in time could finally route to the original source vehicle in the network. The tuning parameter $\lambda \in [0, 1]$ decides the effect of the



(a) Vehicle A city-wide mobility pattern (b) Vehicle B city-wide mobility pattern

Figure 3.2: An example of vehicle city-wide coverage scope as a metric of its mobility pattern

vehicle distance from the associated content location. It is used to let the application decide how much importance should be given to the distance from the information, it is useful for applications relying on location-based information and would like to increase/decrease information importance based on vehicle distance where the distance from information affects its importance.

INPUT: $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$: information association graph
OUTPUT: Updated InfoRank for the next time-slot at time-instant t_{k+1}
for each vehicle $v \in \mathbb{V}$ **do**
  for each content $x \in X_v$ in cache **do**
    Find $d(x, x_k), \tau(t_{k+1}), s_x^v(t_{k+1}), w_x$
    Compute $I_x^v(\bar{t}) \leftarrow \frac{r_x(\bar{t})}{R_x}$
    if $I_x^v(\bar{t}) \neq 0$ **then**
      Update $C_x^v(t_{k+1})$ using (3.1)
    else
      $C_x^v(t_{k+1}) = C_x^v(t_k) + s_x^v(t_{k+1})$,
    end if
  end for
  Find missed interests ratio $I_m^v(t_{k+1})$, Coverage entropy $H^v(t_{k+1})$
  Compute $f_I^v(t_{k+1})$ using (3.2)
end for

Algorithm 1: InfoRank

3.3.4 InfoRank Computation

The vehicle considers its importance with respect to the associated information in order to measure its influence in the networks. Besides information importance in (3.1), we also consider the overall coverage scope as an important parameter to decide a vehicle importance in an urban environment.

Definition 3

(Coverage Entropy) We define $H^v = -\sum_{x \in \mathbb{X}} p(x) \log p(x)$, as the coverage entropy of the vehicle periodically computed with respect to the entire city map (i.e vehicle associated sub-graph $X_v \in G(\mathbb{X}, \mathbb{E}^x)$). The probability $p(x)$ is the visiting frequency to each region $x \in \mathbb{X}$ before the importance computation time t_{k+1}.

The vehicle's coverage in the map can be represented as a set of mobility between regions since the urban map is divided into regions/zones where each vehicle travels between adjacent regions. Therefore, the overall vehicle mobility pattern is mapped as the set of visited regions depending on its daily commute. For example, the vehicles A and B coverage scope are bounded by the set of regions $M^A = \{x_3, x_2, x_2, x_3, x_5, x_2\}$ and $M^B = \{x_1, x_2, x_3, x_5, x_4, x_3\}$, where $x \in \mathbb{X}$. The vehicle A visits the regions x_2,x_3 and x_5 with probabilities $\frac{2}{6}$,$\frac{3}{6}$ and $\frac{1}{6}$, while B visits the regions x_1,x_2 x_3,x_4, and x_5 with probabilities $\frac{1}{6}$,$\frac{1}{6}$,$\frac{2}{6}$,$\frac{1}{6}$ and $\frac{1}{6}$ respectively. The corresponding coverage entropy for the mobility pattern as shown in Figure 3.2 is calculated as:

$$H^A = -\tfrac{2}{6}\log\tfrac{2}{6} - \tfrac{3}{6}\log\tfrac{3}{6} - \tfrac{1}{6}\log\tfrac{1}{6} = 0.439,$$
$$H^B = -\tfrac{2}{6}\log\tfrac{2}{6} - \left(\tfrac{1}{6}\log\tfrac{1}{6} * 4\right) = 0.639,$$

Vehicle A has a narrow coverage scope due to its limited geographical coverage, while B has a wider geographical coverage with respect to the urban map. Therefore, we consider coverage entropy as the coverage metric for the vehicle importance with respect to all locations in the city.

Algorithm 1 shows the steps allowing a vehicle to find the respective InfoRank. For a given location-dependent content in cache, the corresponding information importance is updated for the next time-slot at time instant t_{k+1}. The information-centric centrality function is given as:

$$f_I^v(t_{k+1}) = \frac{(1 + I_m^v(t_{k+1}))^{-\epsilon}}{|X_v|} \sum_{x \in X_v} C_x^v(t_{k+1}) \cdot w_x + H^v(t_{k+1}) \tag{3.2}$$

For all contents $x \in X_v$ associated to v, $I_m^v(t_{k+1})$ are the ratio of missed interest to the total interests received by the vehicle while ϵ is the tuning parameter. Missed interest provides the vehicle reliability regarding successful respond to the incoming interests. $C_x^v(t_{k+1})$ is the respective content importance at time instant t_{k+1}, $w_x = \frac{R_x}{R_T}$ is the edge weight of information association graph $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$ considering the interest satisfied for the content x among all the contents in cache. R_x is the number of responds for x and R_T is the number of responds for all contents in the cache. $|X_v|$ is the cardinality of the sub-graph $X_v \subset \mathbb{X}$, all regions associated to the vehicle $v \in \mathbb{V}$. The term ϵ is used to decide to what extent the received interests which are not satisfied by the vehicle plays a role in its importance. For example, in case a service provider recruits vehicles as gateways where the vehicle naturally tends to receive relatively more interests than it satisfies, in such a case the service provider can tune epsilon to not depending much on the received interests not satisfied (missed) by the vehicle.

InfoRank metrics described above are defined as the local scope of the information relevance with respect to a particular location in time and space. Regular visits to popular locations at well interesting time of the day will increase the vehicle's popularity in the network. However, the vehicle global mobility pattern in a city is bounded

by the regions only known to the vehicle (visited before). Moreover, stale information is automatically deleted from the cache after some time due to the limited size storage buffer at vehicles. One should note that different cache management schemes are developed in ICN which will be addressed in Chapter 6 in details.

3.4 CarRank: A Vehicle Centrality Algorithm

We begin by an example of a smart vehicle equipped with an On-Board Unit (OBU) interconnecting the different sensors and cameras to sense and monitor its vicinity in an urban environment. The data harvested by a vehicle is pre-processed by the on-board processing unit and then relayed to a better ranked vehicle within the vehicular network Set of high ranked vehicles (E.g. buses, taxis, etc.) are responsible for the gathering, storing and publishing of data from source vehicles. The user vehicle broadcasts an "interest" by content name to a better ranked vehicle, any nearby host(s) containing the data responds back to the user request with the desired content using the underlying name-based architecture.

We present CarRank as a novel vehicle centrality metric enabling each vehicle to autonomously find its importance in the network. It is difficult to use the vehicle contact frequency and duration to decide its importance in the network due to the rapid changes in the time evolving vehicular network topology. To overcome this, CarRank simultaneously considers three novel albeit essential parameters, the information importance, the vehicle spatio-temporal availability and its network connectivity. Additionally, the user's interest satisfaction for a content is also considered as a key metric for a vehicle's importance as it regularly responds to user interests. The interests are assumed to be generated and received from the neighboring vehicles using multi-hop interest forwarding. We consider the following local parameters known to the vehicle for analytically finding its importance:

Information Importance

Information importance measures vehicle relevance to users for a particular content, i.e. The interest-response frequency is a vital factor to classify a content's importance. A vehicle associated to contents related to popular locations is considered as an important information hub in the network.

Spatio-Temporal availability

It reflects the social-behavior based on the vehicle's habitual routes as a factor of the daily commute. Spatial availability reflects the vehicle's recursive presence in an area, while temporal availability refers to its relevance in time for a location.

Neighborhood Importance

Neighborhood importance shows vehicle topological connectivity in order to be capable to facilitate different LBS applications. An easily reachable and well connected vehicle in a network topology can better disseminate information within the network.

3.4.1 Information Importance

Vehicles encountering each other constantly receives interests from neighboring vehicles for different location-aware information. Some of such information can be of relatively more importance to the intended users in the network easily classified by the vehicle by the amount and frequency of received user interests. We assume that it is capable of recording the time and position each time it responds as the content provider to user interests. Therefore, a vehicle considers an information as popular if it observes an increase in the number of user interests for it. For this reason, we integrate in CarRank, the information importance metric InfoRank (f_I^v using Equation 3.2) defined in the previous section which takes into account the vehicle latent ability to satisfy more user interests with its natural mobility pattern and availability.

3.4.2 Spatio-temporal Availability

Spatio-temporal availability of the vehicle reflects the driver social behavior. It considers the vehicle physical availability in an area while taking into account different times of the day. For example, we drive the same route around the same time of the day to go to places we visit habitually such as our work place or the gym. Users are likely to be located in the same city neighborhood which is related to their daily routine. The challenge lies in the fact that each user natural mobility scope is bounded by the geographical regions that are only relevant to its daily commute, thus making it difficult to derive a distributed method to find its importance without relying on the complete network topology.

However, to incorporate such social behavior, we borrow tools from information theory to find to what extent the current time and location contribute to the vehicle's importance. Since the vehicle does not have network-wide information to find its relevance, we continue with our proposed interest satisfaction ratio based assumption. The probability of the vehicle v satisfying interests for content for location x at the current time t_k and position x_k is $p_x^v(t_k, x_k) = \frac{R_x^v(t_k, x_k)}{R_T}$, where $R_x^v(t_k, x_k)$ are the interests satisfied for content x at the current time and location in the past and R_T are the total interests satisfied by the vehicle v. The current time in the past refers to the time-slot around the same time in the day for all the days before the present day Y with respect to each content as shown in Figure 3.3. For example, for finding the spatio-temporal availability between 7 AM and 8 AM, it compares the interest satisfied in the same area around 7 AM and 8 AM in the past for all content in cache.

For content associated to location x, the mutual information shared between all the correlating time-slots and the locations is:

$$I_x^v(t_k; x_k) = \sum_{\forall t_k \in T} \sum_{\forall x \in X_v} p_x^v(t_k, x_k) \log \left(\frac{p_x^v(t_k, x_k)}{p_x^v(t_k) p_x^v(x_k)} \right), \tag{3.3}$$

where $p_x^v(t_k)$ and $p_x^v(x_k)$ are the marginal probabilities of the content responds in the current time and the current location, respectively. Now, the vehicle finds its spatio-temporal availability function for all locations:

$$f_{T,X}^v(t_{k+1}) = \frac{1}{|X_v|} \sum_{x \in X_v} I_x^v(t_k; x_k).w_x \tag{3.4}$$

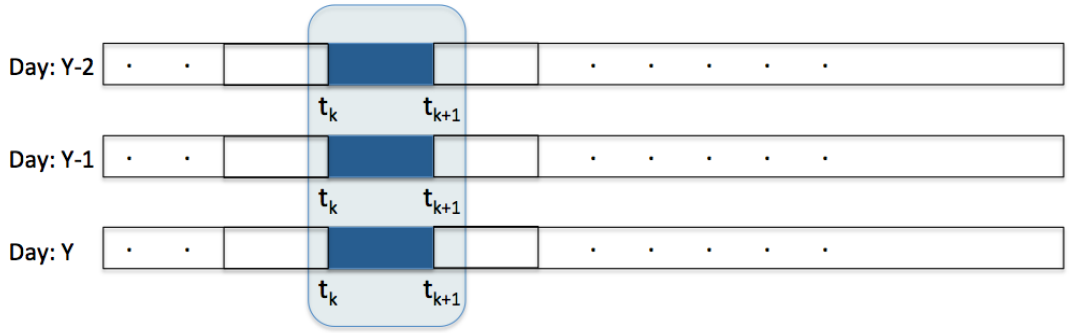


Figure 3.3: Spatio-temporal availaibility in the same time-slot

The function in (3.4) indicates a vehicle's importance at the time and position of CarRank computation. If it correlates more to the associated contents at the current time and location, it counts more towards computing its respective score at the same hour of the day and the same location.

3.4.3 Neighborhood Importance

The neighborhood of the vehicle in a distributed system is important for efficient content distribution and storage. We incorporate the neighborhood information by letting the vehicles in transmission range share their respective importance as well as their connectivity information. The idea is to consider better connected vehicles with better spreading capabilities. This instantiates the use of the vehicles physical topological information. For this purpose, we consider the vehicle's assortativity as its average neighbor degree k_Γ^v. Besides topological connectivity, each neighbor centrality C_Γ^v within communication range at time t_k is also taken into account. The neighborhood importance function for the time-slot t_{k+1} is expressed as:

$$f_\Gamma^v(t_{k+1}) = \frac{1}{k_v} \sum_{\Gamma_v \in \mathbb{V}} C_\Gamma^v(t_k) \cdot k_\Gamma^v \tag{3.5}$$

where k_v is the vehicle degree at time t in the graph $G(\mathbb{V}(t), \mathbb{E}^v(t))$. Since it is impossible to use any network-wide centrality measure unknown to the vehicle at the time of importance computation. Therefore, the function $f_\Gamma^v(t_{k+1})$ in (3.5) considers more information than just the degree of the vehicle while maintaining a local scope, thus, relying only on local information within the vehicle range as shown for the node V_3 in Figure 3.4.

3.4.4 CarRank Computation

The vehicle centrality for the next time instant t_{k+1} is updated as the Exponential Weighted Moving Average (EWMA) function of the current and previous vehicle centrality, where (3.2), (3.4) and (3.5) contributes to the overall CarRank computation:

$$\begin{aligned} C_v(t_{k+1}) = \theta C_v(t_k) + (1-\theta)[\alpha f_I^v(t_{k+1}) \\ + \beta f_{T,X}^v(t_{k+1}) + \gamma f_\Gamma^v(t_{k+1})] \end{aligned} \tag{3.6}$$

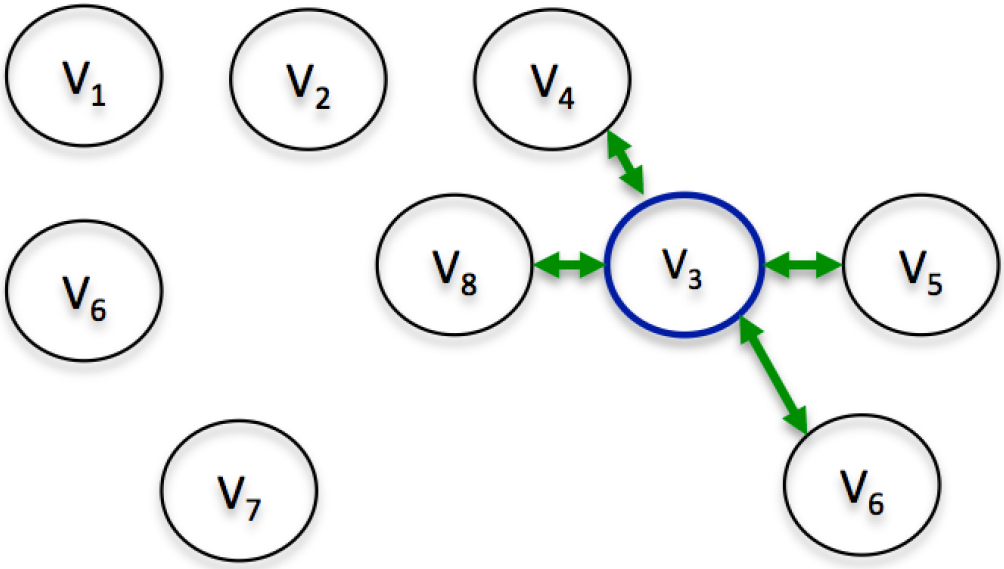


Figure 3.4: Neighborhood Centrality Exchange

Each function's contribution is normalized by the terms α, β and γ, where $\alpha + \beta + \gamma = 1$, where $\theta \in [0, 1]$ allows the vehicle to increase its importance with respect to the previous time-slot. The impact of each parameter differs with respect to different applications. For example, if the vehicle is located in a better connected neighborhood, it can easily spread information. Therefore, the corresponding vehicle weights the information importance along the neighborhood more than the spatio-temporal availability.

The different steps required for a vehicle to find its CarRank score are described in Algorithm 2. Using the importance of all the associated contents, the vehicle finds its CarRank score from the respective information importance function obtained using InfoRank (3.2). We also need the vehicle spatial and temporal availability by finding the mutual information shared between the current time and location for all the associated contents. The vehicle finds its topological connectivity measure from the neighborhood. It exchanges the average neighbor degree along the centrality score with the neighboring vehicles in range. The information importance function, spatio-temporal availability and neighborhood all-together contribute to the vehicle centrality score for the next time-slot.

CarRank metrics described above are defined as the local scope of the information relevance with respect to a particular location in time and space. Regular visits to popular locations at well interesting time of the day will increase the vehicle's importance in the network. However, the vehicle global mobility pattern in a city is bounded by the regions only known to the vehicle (visited before). Thus, the more the number of popular locations visited by the vehicle, the more it increases its centrality. Moreover, Stale information is automatically deleted from the cache after some time due to the limited size storage buffer at vehicles.

INPUT: Information association graph $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$:
OUTPUT: $C_v(t_{k+1})$: Updated CarRank for the next time-slot t_{k+1}
for each vehicle $v \in \mathbb{V}$ **do**
  for each associated content $x \in X_v$ **do**
    Compute associated information importance using InfoRank (3.2)
    Compute mutual information with respect to the content from (3.3)
  end for
  for each neighbor vehicle $\Gamma_v \in \mathbb{V}$ **do**
    $k_{\Gamma}^{v} \leftarrow$ average neighbor degree
    $C_{\Gamma}^{v}(t_k) \leftarrow$ neighbor centrality
  end for
  Find spatio-temporal availability using (3.4)
  Compute neighborhood importance using (3.5)
  Update vehicle centrality from (3.6)
end for
return $C_v(t_{k+1})$

Algorithm 2: CarRank

3.5 Performance Evaluation

The objective of our simulation study is to find answers to the fundamental question: *How well can we identify the top vehicles capable to perform urban data collection?* To address the scalability requirements, we use Network Simulator-3 (NS-3) as a scalable simulation platform for up to three thousand vehicles. The performance of the proposed algorithms are validated by a set of simulation runs under a realistic mobility scenario using traces from Cologne, Germany. To the best of our knowledge, it is considered as the most accurate mobility trace available for Vehicular Networks [42]. The vehicle availability as well as its mobility pattern is extracted using this trace. The simulation parameters are summarized in Table 3.2, followed by a description of the simulation scenarios implemented for the performance evaluation.

3.5.1 Simulation Scenario

We simulate a vehicular urban data collection scenario using the ndnSIM [1] module available for NS-3. ndnSIM integrates the Named Data Networking (NDN) communication model where the name based architecture replaces the traditional IPv4/v6 NS-3 network-layer modules. To incorporate the effect of buildings and environment, we implement a combination of Nakagami propagation loss model for fast fading and Log distance path loss propagation model with three distance fields (near, middle and far) with different exponents. The simulation scenario implements the following two applications:

Provider

We define a provider vehicle to be the content source in the network. The areas visited by a vehicle in a time-slot before the CarRank computation time are considered as

Table 3.2: Simulation Parameters for VISIT

Parameter	Value
Simulation platform	NS-3
Number of nodes	2986
Mobility trace	Cologne, Germany
Area	$6X6km^2$ city center
Duration	1 hour
Communication range	100m
Packet size	1024 bytes
Time granularity	1 sec
Simulation Runs	10

locations associated with the provider.

Consumer

Consumer vehicles are the potential user nodes planning to visit an area. Each consumer vehicle generates an interest for a content associated to a location in the city, which is routed to provider vehicles.

We assume the interests follow a Zipf distribution, where interests for popular contents are more frequent. The city map is divided into zones/areas as voronoi tessellation where vehicles in proximity of each other by average values of their coordinates are co-located within the same voronoi region at the current time-slots. Any provider acting as source for an area upon receiving the interest responds with the desired content. For each vehicle, the vehicle centrality is computed at regular instants using Algorithm 2 to identify the best candidate vehicles for urban data collection. We perform each simulation upto ten times by analyzing different set of nodes as information providers and consumers. The tuning parameters α, β and γ are set to 0.33 and δ, ϵ, λ and θ are set to 0.5 in order to maintain generality since the significance of each parameter depends on the application requirements. The communication range of $100m$ is considered as the worst-case range with the assumption that at least $100m$ is relatively easy to attain. For each simulation scenario, we rank the top vehicles in the network by comparing their CarRank score with the respective Degree, Closeness, Betweenness and Eigenvector centrality score. The "top" vehicles relates to the best ranked vehicles from the 2986 analyzed in our case. However for a municipal administration, the number of vehicles depends on the given budget, vehicle cost, different coverage and redundancy constraints.

Table 3.3: InfoRank in different set of Simulations

Simulation	1		2		3		4		5		6		7		8		9		10			
InfoRank	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	Mean	Confidence Interval
1	1115	1	26	1	1070	1	1248	1	1867	1	57	1	1867	1	1961	1	1248	1	2149	1	1	0
2	318	0.9926	115	0.9828	612	0.9927	2372	0.9913	1902	0.9389	1195	0.9670	373	0.9859	2028	0.9825	57	0.9887	441	0.9831	0.9806	± 0.01
3	1594	0.9858	132	0.9684	1904	0.9908	1103	0.9725	1727	0.9370	2471	0.9555	10	0.9829	1070	0.9744	1245	0.9822	797	0.9830	0.9733	± 0.01
4	1013	0.9691	270	0.9552	2174	0.9630	834	0.9646	1349	0.9356	1322	0.9501	153	0.9788	441	0.9741	483	0.9728	33	0.9821	0.9645	± 0.01
5	757	0.9607	146	0.9459	2144	0.9595	239	0.9469	1051	0.9270	1189	0.9353	2174	0.9781	2174	0.9728	957	0.9667	2726	0.9735	0.9566	± 0.01
6	2366	0.9535	63	0.9410	733	0.9587	2057	0.9397	1105	0.9188	1716	0.9318	56	0.9778	1579	0.9521	140	0.9657	164	0.9712	0.9510	± 0.01
7	306	0.9340	152	0.9380	1927	0.9563	1386	0.9316	238	0.9184	2061	0.9250	238	0.9716	1627	0.9486	950	0.9607	2398	0.9683	0.9452	± 0.01
8	169	0.9280	137	0.9374	613	0.9488	1373	0.9118	70	0.9181	45	0.9237	552	0.9605	2726	0.9392	940	0.9591	496	0.9549	0.9381	± 0.01
9	514	0.9276	79	0.9372	2061	0.9486	1631	0.8665	158	0.9115	13	0.9230	1904	0.9581	103	0.9371	54	0.9578	259	0.9549	0.9322	± 0.02
10	2022	0.9243	392	0.9350	328	0.9442	281	0.8545	415	0.9071	1116	0.9227	140	0.9521	2413	0.9354	1438	0.9538	2343	0.9451	0.9274	± 0.02

Table 3.4: CarRank in different set of Simulations

Simulation	1		2		3		4		5		6		7		8		9		10			
CarRank	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	Mean	Confidence Interval
1	764	1	210	1	1179	1	271	1	36	1	1773	1	2355	1	52	1	2276	1	39	1	1	0
2	1356	0.8182	178	0.6978	907	0.9566	295	0.9657	395	0.6212	2300	0.9756	1917	0.9084	2300	0.9728	444	0.9631	295	0.9836	0.9359	± 0.04
3	294	0.8177	298	0.6770	444	0.8568	595	0.8329	1902	0.5166	1932	0.9708	2276	0.8677	1932	0.8869	1581	0.9480	653	0.8924	0.8843	± 0.05
4	46	0.7831	424	0.6012	2276	0.8511	1179	0.8116	1926	0.4410	46	0.9071	423	0.8308	1695	0.8256	653	0.9440	1568	0.8714	0.8461	± 0.05
5	1454	0.7361	428	0.5701	682	0.8384	1926	0.7878	1147	0.4166	297	0.8311	2554	0.8158	1113	0.8163	1407	0.8858	739	0.8674	0.8144	± 0.05
6	169	0.7289	132	0.5420	2325	0.8285	2300	0.7709	46	0.4127	1804	0.8300	653	0.8114	918	0.8091	907	0.8745	1874	0.8659	0.8079	± 0.05
7	969	0.7287	444	0.5361	653	0.8255	2064	0.7642	1384	0.3991	2187	0.8205	399	0.8083	1720	0.8080	682	0.8549	1147	0.8411	0.7975	± 0.05
8	1384	0.7185	270	0.5149	2527	0.8210	2436	0.7613	895	0.3903	1454	0.8202	2028	0.7703	969	0.8075	399	0.8459	36	0.8270	0.7853	± 0.05
9	949	0.7174	39	0.4952	1581	0.7907	46	0.7375	2251	0.3559	699	0.8003	1310	0.7493	957	0.8028	1239	0.8314	816	0.8245	0.7713	± 0.05
10	1115	0.7157	169	0.4934	399	0.7532	1386	0.6734	477	0.3367	822	0.7918	2221	0.7473	682	0.7984	390	0.8259	1215	0.8152	0.7622	± 0.05

For better analysis of the performance of InfoRank ($\beta = 0$ and $\gamma = 0$ in Equation 3.6) and CarRank respectively in different simulation scenarios, we consider the following performance metrics in comparison with the state of the art centrality schemes (Degree, Closeness, Betweenness and Eigenvector centrality):

- Cumulative Satisfied Interests (CSI) for the top identified nodes by each scheme
- Comparison of top nodes identified by each scheme with their respective centrality scores
- Average aggregated throughput of the identified top ranked nodes by each scheme
- Cache hit rate for the top nodes by each scheme to evaluate CarRank along ICN in vehicular mobility scenarios

3.5.2 Results: Individual Vehicle Ranking

InfoRank and CarRank scores for the top 10 vehicles from ten simulation runs are shown in Table 3.3 and 3.4 respectively. The relative InfoRank and CarRank scores with respect to different vehicles ranking order in different set of simulation runs ar shown to where we can clearly see different set of nodes as top IFVs in each simulation. In each simulation, different vehicles are assigned randomly as providers for a location they already visited before. Similarly, different consumer vehicles are assigned randomly for vehicles planning to visit a location.

The purpose is to show the persistence of InfoRank as well as its efficiency as a ranking scheme. It is important to show that it actually"ranks" with different set of providers and consumers in different scenarios. We normalize the score by assigning unity to the most important vehicle followed by the remaining vehicles to clearly show the ranking order. We will use the same convention to interpret results in the later sections. For each rank, the average score lies within a confidence interval of 0.01 for a confidence level of 95%.

In the first simulation in Table 3.3, the vehicle 1115 is identified to have the top InfoRank score among the selected vehicles in the network as it responds more frequently to the relevant user interests throughout the simulation.

Similarly, for CarRank in Table 3.4, the vehicle 764 is identified to have the top CarRank score among all the vehicles in the network. One reason is that it responded more frequently to the incoming interests throughout the simulation for the respective associated locations. At the same time, its spatio-temporal availability with respect to the associated content and the neighborhood connectivity also contribute towards its score.

The total number of vehicles is about 2986, of which, 100 are randomly chosen as producer and another 100 are chosen as consumers. The most important vehicles relates to the top 5 best ranked vehicles (from the 2986 ones) in our case. It is to note that the number of "most important" vehicles depends on the application requirements, for example in the case of city-wide urban sensing, the municipality is limited by the budget and coverage constraints.

3.5.2.1 Cumulative Satisfied Interests

The term Cumulative Satisfied Interests refers to the total number of user interests satisfied during the simulation duration. CSI is the measure of the vehicle importance with respect to the user interests. Increase in the number of satisfied interest implies a high vehicle association to the particular content for the incoming interests. Figure 3.5 compares the CSI of the five most prominent vehicles identified by InfoRank with those identified using different centrality schemes in an average of ten set of simulations.

Typical ranking schemes only takes into account physical topology towards computing a node importance in the network, ignoring the vehicle relevance with respect to the user interests. Nevertheless, vehicles identified by InfoRank as the top information facilitators satisfied five times more user interests than those identified using other schemes in all simulations, thus heavily outperforming all existing centrality metrics in the dynamic nature of vehicular environment. Such huge difference is due to the consideration of user interest satisfaction as a key metric towards information importance as well as the vehicle importance in the network.

Figure 3.6 shows the CSI score of the top five nodes identified by CarRank and existing centrality schemes in an average of ten set of simulations.

Typical ranking schemes only takes into account physical topology towards computing a node importance in the network, ignoring the satisfied user interests. Nevertheless, the tops vehicles identified by CarRank satisfied more user interests than other schemes in all the ten set of simulations due to the consideration of user interest satisfaction for important information as a key factor towards vehicle centrality in the network.

For an extensive analysis, it is important to observe each vehicle behavior at different time instances by performing a time varying network analysis. Table 3.5 shows CSI for top nodes identified by each scheme at different instants. We consider a time varying behavior by taking network snapshots after each 10 minutes interval in the simulation. We observe some nodes with a higher CSI score than the node identified by CarRank. For instance, The node 969 at 50 minutes identified by Degree and Eigenvector centrality satisfied a total of 354646 interests. There is logical inference to this behavior as CarRank does not only considers the cumulative interest satisfied, but the vehicle neighborhood as well as the spatio-temporal availability as key components to decide its importance in the network. Table 5.3 also reflects the importance of 969 as it is ranked as the $7th$ top vehicle by CarRank as well. The node 764 turn out to be an outlier as it is identified as the the top node throughout the simulation. It is also identified as the top vehicle by Degree centrality and Eigenvector centrality. This is due to its better network connectivity and spatio-temporal availability to respond from the beginning till the end of the simulation.

3.5.2.2 Temporal behavior analysis of the top nodes

It is important to efficiently analyze the time varying behavior of our algorithm due to the dynamic VANET environment to address two questions:

- Is our vehicle centrality successful in identifying vehicles that can be relied for urban sensing over stable time duration?

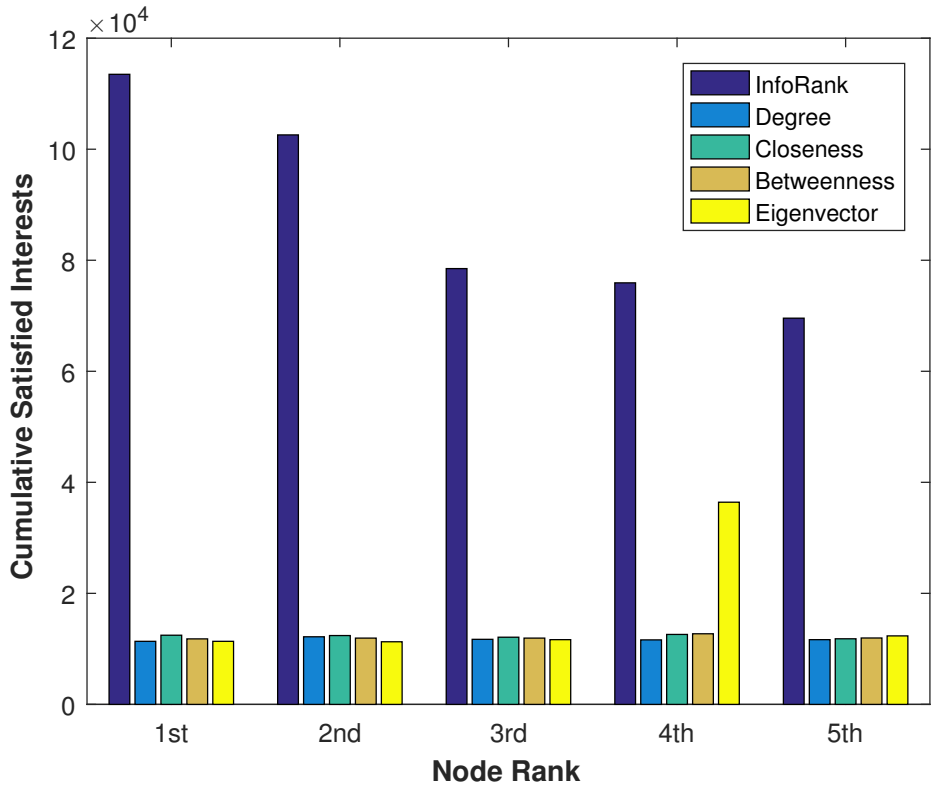


Figure 3.5: Comparison of the Cumulative Satisfied Interests by the top identified vehicles using InfoRank and existing centrality schemes

Table 3.5: Cumulative Satisfied Interests by top vehicles identified by each scheme

Simulation	**CarRank**		**Degree**		**Closeness**		**Betweenness**		**Eigenvector**	
CarRank	ID	CSI	ID	CSI	ID	CSI	ID	CSI	ID	CSI
10min	764	346196	764	346196	230	349	225	1115	764	346196
20min	764	346196	230	349	469	311	947	345500	470	372
30min	764	346196	59	345051	469	311	1939	353698	282	379
40min	764	346196	145	351023	427	392	947	345500	469	311
50min	764	346196	969	354646	1750	348481	464	536	969	354646
60min	764	346196	386	257	1233	343453	38	356	95	169

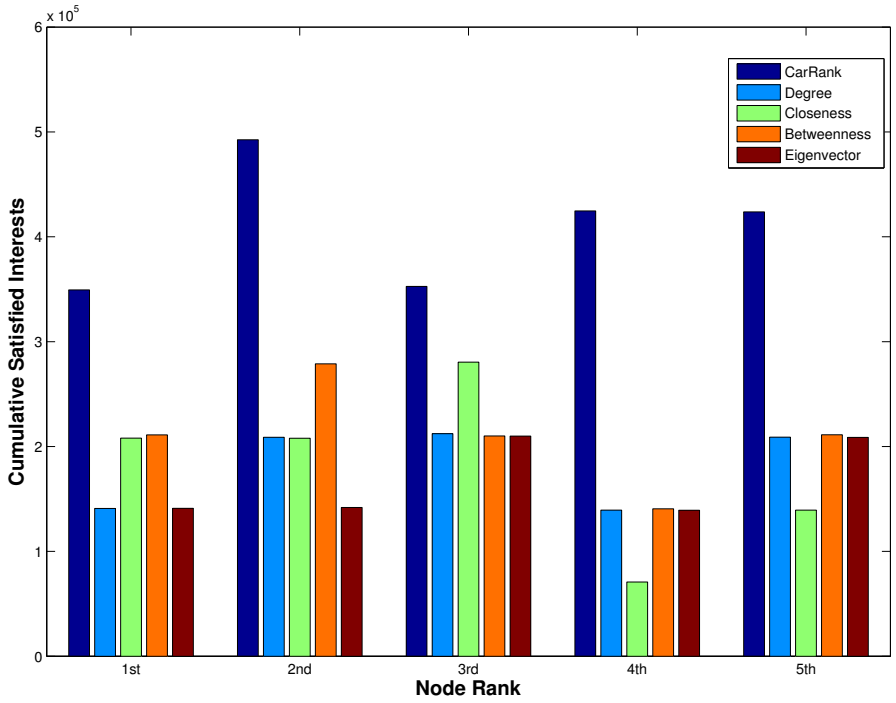
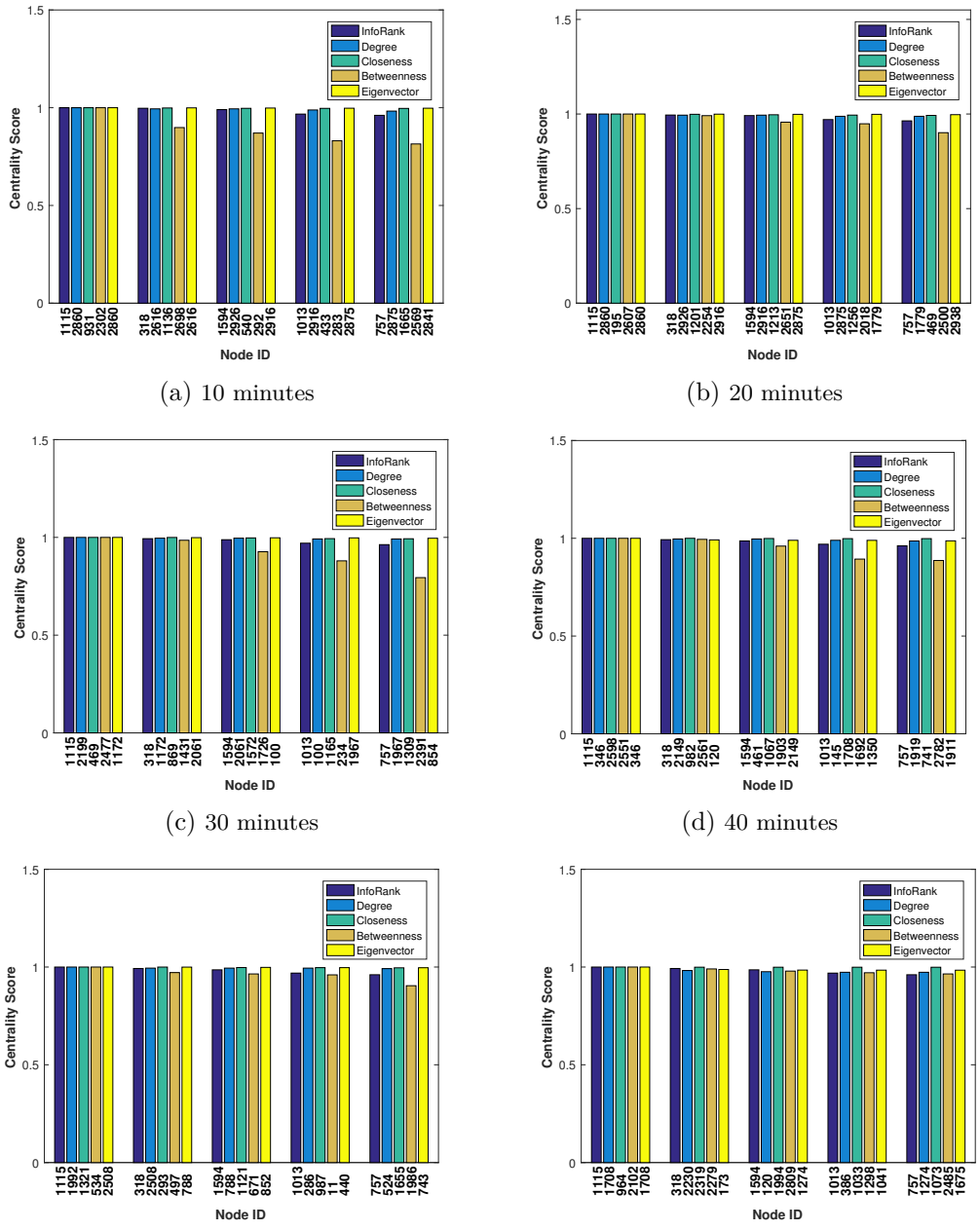


Figure 3.6: Cumulative Satisfied Interests by top identified vehicles using CarRank and existing schemes over an average of ten different simulation scenarios



(a) 10 minutes

(b) 20 minutes

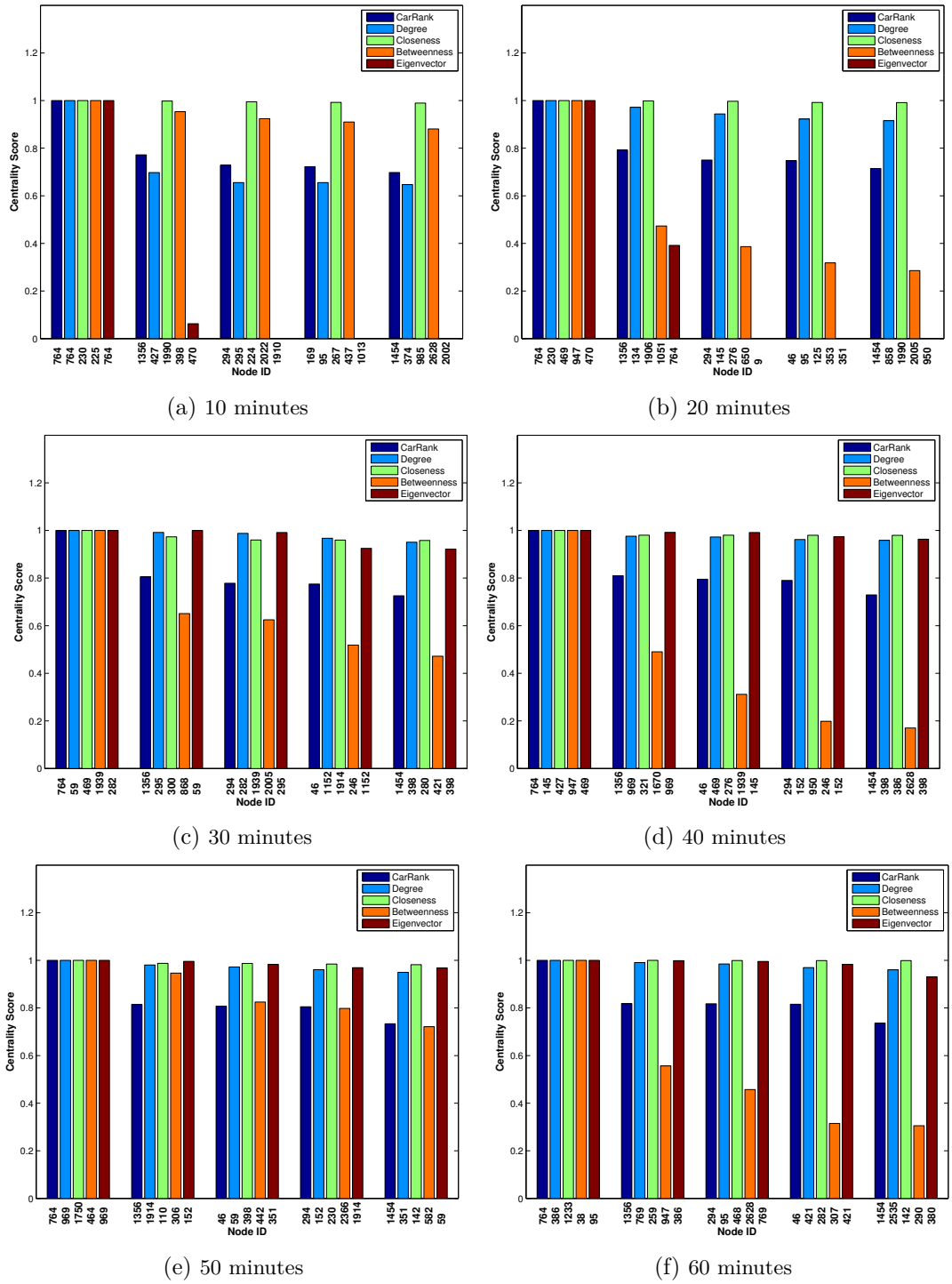
(c) 30 minutes

(d) 40 minutes

(e) 50 minutes

(f) 60 minutes

Figure 3.7: Temporal snapshots after each 10 minutes comparing top identified nodes by each schemes - InfoRank



(a) 10 minutes

(b) 20 minutes

(c) 30 minutes

(d) 40 minutes

(e) 50 minutes

(f) 60 minutes

Figure 3.8: Temporal Snapshots of comparing top nodes identified by all schemes (CarRank, Degree, Closeness, Betweenness, Eigenvector centrality)

- What happens if we use the state of the art schemes (metrics) to identify important vehicles? Why it is not feasible to use existing schemes?

The major purpose is to analyze the efficacy of InfoRank and CarRank as stable ranking algorithms in terms of its capability to rank vehicles. Secondly, we need to identify vehicle to be recruited for a relatively longer time period in order to be relied upon for different services (urban sensing in our case). Therefore we compare it with existing ranking systems with an analysis over time as a novel approach to study whether it is stable unlike the state of the art schemes.

Centrality score of the top five nodes identified by InfoRank and different centrality schemes are shown by the periodic network snapshots after each 10 minutes in Figure 3.7. We consider the top node identified by each scheme as a benchmark by assigning it a unity score. At the beginning, the vehicle 1115 is ranked as the top vehicles by InfoRank, though the other schemes underrated it. Similarly we observe that overall a stable set of vehicles identified by InfoRank, i.e. the vehicle ID 1115, 318, 1594, 1013 and 757, thus validating the feasibility of recruiting stable set of vehicles over longer periods.

After comparing the top nodes identified by each scheme, we observe that unlike InfoRank, other centrality schemes result in different set of top nodes at every snapshot. It is because such schemes only consider the instantaneous shortest paths towards ranking the vehicles at a particular time instant which require the complete topological information. However, such complete network information is not available to an individual vehicle in highly unstable VANETs. InfoRank ensures stable set of top ranked vehicles as it is clear from the time varying VANET analysis that they are not affected by the network dynamics. We are able to rank each vehicle considering relatively stable metrics, which is not the case for other schemes.

The aim of this analysis is to observe CarRank as an efficient ranking algorithm to identify vehicle to be recruited for a relatively longer time period in order to be relied upon for different LBS applications. The time varying behavior of the relative score of the top five nodes identified by all schemes are shown by periodic network snapshots after each 10 minutes interval in Figures 3.8a to 3.8f. We consider the top node identified by each scheme as benchmark by assigning it a unity score. We identified an outlier node 764, persistently ranked as top vehicles by CarRank, thought the other schemes underrated it. This is because we consider relatively stable factors such as the importance of associated information and the vehicle spatio-temporal availability besides the topological information. Vehicles also change places along the ranking order. For example, the node 294, ranked 3rd at 10, 20 and 30 minutes swap place with the node 46 around 40 minutes and finally retake the 3rd place in Figure 3.8f.

An interesting results was observed in Figure 3.8a and Figure 3.8b: Only one node yields a high Eigenvector centrality score followed by other nodes with a negligible Eigenvector centrality score. We investigate this effect and found that the principle eigenvalue yields the top nodes where the eignvector is shifted towards the principle component. Thus, providing one major central node. This shows that the famous Eigenvector centrality fails to assign significant score to a large fraction of nodes in a large network, while CarRank does not present such behavior. Other centrality schemes result in different set of top nodes at every snapshot. It is because such schemes only consider the instantaneous shortest paths towards ranking the vehicles at

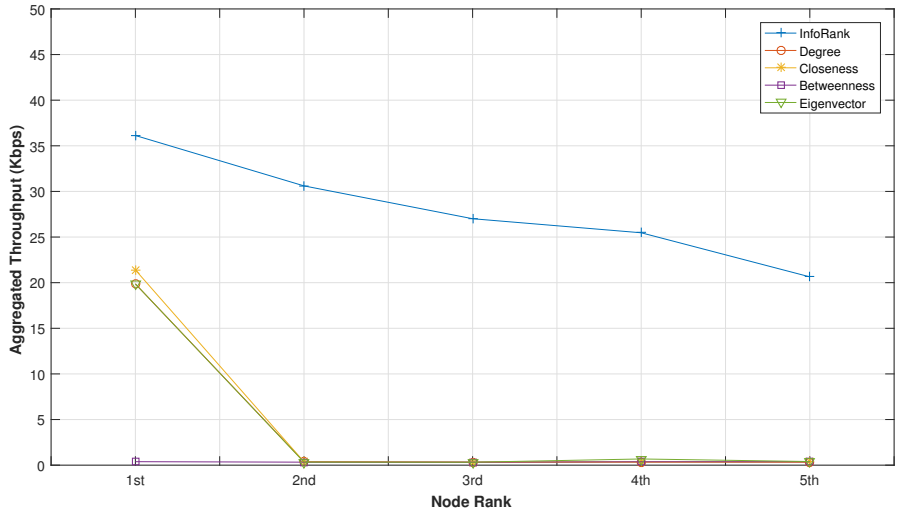


Figure 3.9: Comparison of the average aggregated per node throughput achieved by the top identified vehicles using each centrality scheme

a particular time instant which require the complete topological information. However, such complete network information is not available to an individual vehicle in highly unstable network. CarRank ensures more stable set of top vehicles as it is clear from the time varying network analysis that it is not affected by the network dynamics since we are able to rank each vehicle considering relatively stable metrics, which is not the case for other schemes.

3.5.2.3 Throughput

We also evaluate InfoRank and CarRank by analyzing the throughput achieved at the identified important nodes in the network. Figure 3.9 shows the aggregated per node throughput of the most prominent nodes identified by InfoRank compared to other centrality scheme. The average aggregated throughput (Kbps) is computed over the entire simulation duration for the ten set of simulations. The top nodes identified by InfoRank turns out to be an outlier for the amount of throughput compared to the nodes identified by other schemes. We observe in Figure 3.9 that the throughput of only the top most vehicles identified by Degree, Closeness and Eigenvector centrality is about half the amount of those identified by InfoRank. It is also clear that all other schemes results in a negligible amount of the throughput along the downward ranking order as the depend on the physical topology which is greatly affected by high vehicle mobility and intermittent connectivity. Moreover, we also notice that the nodes identified by betweenness centrality yields no substantial throughput along the top ranked order. Another reason for this behavior is the constant user interest satisfaction ratio from the best vehicles identified using InfoRank during the simulation duration.

For CarRank, Figure 3.10 shows the aggregated per node throughput of the top nodes identified by each scheme. The average aggregated throughput (Kbps) is com-

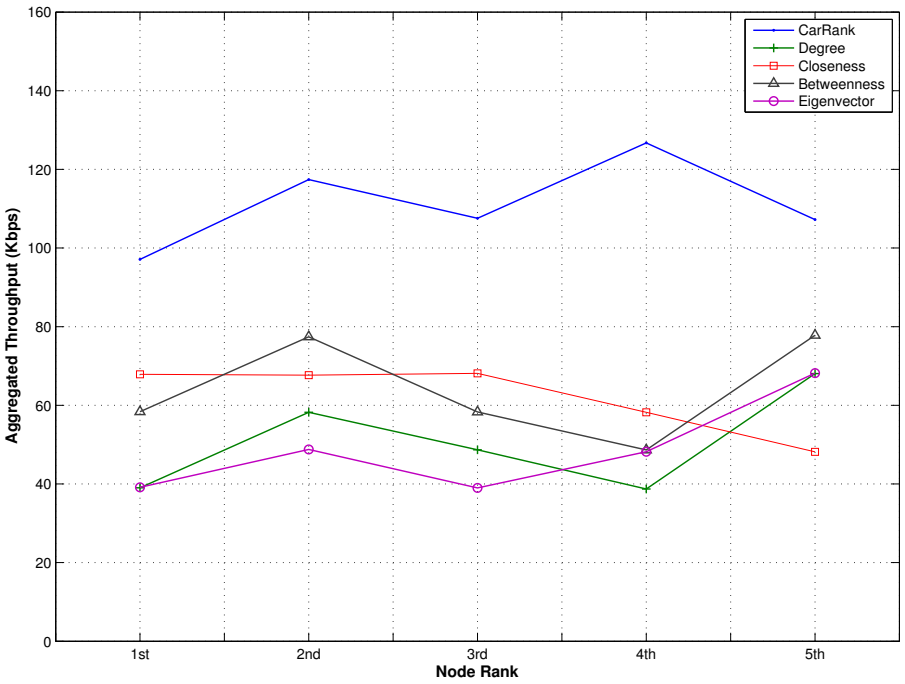


Figure 3.10: Average aggregated throughput by the top identified nodes using each scheme in ten simulations

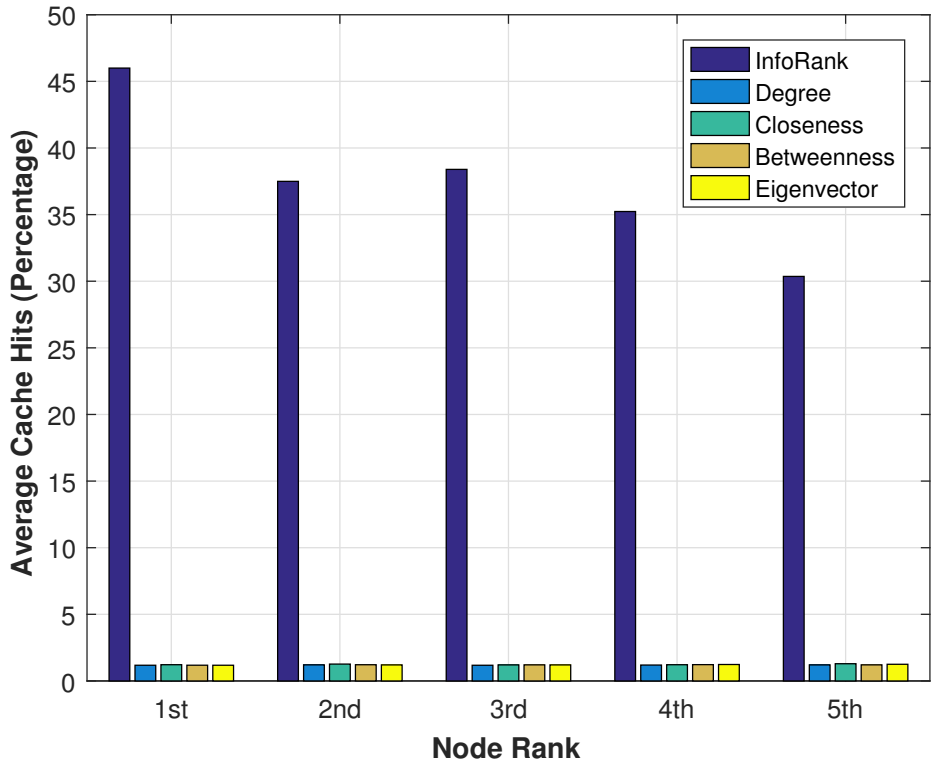


Figure 3.11: Comparison of the average cumulative cache hit rate at the top identified vehicles using each centrality scheme

puted over the entire simulation duration for ten set of simulations. The top nodes identified by CarRank yields more throughput compared to other schemes. We also observe that the throughput of the fourth node is relatively higher, thus inferring a variation between different ranks. Similar variations are seen for Degree, Betweenness and Eigenvector centrality. However, Closeness centrality follow a decrease along the vehicles ranking order. CarRank outperformed all schemes as it incorporates additional factors towards vehicle importance computation such as the information importance and the spatio-temporal availability, while other schemes rely only on topological measures (node degree or shortest paths) towards vehicle importance computation.

3.5.2.4 ICN Evaluation - In-Network Caching

We evaluate the ICN built-in feature of In-Network caching at the intermediate nodes by computing the cache hit rate at the top nodes identified by each scheme as shown in Figure 3.11. A second successful response by a node for the same content is considered a cache hit. The cumulative cache hit rate is computed for the entire simulation duration for the ten set of simulations using both InfoRank and CarRank.

The most important nodes identified by InfoRank in Figure 3.11 yield a higher hit rate of around 45% compared all the other schemes which resulted in negligible amount of around 2% on average cache hits during the simulation. This is because

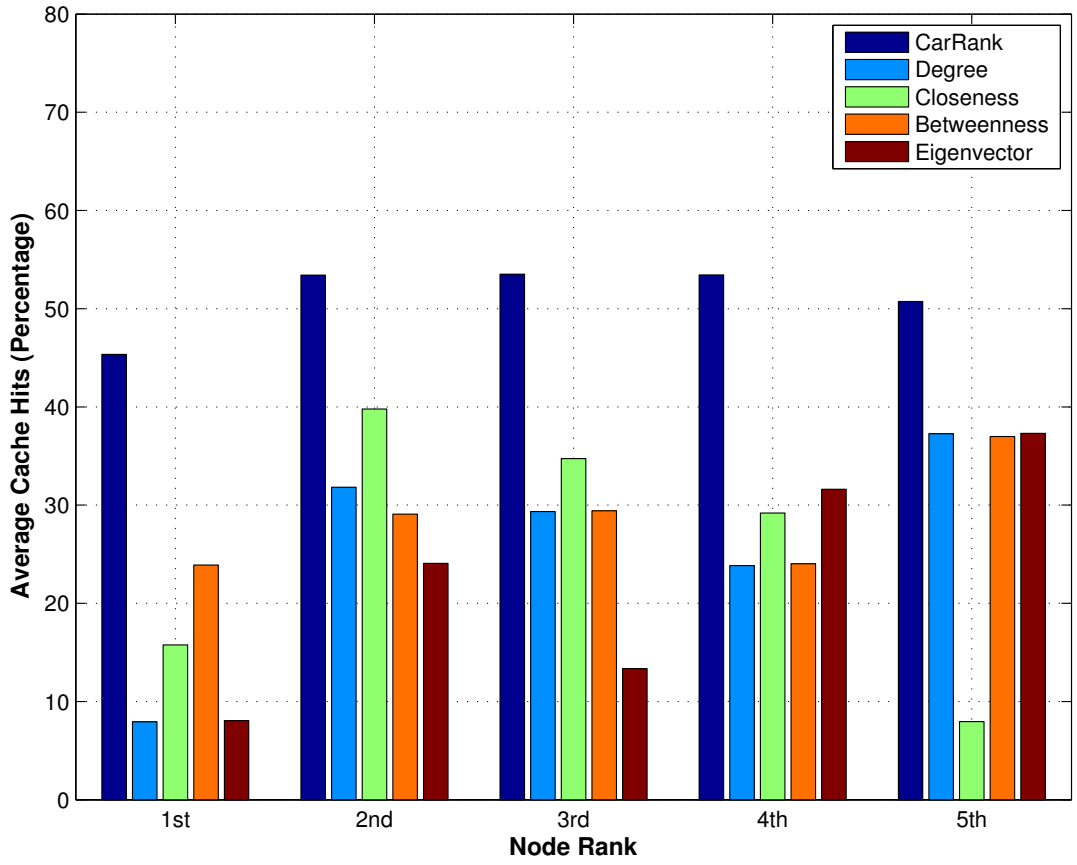


Figure 3.12: Average cumulative cache hit rate by the top identified nodes using each scheme in ten simulations

InfoRank considers information importance as a key factor, thus, the vehicle containing important information responds and subsequently cache more frequently compared to other vehicles. This proves that In-Network caching offered by ICN in InfoRank implementation overcomes the mobility and intermittent connectivity constraints in VANETs for efficient content access.

In the similar regard, Figure 3.12 shows the cache hit rate for top nodes identified by CarRank and state of the art centrality schemes. The top nodes identified by CarRank yield a higher hit rate than all the other schemes in all the ten simulations. This is because CarRank considers content popularity as a key factor, thus, the vehicle containing important information responds and subsequently cache more frequently compared to other vehicles. Moreover, the top node identified by CarRank cached more important content due to their better neighborhood and spatio-temporal availability. This proves that In-Network caching offered by ICN in CarRank implementation overcomes the mobility and intermittent connectivity constraints in vehicular network for efficient content access.

3.5.3 Discussion

Finally, we are able to comment on the question we posed in the beginning of this section: *How well can we identify the top vehicles capable to perform large scale urban data collection?* From the simulation results using scalable network scenario, it is evident that a relatively stable set of vehicles are identified by our novel centrality scheme compared to the other schemes. It is also clear that our novel centrality metrics identify nodes which satisfied more user interests with higher aggregated per node throughput and more cache hit rate compared to the other schemes. Thus, the overall comparative analysis of our centrality with different network ranking schemes from the literature suggests it as an efficient vehicle ranking algorithm compatible with the mobility and intermittent connectivity challenge in vehicular networks.

3.6 Conclusions

This chapter address the first step in efficient large scale urban data collection to find eligible candidates. It is challenging to identify the appropriate vehicles at the right time and right place for efficient urban sensing. There exist network-centric centrality schemes to identify important node in a network, however such centrality schemes cannot be applied to the dynamic vehicular environment. Therefore, we modeled the vehicle identification as a node identification problem in a network where novel centrality schemes are proposed.

To do so, we first identified socially important vehicles using two innovative vehicle centrality schemes, “InfoRank” and “CarRank”, allowing smart vehicles to autonomously rank themselves based on user relative importance. The vehicle considers the importance of location-aware information, its neighborhood topology along its spatio-temporal availability to find its importance in the network. Results by comparing with state of the art centrality schemes revealed that our novel centrality metrics are best suited to efficiently identify the best candidates vehicles compared to other schemes with a success rate of 60% compared to the 20% of benchmark centrality

schemes. In the next chapter, we go further in the data collection process by presenting the next phase of the proposed VISIT scheme. More precisely, we formulate an optimization model as well as the related heuristic algorithms for the efficient selection of best candidate vehicles for data collection under limited budget, coverage and redundancy constraints.

Chapter 4

VISIT for Data Collection: An Optimal Vehicles Selection Scheme

4.1 Introduction

Once the eligible candidates are identified, there is a challenge to select the best candidate vehicles to achieve urban data collection with a desired coverage. Moreover, assuming the large number of vehicles on urban roads it is difficult to recruit all vehicles to perform data collection under a given budget limitation, requiring an optimal selection/recruitment of an appropriate set of vehicles.

In literature, there exist algorithms for recruiting vehicles for urban sensing, however, they are unable to provide a scalable spatio-temporal coverage using vehicles, thus requiring the use of high cost infrastructure support. In order to solve the vehicle selection problem in polynomial time, we propose ROVERS as Recruitment of Optimal Vehicles for Efficient Road Sensing as a set of algorithms to optimize the selection of the best ranked set of vehicles identified by CarRank and InfoRank for urban data collection. Unlike existing schemes, ROVERS better solves the problem by maintaining city-wide coverage with the minimum cost where the selected set of vehicles in our simulation satisfied 60% of the user interest compared to less than 10% for the existing algorithms.

The chapter organization is the following. The next section talks about the context and motivation for the selection of best set of eligible candidates to perform large scale urban data collection. Section 4.3 describes the formulation of the optimization problem for our approach, ROVERS, along the algorithms to select the best candidate vehicles for urban data collection. Performance evaluation and results to validate the proposed optimal vehicle selection scheme are explained in Section 4.4 and finally Section 4.5 provides concluding remarks on ROVERS.

4.2 Context and Motivation

Given an urban environment with thousands of vehicles on the road, with their high processing, storage and communication capabilities. Once we are able to identify eligible vehicles using the novel metrics defined in the previous chapter, the next step for smart city application provider is to recruit a set of vehicles to perform urban sensing and vicinity monitoring in order to improve citizen lifestyle while offering new services. More precisely, this chapter targets the following question?

- Which set of vehicles can be selected (out of the good candidate vehicles) as appropriate candidates to optimally achieve large scale urban data collection under a given budget by a smart city administration or an application provider and a set of city-wide spatio-temporal coverage requirements with minimum redundancy ?

To address this, we believe that of all the vehicles, only a set of appropriate vehicles can be selected based on their daily commute while considering the popularity of its frequently visited neighborhoods. Therefore, in the following sections we devise ROVERS as an optimization model along with the associated heuristic algorithms for the selection/recruitment of a subset of such eligible vehicles for the above described smart-city usage under an associated budget and coverage requirements.

The major contribution to this chapter can thus be summarized as follows:

- ROVERS as a model comprising optimal recruitment algorithms to select the minimum set of important vehicles identified by InfoRank and CarRank for urban data collection while maintaining acceptable city-wide coverage with less redundant data at the minimum cost.

The objective here is to show that the proposed algorithms are well suited to help in the efficient selection of the best vehicles for urban data collection.

4.3 Recruitment of Optimal Vehicles for Efficient Road Sensing (ROVERS)

As mentioned earlier, our objective here is to recruit the best set of important vehicles identified by CarRank for different LBS applications while maintaining a certain level of city-wide coverage with the minimum cost. We assume a monetary reward is paid as an incentive to each vehicle proportional to its contributions relatively to the spatio-temporal coverage based on their centrality and the fuel cost. Therefore, we propose an optimal vehicle recruitment scheme where the objective is to minimize the recruitment cost for the selected set of vehicles with an acceptable level of coverage for different parts of the city. We also define the centrality constraint for an individual vehicle as well as the overall centrality of the selected set of vehicles to be higher than a certain threshold. Moreover, we consider a maximum threshold amount of budget as payoff possible to an individual vehicle to ensure fairness between vehicles. Similarly, the overall budget assigned for all vehicles should not exceed a certain threshold amount dedicated by the municipality or application provider.

Table 4.1: VISIT II - List of Notations

Notation	Description
$\mathbb{V}$	Set of vehicles
$\mathbb{X}$	Set of locations/regions
$\mathbb{E}^x/\mathbb{E}^v$	Set of edges between locations/vehicles
$\mathbb{E}$	Set of edges between vehicles and locations
$\bar{t}_k$	Time-slot k for centrality computation
t_k/t_{k+1}	Current time instant/next time instant
X_v	Set of locations associated to vehicle v
$V \subset \mathbb{V}$	Selected set of IFVs
$A^{\text{th}}_{max}/A^{\text{th}}_{min}$	maximum/minimum allowed coverage threshold
A_x	Individual location coverage vector for x
C^{th}	Pre-defined vehicle centrality threshold
R^{th}	Overall vehicle centrality benchmark
R_V	Overall vehicle centrality score
B_V	Budget assigned for all vehicles
B_v	Individual vehicle payoff
B^{th}	Individual vehicle maximum dedicated budget
$\mathbb{B}$	Overall maximum dedicated budget
U	Temporary set of removed IFVs
W	Set of vehicles not selected as IFVs

4.3.1 Problem Formulation

We formulate the vehicle selection as a discrete optimization problem where the objective function F :$\mathcal{V} \rightarrow \mathbb{R}_+$ finds the set of important vehicles, $\mathcal{V} = \{V : V \subset \mathbb{V}\}$, that minimize the overall recruitment cost. The necessary constraints are defined as:

4.3.1.1 Spatio-temporal Coverage

We define the area coverage vector $A_x(\overline{t_k})$ as the number of vehicles covering the area $x \in \mathbb{X}$ at time-slot $\bar{t}_k$. The coverage requirement specifies the number of vehicles needed in location x at a particular time-slot depending on the location importance. We also consider redundancy by limiting the number of vehicles available to sense in a particular location. The spatio-temporal coverage constraint is divided into two parts, (i) to avoid redundant data collection by limiting the number of vehicles covering an area (ii) to improve the quality of service by designating more vehicles to cover a specific area:

- Acceptable non-redundant coverage: A location x should be covered by at least A^{th}_{min} and at most A^{th}_{max} number of vehicles at time-slot $\overline{t_k}$, where A^{th}_{min} specifies the lower bound for the spatio-temporal coverage and A^{th}_{max} provide a bound on the maximum number vehicles. For instance, $A^{\text{th}}_{min}(\overline{t_k}) = 1$ as the lower bound requiring at least one vehicle covering each location in the time-slot $\overline{t_k}$.

- Desired coverage: Each location x should be covered by A_x^{th} number of vehicles at time-slot $\overline{t_k}$, where A_x^{th} is specified as the desired number of vehicles required in location x depending on its popularity.

4.3.1.2 Vehicle Centrality

To ensure the selection of the important information hubs, we define a two-tier vehicle centrality constraint as:

- Individual centrality: Each of the selected vehicles should have a centrality C_v higher than a given individual threshold centrality C^{th}.
- Overall centrality: The overall centrality score R_V of all the selected vehicles should be higher than a given centrality benchmark R^{th}, where $R_V = \sum_{v \in V} C_v$.

4.3.1.3 Dedicated Budget

The maximum payable amount for the selected vehicles is limited by an upper bound depending on the dedicated budget. Similar to the centrality constraint, we define a two-tier budget constraint as:

- Individual budget: In order to achieve fairness between vehicles, The payoff B_v for an individual vehicle should not exceed the maximum budget B^{th} dedicated for a single vehicle.
- Overall budget: The overall assigned budget B_V should also not exceed the budget $\mathbb{B}$ dedicated for all vehicles, where $B_V = \sum_{v \in V} B_v$.

The optimization problem needs to take into account the two coverage possibilities, therefore we formulate OPT_1 for the acceptable spatio-temporal coverage as follows:

$\underline{\text{OPT}_1}$: Acceptable non-redundant coverage

$$
\begin{aligned}
\underset{V}{\text{minimize}} \quad & F(V) \\
\text{subject to} \quad & A_{min}^{\text{th}} \geq \min_{x \in \mathbb{X}} A_x(\bar{t}_k) \geq A_{max}^{\text{th}} && (\text{C}_1) \\
& \min_{v \in V} C_v \geq C^{\text{th}} && (\text{C}_2) \\
& R_V \geq R^{\text{th}} && (\text{C}_3) \\
& \max_{v \in V} B_v \leq B^{\text{th}} && (\text{C}_4) \\
& B_V \leq \mathbb{B} && (\text{C}_5)
\end{aligned}
$$

Similarly, we formulate OPT_2 as the optimization problem considering the possibility to achieve the desired level of coverage for each location as:

$\underline{\text{OPT}_2}$: Desired coverage

$$
\begin{aligned}
\underset{V}{\text{minimize}} \quad & F(V) \\
\text{subject to} \quad & A_x(\bar{t}_k) \geq A_x^{\text{th}}(\bar{t}_k),\ \forall\, x \in \mathbb{X}, \forall\, \bar{t}_k && (\text{C}_6)
\end{aligned}
$$

and (C_2), (C_3), (C_4), (C_5)

The constraint (C_1) in OPT_1 concerns the number of vehicles able to cover the location x with a minimum redundancy. The constraint (C_2) and (C_3) address the vehicle centrality constraint, where (C_4) and (C_5) deals with the budget constraints. Similarly, the constraint (C_6) in OPT_2 address the desired spatio-temporal coverage requirements.

1: **INPUT:** Vehicles set $\mathbb{V}$
2: **OUTPUT:** Selected vehicles set $V \subset \mathbb{V}$
3: Initialize $W = \phi$,
4: **for** each vehicle $v \in \mathbb{V}$ **do**
5: **while** $(R_V \geq R^{th})$ and $(B_V \leq \mathbb{B})$ and $(A^{\text{th}}_{min} \geq A_x(\bar{t}) \geq A^{th}_{max}), \forall \bar{t}, \forall x \in X$ **do**
6: $A_x(\bar{t}) \leftarrow A_x(\bar{t}) - A_{x_v}(\bar{t})$
7: $R_V = R_V - C_v$
8: $B_V = B_V + B_v$
9: $W \leftarrow W \cup v$
10: **end while**
11: **end for**
return V

Algorithm 3: Optimized Vehicles Removal

4.3.2 Algorithm: Optimized Set of Vehicles selection

The selection algorithm for the best ranked vehicles in the network are summarized in Algorithm 3 and 4. We assume vehicles $\mathbb{V}$ are subscribed to provide different LBS services in a smart city. In Algorithm 3, we start by greedy removal of vehicles from the set of all vehicles $\mathbb{V}$ to the set W until sufficient vehicles remain as the subset $V \subset \mathbb{V}$. Line 5 ensures the overall centrality, budget, coverage and redundancy constraints are respected. Line 6-9 removes the vehicle, subtract its centrality from overall centrality score as well as remove its cost from the overall cost. The removed vehicles are then added to the set W until a local minima is obtained.

In order to search for another local minima or the global minima, vehicles from the removed set W are iteratively added back step-wise with an increase in the step size at each iteration as shown in Algorithm 4. The removed vehicles from the set W are revisited to be added back to the selected vehicles set V using the temporary revisiting vehicle set U. The set U contains randomly selected vehicles from the set of removed vehicles W where the number of selected vehicles are specified by the step size. The condition in line 7 ensures vehicles exists in the removed set as well as defining the upper bound on the step size for the number of vehicles considered to add back at each iteration. The respective vehicle coverage, centrality and cost are updated for the set of vehicles added back in line 9-11. The algorithm continue to optimize until the required set of vehicles V which minimize the objective function is found.

For the desired coverage in OPT 2, the maximum bound A^{th}_{max} for the coverage requirements is removed and the minimum bound A^{th}_{min} is replaced by $A^{\text{th}}_x(\bar{t}_k)$ in line

INPUT: Removed vehicles set $W \subset \mathbb{V}$, Selected vehicles set V
OUTPUT: Updated removed vehicles set $W \subset \mathbb{V}$, Selected vehicles set $V \subset \mathbb{V}$
Initialize step $= 0$, vehicles revisiting set $U = rand(u, step)$
step $=$ step $+ 1$
while $(R_V \geq R^{th})$ and $(B_V \leq \mathbb{B})$ and
$(A^{\text{th}}_{min} \geq A_x(\bar{t}) \geq A^{th}_{max})$, $\forall \bar{t}, \forall x \in X$ **do**
  for each vehicle $u \in U$ **do**
    if $(|W| \geq 0)$ and (step $\leq |V|$) and $(C_v \geq C^{th})$ and $(B_v \leq B^{th})$ **then**
      $V \leftarrow V \cup u$
      $A_x(\bar{t}) \leftarrow A_x(\bar{t}) + \sum_{i=1}^{step} A_{x_u}(\bar{t})$
      $R_V = R_V + \sum_{i=1}^{step} C_u$
      $B_V = B_V - \sum_{i=1}^{step} B_u$
      $W \leftarrow W - U$
    end if
  end for
end while
return V, W

Algorithm 4: Optimized Vehicles Revisiting

5 for both Algorithms 3 and 4.

4.4 Performance Evaluation

In this section we discuss the performance evaluation using the urban sensing simulation scenario defined in the previous chapter along the results obtained by ROVERS after applying the optimization model using Algorithms 3 and 4. The algorithms finds the best set of IFVs for urban sensing under the constraints associated to the spatio-temporal coverage, vehicle centrality and the available budget. The city map is divided into zones/areas as voronoi tessellation where vehicles in proximity of each other by average values of their coordinates are co-located within the same voronoi region at the current time-slots. For each vehicle, the vehicle centrality is computed at regular instants using InfoRank and CarRank respectively followed by the optimization Algorithms 3 and 4 to select the set of best ranked vehicles for urban data collection under the given coverage, redundancy and budget constraints. We consider the following performance metrics in comparison with the state of the art centrality schemes for the selected set of vehicles:

- Cumulative Satisfied Interests (CSI) by the selected set of best vehicles
- Average aggregated throughput achieved by the selected set of best vehicles

4.4.1 Results: Best set of vehicles selection

4.4.1.1 Cumulative Satisfied Interests

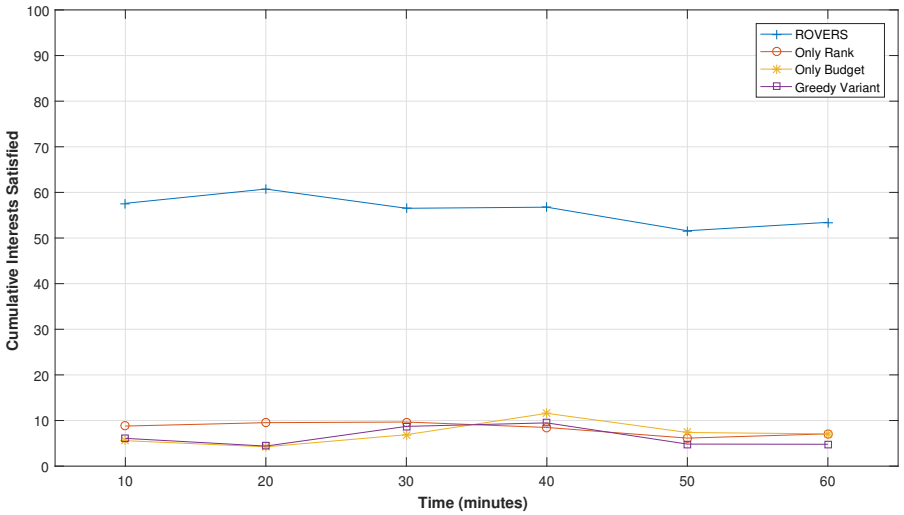


Figure 4.1: Temporal analysis of the average CSI by the optimized set of IFVs selected using each algorithm

In order to validate the optimized set of selected IFVs, we find the CSI by the vehicles identified as the best set of IFVs using our proposed IFV optimization stated in OPT_1 and OPT_2. We compare the CSI of the proposed optimizations algorithm with (1) the NP-Hard budgeted maximum coverage optimization problem [22], where the objective function targets to select the set of vehicles with the minimum cost (only low cost), while achieving the maximum coverage, (2) set of IFVs selected based on high vehicle centrality score as the objective function (only best ranked) and (3) the greedy approach of our optimization algorithm to select the best set of IFVs based on their high centrality score with the minimum cost (best ranked and low cost).

We also compare the best set of IFVs identified by ROVERS (InfoRank and Carrank) by applying our optimization algorithms using the state of the art centrality schemes. Figure 4.1 depict the ratio of the cumulative interests satisfied to the total received interests by the set of best ranked IFVs at different time instants. From the temporal analysis, we observe that the selected set of IFVs by ROVERS using our approach satisfied 60% user interests compared to an average of below 10% by other selection algorithms at different time-instants. Additionally, comparison results from applying our optimized IFVs selection algorithms along ROVERS and existing centrality scheme are shown in Figure 4.2.

Interestingly, compared to the 60% user interests satisfied by the set of nodes identified using InfoRank and CarRank, none of the other centrality schemes succeeded in identifying a set of IFVs which satisfied more than 10% of the received user interest. We also observe that the set of IFVs identified by the famous eigenvector centrality failed to satisfy any interest generated in the network. ROVERS outperforms existing

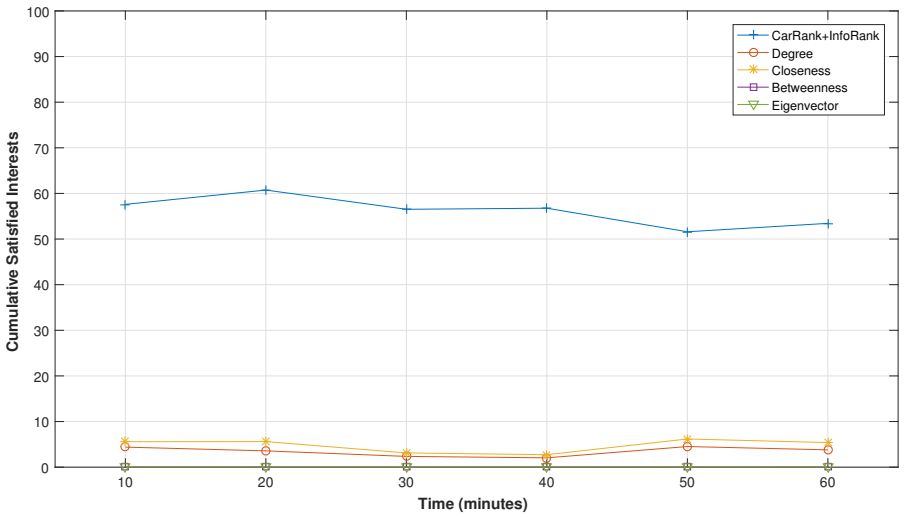


Figure 4.2: Temporal analysis of the average CSI by the optimized set of IFVs selected using each centrality scheme

metrics as the optimal set of best ranked IFVs by ROVERS are co-located at important locations with respect to the user interests. Similarly, Algorithm 3 and 4 ensures that the best ranked IFVs by our scheme provide the desired coverage with the least possible cost. On the other hand, other centrality schemes ignore metrics associated to user interest and spatio-temporal coverage for the identifications of important nodes.

4.4.1.2 Throughput

In addition to individual per node throughput for the top ranked IFVs, we analyze the aggregated throughput of the set of best ranked IFVs selected for urban sensing. Figure 4.3 compares the throughput achieved by the set of best ranked IFVs identified using our optimization Algorithms 3 and 4 with other approaches. The efficiency of ROVERS using our proposed IFV selection algorithms is evident since we observe that the selected set of best IFVs achieved around three times more throughput compared to those identified by other approaches. It is due to high relevance to the user interests and the availability to satisfy interests (provide coverage) at different geographical locations at lower cost.

Similarly, the aggregated throughput of the selected set of vehicles by applying our optimized IFV selection algorithm using state of the art centrality schemes is shown in Figure 4.4. We observe that ROVERS results in more aggregated throughput (maximum 83 Kbps at around 10 minutes) compared to any other centrality scheme, where none of the existing metric yields a set of vehicles exceeding a maximum throughput of 20 Kbps. Thus, it clearly indicates that Algorithms 3 and 4 with ROVERS for selecting the appropriate set of IFVs outperforms the state of the art centrality schemes.

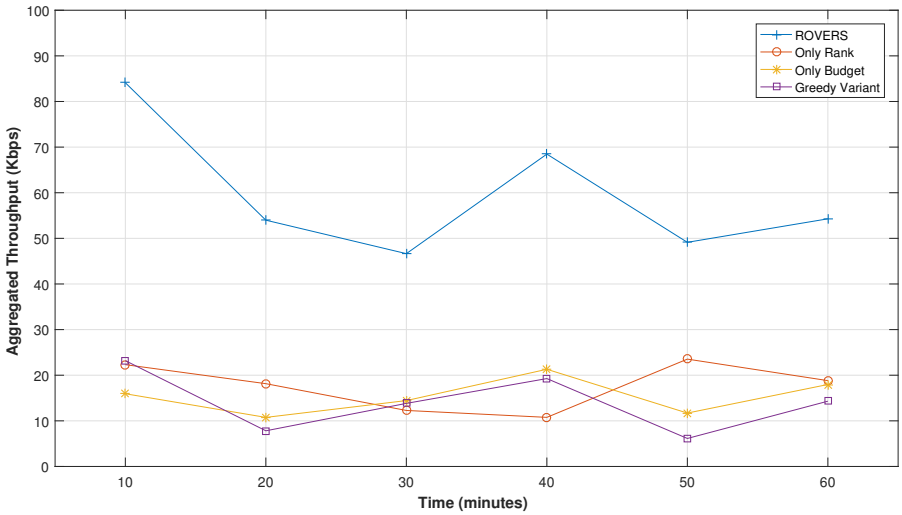


Figure 4.3: Temporal analysis of the throughput achieved by the optimized set of IFVs selected using each algorithm

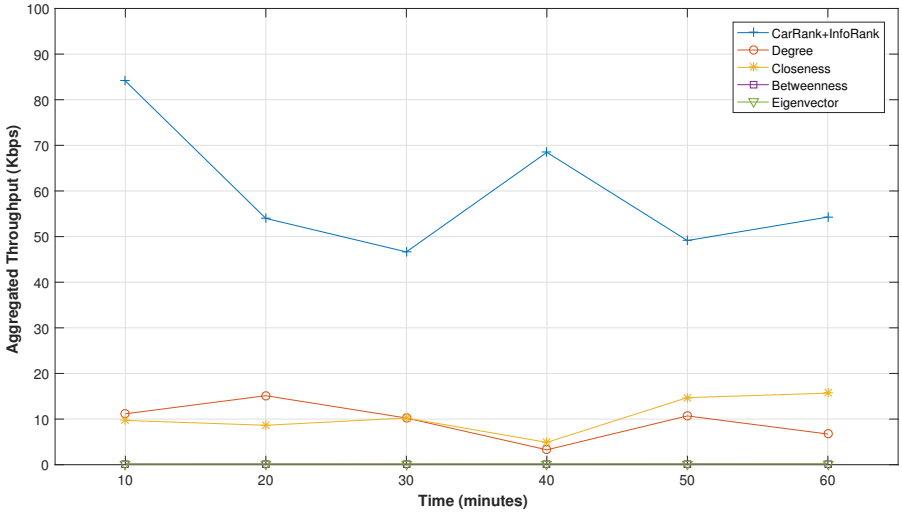


Figure 4.4: Temporal analysis of the throughput achieved by the optimized set of IFVs selected using each centrality scheme

4.5 Conclusions

Once we are able to identify eligible candidate vehicles, the selection of appropriate vehicles for the collection, storage and distribution of large amount of data from the fleet of vehicles on urban streets is a challenging task. There exists selection algorithms for maximum coverage with limited budget, however, none of them provided a scalable solution for large coverage with minimum cost.

In order to solve the problem in polynomial time, we modeled the vehicle selection as an optimization problem.This chapter proposed an optimum selection system ROVERS to efficiently recruit the set of best ranked vehicles for urban data collection while achieving a desired coverage and avoiding redundancy within a limited budget. Results by comparing with state of the art centrality schemes revealed that our centrality system is best suited to efficiently select best suitable vehicles which satisfied 65% user interests compared to the average of 15% user satisfied interests by vehicles selected using other algorithms. The identified ROVERS collect data from different urban neighborhoods. Now there is a need to store the gathered data at different city-wide information hubs available for the users in an urban environments. Therefore, the next chapter focuses on a novel concept of the identification and recruitment of vehicles as distributed content caches in an urban environment.

Chapter 5

SAVING as Data Storage: A Social Choice Game for Urban Cache Recruitment

5.1 Introduction

The growth in the number of mobile devices today result in an increasing demand for large amount of rich multimedia content to support different applications. However, the connection-centric nature of the current cellular networks are struggling with such increasing content demand from mobile users in an urban environment [29] [14]. In the previous chapter we presented the idea of using vehicles for data collection in an urban environment. At the same time such smart vehicles today can be used as content caches with their high computing, caching and communication capabilities to supplement the infrastructure networks to store the gathered data. Therefore, we leverage the recent shift towards content centric networking (CCN) by recruiting vehicles to cache content closer to the mobile user to avoid bandwidth and cost.

This chapter aims to identify and select the appropriate set of vehicles for urban content caching under spatio-temporal coverage and budget requirements by a service provider. To do so, we first model the vehicle identification as an autonomous node identification problem in a network. Such identification is typically achieved by well known node centrality schemes. However, such scheme cannot be used in the dynamic nature of vehicular network. We proposed InfoRank and CarRank as new centrality in Chapter 3 to identify vehicles for urban data collection, however both address the vehicles identification on a local spatio-temporal scope needed for short range and immediate neighborhood information collection and dissemination. However for a city wide content access, we need vehicles to be reachable globally for a city wide content caching, thus requiring metrics with global scope. Therefore, using concept from complex theory, we present GRank allowing a vehicle to classify its eligibility for distributed content caching in dynamic urban environment.

The second issue is that the selection of appropriate set of vehicles is an NP-complete problem under budget and coverage constraints. Therefore, in order to solve it in polynomial time, we formulate a social welfare game to optimally select the best

set of appropriate vehicles for a desired coverage within budget limitations.

The third issue arise with the existence of nodes with rational behavior opting to cover desired location to gain maximum reward. To cater this, we present a game theoretic based solution to ensure incentive-compatible fairness. A social welfare optimization is achieved which exploit the vehicle rationality towards satisfying a social norm where the Nash Equilibrium optimizes for the social welfare. By doing so, we are successful in identifying and selecting set of vehicles that achieve an average of 45% cache hit rate compared to 10% by using benchmark algorithms.

The chapter is organized as follows. The next section discusses the context and motivation for data storage at important information hubs. Section 5.3 present the novel metric GRank to identify such distributed urban information hubs followed by the social welfare game for efficient recruitment of such vehicles for city-wide storage in Section 5.4. Performance evaluation and results to validate our proposals are discussed in Section 5.5 and the Section 5.6 concludes the chapter.

5.2 Context and Motivation

The advancement of mobile devices such as smartphones, tablets and other portable gadgets bring along an explosive growth towards the utilization of high bandwidth consuming applications for such devices. It is now challenging for current network infrastructure to facilitate content availability for mobile users having multiple devices while offering attractive tariff plans supporting unlimited bandwidth. Despite the revolution towards 5G networks, the connection-centric nature of current mobile networks makes it difficult to cope with the tremendous content demand from mobile devices. The recently proposed Content-Centric Networking (CCN) [2] paradigm cater the issue by decoupling the content provider-consumer and support in-network caching at intermediate nodes to improve content availability with minimum delays. However, the processing, storage and power limitations of mobile devices greatly impede content caching at such nodes.

Therefore, we advocate to exploit vehicles as robust content caches with their relatively high processing, storage and communication capabilities where the named-data networking caters mobility and intermittent connectivity [40]. A mobile service provider can recruit vehicles as distributed content caches taking into account their mobility and availability within the urban environment. However, recruiting appropriate vehicles of the thousands of vehicles on urban road is a difficult task. First, how to classify a particular vehicle as an eligible candidate for content caching at different urban locations? Second, Once eligible candidate vehicles are identified, how to optimally select the best set of vehicles for city-wide content caching with desired coverage requirements under a given budget? Third, how to ensure fairness among recruited vehicles compensating for their coverage cost while dealing with vehicle individual-rationality?

Since InfoRank and CarRank were proposed for data collection requiring a local scope for information diffusion/collection. In order to find city-wide information hubs to cache content and to be able to reachable from different locations, we need metrics that go beyond local neighborhood to a more global scope. Therefore, this chapter address the above questions by first presenting an innovative vehicle ranking algo-

rithm, GRank, which allows a vehicle to use a new stable metric named "Information communicability" to autonomously rank different locations in the city and rank itself accordingly. Using GRank, the vehicle finds each location reachability and popularity taking into consideration the user interest satisfaction related to the location. It also considers its mobility pattern between different citywide neighborhoods to be available for content caching in each neighborhood. Vehicles available in popular neighborhoods classify themselves as important information hubs with higher *vehicle centrality* score in the network.

We then devise an incentive-driven social welfare game to optimally select the best set of vehicles identified by GRank for a given urban coverage and budget requirements. We formulate a Bayesian game with vehicles as rational players intending to cache content at potentially high rewarding neighborhoods. To cater the player rational behavior, we define a social norm where the vehicle utility is derived to satisfy a social welfare function as the Bayesian Nash equilibrium condition. Moreover, we consider a preference order over urban neighborhoods for each player where it receives a more reward for neighborhood closely related to its daily commute and vice-versa. Similarly, to ensure fairness, we define the equivalent utility for the service provider where reward is offered according to a preference order over different urban neighborhoods based on user interests.

An optimization problem is modeled to maximize the utility for both the vehicle and the service provider vis-à-vis the spatio-temporal coverage requirements and a available budget. We formulated the problem as optimization under budget and coverage constraint and we propose set of heuristics to optimally solve the NP hard maximum coverage problem. The vehicle selection algorithm is then proposed to recruit the set of vehicles with maximum utility as a function of their rational behavior and service provider interest to cover different urban neighborhoods under dedicated budget. Defining the problem as above, the proposed social welfare optimization seems to us as the best approach to propose the identification and recruitment of vehicles as urban content caches in the network. The major contributions are ths summarized as:

- An innovative vehicle ranking algorithm "GRank" is proposed allowing a vehicle to classify its eligibility as an efficient urban content cache in the network by first ranking different urban neighborhoods and then rank itself accordingly.
- An incentive driven social welfare game is formulated to fairly select among the best ranked vehicles the eligible candidates to cache content for different urban neighborhoods catering individual rationality.

The objective here is to show that GRank along the social welfare optimization game is an efficient vehicle selection system for content caching as the selected vehicles satisfied three times more interests compared to other schemes.

5.3 Autonomous Information Hub Identification - GRank

We present GRank as a global centrality measure enabling each vehicle to autonomously find its importance independently from a centralized infrastructure. It is difficult to use the vehicle contact frequency and duration to decide its importance in the network

due to the rapid changes in the time evolving vehicular network. Similarly, it is not always necessary for two nodes to communicate through shortest paths but they can possibly follow non-shortest paths in a dynamic VANET topology. Therefore, inspiring from the concept of communicability in complex networks [10], we introduce a new metric called "Information communicability", where each vehicle periodically finds its importance in the network based on the reachablity of the associated information with respect to the satisfied user interests. The interests are assumed to be generated and received from the neighboring vehicles using multi-hop interest forwarding.

Definition 1

(Information Association Profile) The information association profile for vehicle v from the sub-graph $\Theta_v \subset \mathbb{X}$ is the $|N| \times 1$ probability vector $P_x = [p_{x_1}, ...p_{x_N}]^T$, for N locations, where $(\cdot)^T$ is the matrix transpose and p_{x_i} is the probability of satisfying user interests for information associated to location $x_i \in \Theta_v$, where $\sum_{x_i \in \Theta_v} p_{x_i} = 1$.

Information type comprises different Intelligent Transportation Systems (ITS) applications (Safety warnings, Road congestion information, Infotainment...) with varying content popularity and priority. The regions are clustered using Voronoi tessellation [32] where the vehicles concentrated in an area are associated to the set of roads in a single Voronoi region $x \in \mathbb{X}$.

For temporal VANET analysis, we divide the time $T = (\overline{t_1}, \overline{t_2}, ...)$ into a sequence of regular time-slots, where the k^{th} time-slot is represented as $\overline{t_k} = [t_k, t_{k+1})$. Each vehicle finds its centrality at the time instant t_{k+1} based on the known information in the current time-slot, where t_k is the time instant at the beginning of the time-slot $\overline{t}_k$. We will refer to content, information or location interchangeably in the text since we associate content to locations in the urban map.

5.3.1 Information Global Centrality

In this section, we define the parameters each vehicle use to analytically find the information popularity for each location in the city. It periodically compute and exchange its own knowledge about the information popularity for each city location. It also uses the corresponding information popularity received from neighboring vehicles to rank all locations.

Definition 2

(Information Walk) We define an information walk of length k as a sequence of locations $x_0, x_1, ..., x_{k-1}, x_k$, such that, for each $i = 1, 2, ..., k$, there is a link from x_i to x_{i+1} in Θ_v.

It reflects the vehicle commute between the two regions taking into consideration all the possible regions it can visit in the way. For example, the set of intermediate regions with all possible set of information walks between location x_A and x_B are shown in Figure 5.1. The idea of using walk instead of path allows revisiting intermediate regions several times since it is sometimes inevitable to avoid traversing an intermediate region due to the static urban road structure.

Table 5.1: SAVING I - List of Notations

Notation	Description
$\mathbb{V}$	Set of vehicles
$\mathbb{X}$	Set of locations/regions
$\mathbb{E}^x/\mathbb{E}^v(t)$	Edges set between locations/vehicles at time t
$\mathbb{E}$	Edges set between vehicles and locations
$\bar{t}_k$	k^{th} time-slot under consideration
M	Total number of vehicles
N	Total number of locations
Θ_v	Set of locations associated to vehicle v
$p(\theta_v)$	probability distribution over types/ locations
P_x	Information association profile for a vehicle
p_{x_i}	Probability to satisfy interests for location x_i
k	number of regions between two regions
W	Vehicle adjacency matrix of information association profile for all locations
$C^v_{x_i x_j}$	Information communicability between locations x_i and x_j
Γ_v / Γ_x	Set of neighbors for vehicle v / locations x
$f^{\Gamma_v}_{x_i}$	Neighbors communicability function
$f^v_{x_i}$	Information centrality function
P_v	Vehicle mobility transition probability matrix
p_{x_j,x_i}	State transition probability between regions
π_x	Steady state probability of being in region x
$\rho^v_{x_i}$	Location importance parameter for location x_i
$G^v_{x_i}$	Information global centrality for each location
α	Tuning parameter for neighbors contribution
β	Tuning between the recent location popularity and overall location popularity
γ	Tune past GRank score with the current score
f_v	Vehicle centrality function
C_v	Vehicle global centrality
A	Vehicle actions as a function of its urban coverage
s	Vehicle strategy as a function of of its actions
u/U	Utility function for vehicle / service provider
s	Vehicle strategy as a function of of its actions
$R^v_{x_i}$	Reward vehicle v gets for location x_i
R_{x_i}	Reward dedicated by service provider for x_i
$b^v_{x_i}$	Vehicle cost to cover location x_i
B_v	Total Budget dedicated by service provider
L	Preference ordering set for different locations

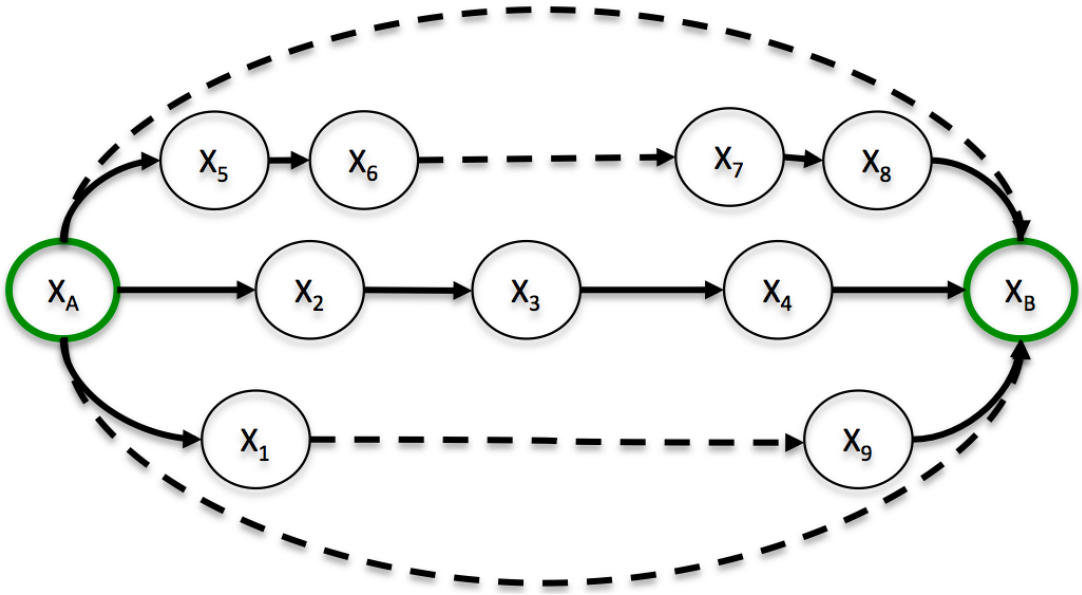


Figure 5.1: Information Walk

Definition 3

(Information Communicability) For each vehicle v, the information communicability for associated location x_i is the information reachability through the weighted sum of all possible information walks starting at location x_i and ending at x_j, $\forall x_i, x_j \in \mathbb{X}$:

$$C^v_{x_i x_j}(t_{k+1}) = \sum_{k=0}^{\infty} \frac{(W^k)_{x_i x_j}}{k!} = (e^W)_{x_i x_j}, \tag{5.1}$$

The term W is the weighted adjacency matrix with entries as area weight $w_{ij} = p_{x_i} p_{x_j}$, where p_{x_i} and p_{x_j} reflects the vehicle information association profile for location x_i and x_j considering the vehicle association to both regions. $(W^k)_{x_i x_j}$ uses the k^{th} power of the adjacency matrix in order to take into account the number of information walks of length k starting at location x_i and ending at location x_j.

The matrix function e^W can be defined using the Taylor series expansion: $e^W = I + W + \frac{W^2}{2!} + \cdots + \frac{W^k}{k!} + \cdots$. The identity matrix in the expansion does not affect the communicability as it only adds a constant to every value of the communicability measures. As a special case, the vehicle can find the reachability of its home location using self-communicability as the sub-graph centrality by replacing x_j in (5.1) by the home location x_i.

The information communicability for a location considers all the possible routes to it, however information walks with more number of intermediate regions should be given lower priority. To implement such penalty, taking $k!$ in the denominator ensures that longer routes are not weighted more than the shorter ones. The advantage is to prefer possible shorter routes, not necessarily the shortest path, for communicating between two regions.

Neighbor Communicability

The vehicle receives information communicability for each location from the neighboring vehicles which is used to find information centrality function. The neighbors communicability function is defined as;

$$f_{x_i}^{\Gamma_v}(t_{k+1}) = \frac{1}{|\Gamma_v| \cdot |\Theta_v|} \sum_{u \in \Gamma_v} \sum_{x_j \in \Theta_v} C_{x_i x_j}^{u}(t_{k+1}) \cdot C_u(t_{k+1}), \tag{5.2}$$

The neighbors belong to set of vehicles $\Gamma_v \subset \mathbb{V}(t_{k+1})$ at time instant t_{k+1}, where $C_{x_i x_j}^{u}(t_{k+1})$ is the neighbor information communicability for location x_i and $C_u(t_{k+1})$ is the respective neighbor vehicle centrality. A well popular neighborhood can easily be identified as the location where vehicles with high centrality exchange information more frequently. The information centrality function at the vehicle is given as:

$$f_{x_i}^{v}(t_{k+1}) = \frac{\alpha}{|\Theta_v|} \sum_{x_j \in \Theta_v} C_{x_i x_j}^{v}(t_{k+1}) + (1-\alpha) \cdot f_{x_i}^{\Gamma_v}(t_{k+1}), \tag{5.3}$$

The vehicle uses the neighbor communicability function in (5.2) to find each location information centrality function, where $\alpha \in [0, 1]$ is the tuning parameter to adjust the neighbors contribution.

Vehicle mobility pattern

The road sections in each region directly depends on the corresponding set of road sections in the adjacent regions. Therefore, the vehicle mobility pattern is modeled as a Markov chain where each region $x \in \mathbb{X}$ is represented as a state. The entries p_{x_i,x_j} of the vehicle state transition probability matrix P_v represents the transition probability between region x_i and the neighboring region $x_j \in \Gamma_{x_i}$ where Γ_{x_i} is the set of neighboring regions for x_i. For each vehicle, the transition probability matrix for the mobility between N regions is:

$$P_v = \begin{pmatrix} p_{x_1,x_1} p_{x_1,x_2} & \cdots & p_{x_1,x_N} \\ \vdots & \ddots & \vdots \\ p_{x_N,x_1} p_{x_N,x_2} & \cdots & p_{x_N,x_N} \end{pmatrix}$$

The vehicle steady state probability π_{x_i} to be in region x_i as a factor of the neighboring regions given by the relation $\pi_{x_i} = \sum_{x_j \in \Gamma_{x_i}} p_{x_j,x_i} \pi_{x_j}$. The steady state probability π_{x_j} is the probability of the vehicle availability in the j^{th} neighboring region, where $\sum_{x \in \mathbb{X}} \pi_x = 1$, and $\sum_{x_j \in \mathbb{X}} p_{x_j,x_i} = 1$.

The vehicle mobility pattern described above associates the vehicle availability in the region which helps in finding each region importance with respect to the vehicle. For instance, higher steady state probabilities of being in a region yields more association. It is also noteworthy that a vehicle might be unable to visit a neighboring region in case of pre-defined lane rules defined by the city administration. Therefore, by modeling the vehicle mobility pattern as a markov chain, we take into account the

regions where it will subsequently visit being in the current region, independent of the previous travel pattern and the underlying road structure.

The importance of a location x_i with respect to the vehicle availability pattern as well as the associated information reachability is given as $\rho^v_{x_i}(t_{k+1}) = \pi_{x_i} \cdot f^v_{x_i}(t_{k+1})$, where π_{x_i} yields the probability of the vehicle availability in the region and $f^v_{x_i}$ considers the respective information centrality function using (5.3). The information global centrality for each location can now be defined as:

$$G^v_{x_i}(t_{k+1}) = \beta \cdot \rho^v_{x_i}(t_{k+1}) + (1-\beta)\rho^v_{x_i}(t_k), \tag{5.4}$$

where $\rho^v_{x_i}(t_k)$ is maintained by each vehicle as the location importance value at the time t_k. The parameter $\beta \in [0,1]$ is used to tune between the recent location popularity and overall location popularity. Thus, we can identify popular locations at time instant t_{k+1} in the city with the maximum global centrality $\arg\max_{v\in\mathbb{V},x_i\in\mathbb{X}} G^v_{x_i}(t_{k+1})$ with respect to all vehicles. However, popularity of locations depends on several factors such as the information-type depending on the application requirements as well as time of the day. Similarly, we can use the maximum location importance $\max_{v\in\mathbb{V},x\in\mathbb{X}} \rho^v_{x_i}$ to identify popular neighborhoods for a longer time span.

5.3.2 GRank Computation

GRank allows the vehicle to consider its importance with respect to all locations in the city in order to measure its influence in the network. Each vehicle information association profile plays an important role in deciding its importance in the network. Thus, it should receive and subsequently satisfy more frequently the interests for information regarding the popular regions in the city. The vehicle centrality function at the time instant t_{k+1} is given as the average information global centrality for all associated locations:

$$f_v(t_{k+1}) = \frac{1}{|\Theta_v|} \sum_{x_i \in \Theta_v} G^v_{x_i}(t_{k+1}),$$

where the vehicle regularly updates its global centrality as the Exponential Weighted Moving Average (EWMA) of the vehicle centrality:

$$C_v(t_{k+1}) = (1-\gamma)C_v(t_k) + \gamma f_v(t_{k+1}), \tag{5.5}$$

where $\gamma \in [0,1]$ is a tuning parameter to tune between the past vehicle centrality score and the score in the current time-slot.

5.3.3 Summary

The steps required for a vehicle in order to find its global importance at regular time instants are summarized in Algorithm 5. For each associated location, the vehicle first finds the information communicability as the information reachability with respect to the location. Similarly, it exchanges the information communicability for each location with the neighboring vehicles in range along its respective local vehicle centrality. The vehicle uses its own as well as the neighbors information communicability to find the information centrality function for each location using (5.3). It then finds

INPUT: $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$: information association graph , Information global centrality, GRank in previous time-slot
OUTPUT: Updated GRank for the next time-slot at time-instant t_{k+1}
for each vehicle $v \in \mathbb{V}$ **do**
  for each associated location $x_i \in \Theta_v$ **do**
    find information communicability $C^v_{x_i x_j}, \forall x_i x_j \in \mathbb{X}$ using (5.1)
    for each vehicle neighbor $\Gamma_v \in \mathbb{V}$ in range **do**
      receive neighbor communicability $C^{\Gamma_v}_{x_i x_j}$, neighbor centrality C_{Γ_v},
    end for
    compute neighbors communicability function $f^{\Gamma_v}_{x_i}$ using (5.2)
    find information centrality function $f^v_{x_i}$ using (5.3), then location importance $\rho^v_{x_i}$
    compute information global centrality $G^v_{x_i}$ using (5.4)
  end for
  compute f_v, C_v
end for
return $C_v(t_{k+1})$

Algorithm 5: GRank

each location's importance taking into account its availability in that location. Each location's importance is then used to find the information global centrality. Each vehicle regularly updates the global information centrality in order to find its respective global centrality in the network. The vehicle global centrality is the average global information centrality with respect to all locations in the city. Thus, highly reachable and available vehicles with respect to all locations are considered as the most central vehicles in the city.

5.4 Information Hubs Recruitment

The information hub identification metric GRank enables each vehicle to classify itself as eligible candidate for content caching in an urban environment. In this section we extend the recruitment process by selecting the best set of vehicles identified for urban content caching. We formulate an incentive-driven social welfare game to avoid a rational vehicle trying to be selected as content cache in a neighborhood not related to its daily commute. To do this, we design the vehicle utility to satisfy a social welfare function considering user relevant preference for different urban neighborhoods. Moreover, to ensure fairness, the vehicle receives a high payoff for a neighborhood only in case it frequently visits/cache content for a particular neighborhood.

5.4.1 Game Formulation

The Bayesian game with incomplete information is defined as $< M, A, \Theta, p, u, G^T >$, where M number of vehicles play the stage game G for T periods represented as G^T, we define;

- $\Theta = (\Theta_1, ..., \Theta_m)$, where for each player v, Θ_v is the (finite) set of possible types

as the associated locations $\Theta_v \in \mathbb{X}$ in an urban environment. The vehicle provide coverage for location x_i declare its type as the respective location.

- $A = (A_1, ..., A_m)$, where A_v is the (finite) set of possible actions available to each player achieving city-wide spatio-temporal coverage for urban regions. The vehicle actions regarding location x_i provide the spatio-temporal coverage information for the location.

- $p = \Theta \rightarrow [0, 1]$ is the common prior over types, for finite space , we assume $p(\theta_v) > 0 \ \forall \ \theta_v \in \Theta_v$, where $p(\theta_v)$ represents the probability distribution of a player to provide coverage for its associated locations.

- Payoff function $u = (u_1, ..., u_m)$ where for each vehicle, $u_v : A \times \Theta \rightarrow \mathbb{R}$

Pure Strategy

A Bayesian pure strategy for vehicle in G is defined as $s_v : \Theta_v \times A_v$ as a mapping from its associated location (type) to the action it would play for the respective location (type) regarding its coverage. Similarly as an extension, the **mixed strategy** $s_v : \Theta_v \times \Pi(A_v)$ is a mapping from a vehicle associated location (type) to a probability distribution over its action choices regarding each location (type) coverage.

Let $s_w(a_w \,|\theta_w)$ denotes the probability under mixed strategy s_w that vehicle $w \in M$ plays action a_w, given that its type is θ_w. Given the probability distributions $p(\theta_1, ...\theta_m)$ over types, we can now find conditional distribution $p(\theta_{-v} \,|\theta_v)$ using Bayes rule since the vehicle knows its associated locations and evaluates its expected payoff according to the conditional distribution $p(\theta_{-v} \,|\theta_v)$ where $\theta_{-v} = (\theta_1, ..., \theta_{v-1}, \theta_{v+1}, ..., \theta_m)$.

Ex-interim expected utility

We define the ex-interim expected utility in the Bayesian game for a vehicle v with type θ_v and strategies represented by the mixed strategy profile s_v as:

$$EU_v(s \,|\theta_v) = \sum_{\theta_{-v} \in \Theta_{-v}} p(\theta_{-v} \,|\theta_v) \sum_{a \in A} \left(\prod_{w \in M, w \neq v} s_w(a_w \,|\theta_w) \right) u_v(a, \theta_{-v}, \theta_v), \tag{5.6}$$

The ex-interim notion is considered here since the players know their own types, but not the types of other players. In this case, it can form a belief with some probability for other players types. Breaking down the Equation 5.6, the first term $\sum_{\theta_{-v} \in \Theta_{-v}} p(\theta_{-v} \,|\theta_v)$ is the sum of the probabilities of types where the player type θ_v provides the belief about the probability of other players types θ_{-v}. The middle term, $\sum_{a \in A} \left(\prod_{w \in M, w \neq v} s_w(a_w \,|\theta_w) \right)$ concerns the probabilities of other players actions summed over possible actions and the third term $u_v(a, \theta_{-v}, \theta_v)$ is the payoff as the function of actions, given each player types. We can now derive the Bayesian Nash Equilibrium as:

Bayesian Nash Equilibrium (BNE)

A Bayesian equilibrium is a mixed strategy profile s that satisfies $s_v \in \arg\max_{s'_v} EU_v(s'_v, s_{-v}\,|\theta_v)$ for each v and $\theta_v \in \Theta_v$.

In case of the probability $p(\theta_v) > 0 \;\forall\; \theta_v \in \Theta_v$, for each vehicle v as defined above, then the ex-ante equilibrium condition states:

$$s_v \in \arg\max_{s'_v} EU_v(s'_v, s_{-v}) = \arg\max_{s'_v} \sum_{\theta_v} p(\theta_v) EU_v(s'_v, s_{-v}\,|\theta_v)$$

At the equilibrium conditions, each vehicle will maximize its expected utility regardless of its types. Thus, each vehicle will choose strategies that maximizes the vehicle expected utility in all associated locations. In order to reach the above mentioned equilibrium conditions, we define below the utility each vehicle will compute as its individual rationality.

5.4.2 Vehicle Utility as a Social Norm

In this section we derive the vehicle utility where each vehicle's rational approach to maximize its expected utility also satisfies the equilibrium conditions. The goals for a rational player to provide caching services in an urban environment are summarized as:

- Minimize cost: The obvious objective for each vehicle is to minimize its cost which is proportional to its (a) fuel consumption and (b) mobility pattern in an urban environment.
- Increase reward: The incentive for each vehicle can be well addressed by offering monetary rewards proportional to their participation to provide caching services. Each rational vehicle will try to cover neighborhoods with great amount of reward offered by the service provider in order to increase its reward.
- Increase centrality: A rational vehicle will increase its centrality by deliberate attempts of frequently covering important city neighborhoods due to their user-relevance importance. It is to note that vehicles in important neighborhoods receive high centrality score due to high content demand.
- Minimize commute perturbation: The aim of a rational vehicle is to gain more reward by providing less effort. Thus the vehicle will confine its mobility scope to fewer high rewarding neighborhoods.

The reward a vehicle can receive for the location x_i is given as $R^v_{x_i} = G^v_{x_i} \cdot R_{x_i}$, where $G^v_{x_i}$ is the information global centrality in Equation 5.4 and R_{x_i} is the reward dedicated by the service provider for an individual vehicle providing caching service for location x_i. The corresponding utility received by a vehicle v for location x_i in time-slot $\overline{t_k}$ is defined as:

$$u_v(x_i, \overline{t_k}) = p(\theta_v) \cdot C_v \cdot R^v_{x_i} - b^v_{x_i}, \tag{5.7}$$

where $p(\theta_v)$ is the prior probability of vehicle availability to cover as it a function of its type and C_v is the vehicle centrality from Equation 5.5. The vehicle centrality C_v ensures the best vehicles identified by GRank are selected as content caches. The vehicle cost $b^v_{x_i}$ is proportional to its fuel consumption with respect to its city-wide mobility in order to cover location x_i. The overall vehicle utility as a sum of its individual-rationality utilities at all time-slots with respect to all locations in Equation 5.7 is given as $u_v = \sum_{x_i \in \Theta_v} \sum_{\overline{t_k} \in T} u_v(x_i, \overline{t_k})$. The above vehicle utility is proportional to (a) its global centrality, (b) information global centrality, (c) reward service provider assigned for the respective locations while (d) considering the impact of vehicle cost with respect to its commute.

5.4.3 Social Welfare Optimization

The recruitment of set of best candidate vehicles as city-wide content caches is modeled as a discrete optimization problem where the social optimum is to maximize the expected utility, both for the service provider and the vehicle. In order to do so, we first define the social choice and the social welfare function as:

Definition 3

(Social Choice Function) Assume a set of vehicles $M = 1, 2, ..., m$, and the set of locations as possible types Θ. Let L be the set of strict total orders on Θ, where the elements of L are preferences of a vehicle v over its types Θ_v. A social choice function (over M and Θ) is a function $f : L^m \to \Theta$ that aggregates to a group choice.

Definition 4

(Social Welfare Function) Let M, Θ, L be as defined above. A social welfare function (over M and Θ) is a function that aggregates preferences for locations to a group preferences $F : L^m \to L^-$,where L^- is the set of weak total orderings (that is, total pre-orders) on Θ.

Similar to the vehicle utility, the service provider utility for each location x_i is defined as: $U_s(x_i, \overline{t_k}) = p(\theta_s) \cdot R_{x_i} - \sum_{v' \in V_{x_i}} B_{v'}$, where the probability distribution $p(\theta_s)$ over types signify the service provider preference for different urban neighborhoods. For the service provider , R_{x_i} is the reward dedicated for location x_i and $B_{v'}$ is the consumed budget by the set of vehicles V_{x_i} selected for caching services in the location. The city-wide utility for the service provider is given as the sum of the individual utilities from all urban neighborhoods:

$$U_s = \sum_{x_i \in \Theta_v} \sum_{\overline{t_k} \in T} U_s(x_i, \overline{t_k}) \tag{5.8}$$

The social welfare optimization needs to characterize the preferences by the service provider with respect to the overall preferences of each vehicle for different neighborhoods. Therefore the objective function needs to maximize the expected utility for both, EU_v for the vehicle and EU_s for the service provider in accordance with the

social welfare function over preference orderings in Definition 4. The social welfare optimization problem to maximize the expected utility $EU = \{EU_v, EU_s\}$ can be stated as:

$$\begin{aligned}
&\text{maximize} && F(EU) && \forall x_i \in \mathbb{X}, \overline{t_k} \in T && \\
&\text{subject to} && \sum_{x_i}\sum_{\bar{t}} A^{\text{th}}_{min}(x_i,\bar{t}) \geq \sum_{x_i}\sum_{\bar{t}} A(x_i,\bar{t}) \geq \sum_{x_i}\sum_{\bar{t}} A^{\text{th}}_{max}(x_i,\bar{t}) && && (\text{C}_1)\\
& && \max \sum_{x_i}\sum_{\bar{t}} B_v(x_i,\bar{t}) \leq \sum_{x_i}\sum_{\bar{t}} B^{th}_v(x_i,\bar{t}) && && (\text{C}_2)
\end{aligned}$$

The spatio-temporal coverage constraint in C_1 is a function of the actions concerning the city-wide coverage for different locations at different times while avoiding redundancy. The coverage requirements $\sum_{x_i}\sum_{\bar{t}} A(x_i,\bar{t})$ specifies the level of coverage needed in location x_i at a time-slot $\overline{t_k}$ depending on the location importance. $A^{\text{th}}_{min}(x_i,\bar{t})$ defines the minimum threshold for the coverage requirements and $A^{\text{th}}_{max}(x_i,\bar{t})$ represents the maximum coverage requirements with redundancy avoidance. The second constraint C_2 deals with the budget requirements. The total amount B_v a service provider can pay vehicles to cover different urban neighborhoods following the social preference ordering is limited by the threshold dedicated budget B^{th}_v. It is to note that different budget can be dedicated for different locations at different timings reflecting the spatio-temporal importance of different neighborhoods.

5.4.4 Algorithm: Vehicle Selection as Content Caches

The vehicle selection process can be treated as the maximum coverage problem, which has been proved to be NP-hard. Thus, we address our vehicle selection as NP-hard by presenting a greedy algorithm to recruit vehicles with a polynomial time complexity with a $(1 - 1/e)$ approximation solution. Algorithm 7 summarize the selection of vehicle as content caches in an urban environment without perturbing their daily commutes. The budget conditions in Line 5 ensures the assigned budget for each location at different times do not surpass the total budget dedicated by the service provider. Similarly the coverage constraints in Line 5 as a function of vehicle actions respects the minimum and maximum limits regarding the spatio-temporal coverage of a location. For a location x_i at time $\overline{t_k}$, the vehicles are selected to maximize the objective function where the vehicle with strategy profile s_v is ensured to satisfy BNE conditions.

The coverage provided by such vehicle actions are added to the overall coverage as well as its associated cost is payed from the dedicated budget in Line 7 and 8 respectively. As shown in Line 9, the vehicle is then added to the set of selected vehicles V for urban caching services. The selection process converges when (a) there are no eligible vehicles left to select (b) The maximum budget threshold for service provider is reached to recruit further vehicles.

INPUT: Vehicles set $\mathbb{V}$
OUTPUT: Selected vehicles set $V \subset \mathbb{V}$
Initialize $V = \phi$,
for each location $x_i \in \mathbb{X}$, time-slot $\overline{t_k} \in T$ **do**
  while $(B_v(x_i, \overline{t_k}) \leq B_v^{th}(x_i, \overline{t_k}))$ and
  $(A_{min}^{\text{th}}(x_i, \overline{t_k}) \geq A(x_i, \overline{t_k}) \geq A_{max}^{th}(x_i, \overline{t_k}))$, $\forall \overline{t_k}, \forall x \in X$ **do**
    Find $v \in \mathbb{V}$ for which strategy profile s_v maximize $EU(x_i, \overline{t_k})$
    $A(x_i, \overline{t_k}) \leftarrow A(x_i, \overline{t_k}) + A_v(x_i, \overline{t_k})$
    $B_v(x_i, \overline{t_k}) = B_v(x_i, \overline{t_k}) - b_{x_i}^v$
    $V \leftarrow V \cup v$
  end while
end for
return V

Algorithm 6: Optimal Vehicles Selection

Table 5.2: SAVING I - Simulation Parameters

Parameter	Value
Simulation platform	NS-3
Number of nodes	2986
Mobility trace	Cologne, Germany
Area	$6X6km^2$ city center
Duration	1 hour
Communication range	100m
Packet size	1024 bytes
Time granularity	1 sec
Simulation Runs	5

5.5 Performance Evaluation

5.5.1 Simulation Scenario

The performance of GRank along the proposed social welfare optimization is validated by a set of simulations under a realistic mobility scenario using traces from Cologne, Germany as an accurate mobility trace available for a vehicular environment. The number of vehicles in each region vary at different times of day. We analyze up to 2986 vehicles over the entire simulation duration with 1 s time granularity. The Cologne city center is simulated for 1 hour by clustering $6X6km^2$. The number of regions can vary between different cities depending on the size, although we divide Cologne into 36 neighborhoods. Urban roads with vehicle communication range around $300m$ is considered. The Nakagami path loss model is combined with a log-distance propagation model to cater for the impact of buildings and other obstacles. We associate each vehicle with a different set of location-dependent content as its cached content. Each vehicle is enabled to randomly generate interest with varying frequency at different

Table 5.3: GRank in different set of Simulations

Simulation	1		2		3		4		5		
GRank	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score	Mean
1	2687	1	483	1	1999	1	238	1	1239	1	1
2	259	0.9988	957	0.9906	9	0.9996	1105	0.9893	1407	0.9988	0.9954
3	768	0.9982	1932	0.9898	1215	0.9993	925	0.9889	34	0.9971	0.9946
4	1301	0.9980	103	0.9896	2471	0.9991	1867	0.9889	401	0.9965	0.9942
5	797	0.9971	950	0.9896	2204	0.9990	1902	0.9888	1195	0.9964	0.9941
6	216	0.9971	318	0.9788	57	0.9988	1349	0.9877	653	0.9963	0.9917
7	92	0.9969	2728	0.9770	808	0.9983	313	0.9866	444	0.9963	0.9910
8	76	0.9967	268	0.9738	849	0.9982	46	0.9862	508	0.9959	0.9901
9	658	0.9966	54	0.9735	1940	0.9968	1682	0.9860	108	0.9958	0.9897
10	1603	0.9965	150	0.9729	771	0.9966	158	0.9860	1904	0.9957	0.9895

time intervals for different (predefined) content as the content consumer.

5.5.2 Results: Individual Vehicles Ranking

GRank score for the top 10 vehicles from ten simulation runs are shown in Table 5.3. For each rank, The average score lies within a confidence interval of 0.01 for a confidence level of 95%. We rank all the vehicles in the simulation scenario, however, we are interested only to identify the top vehicles in each simulation. Therefore, GRank score is normalized with respect to the top identified node, i.e. the top node will have unity score followed by the relative score of other vehicles. We will use the same convention to interpret results in the later sections.

In the first simulation, due to its natural mobility pattern, the vehicle 2687 is identified to have the top GRank score among all vehicles in the network as it is able to satisfy more frequently the incoming interests throughout the simulation. It is also noteworthy that GRank will result in different set of top nodes by tuning the parameters α, β and γ depending on the application requirement.

For better analysis of GRank in different simulation scenarios, we consider the following performance metrics in comparison with the state of the art importance computation schemes (Degree, Closeness, Betweenness and Eigenvector centrality):

- Cumulative Satisfied Interests (CSI) for the top identified nodes by each scheme
- Comparison of top nodes identified by each scheme with their respective centrality scores
- Average aggregated throughput of the identified top ranked nodes by each scheme
- Cache hit rate for the top nodes by each scheme to evaluate GRank along CCN in VANETs

5.5.2.1 Cumulative Satisfied Interests

Cumulative Satisfied Interests refers to the total number of user interests satisfied during the simulation duration. Figure 5.2 shows the CSI score of the top five nodes identified by all these schemes in an average of ten set of simulations with random

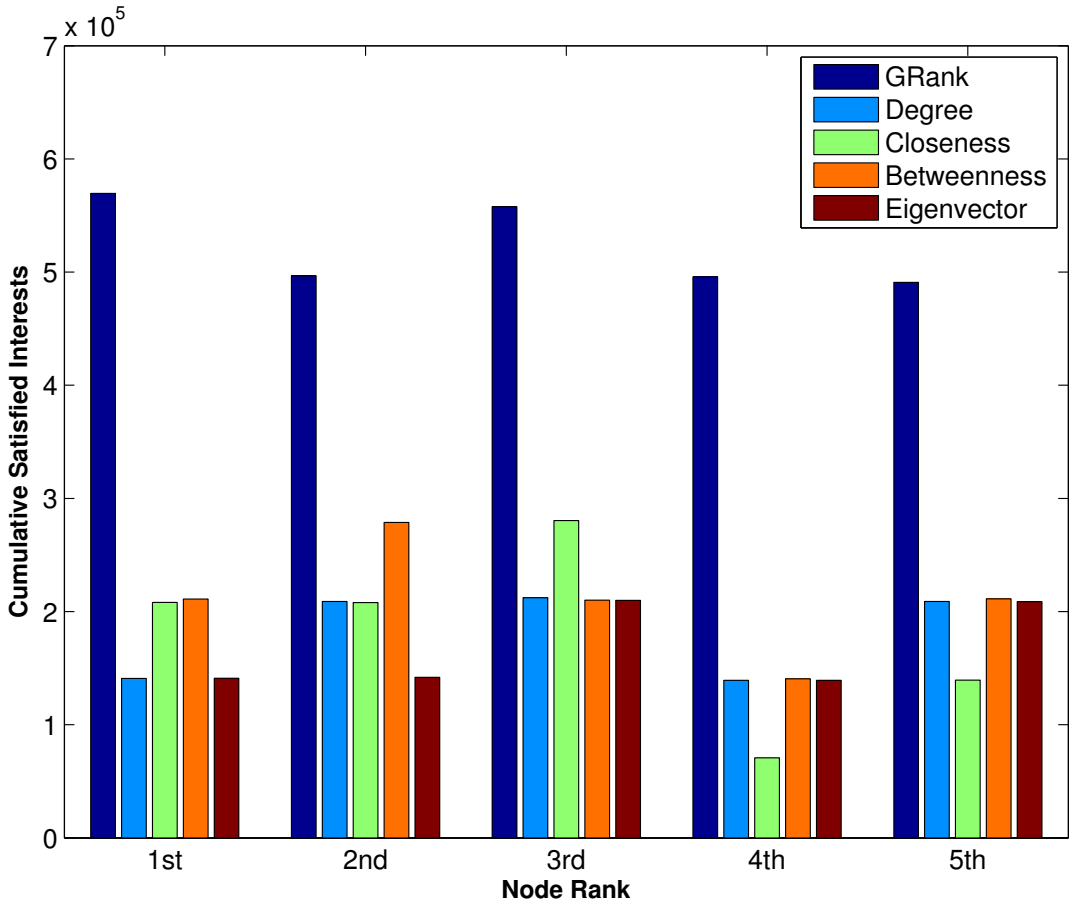


Figure 5.2: Cumulative Satisfied Interests by top identified vehicles using GRank compared with exisiting centrality schemes over an average of ten different simulations

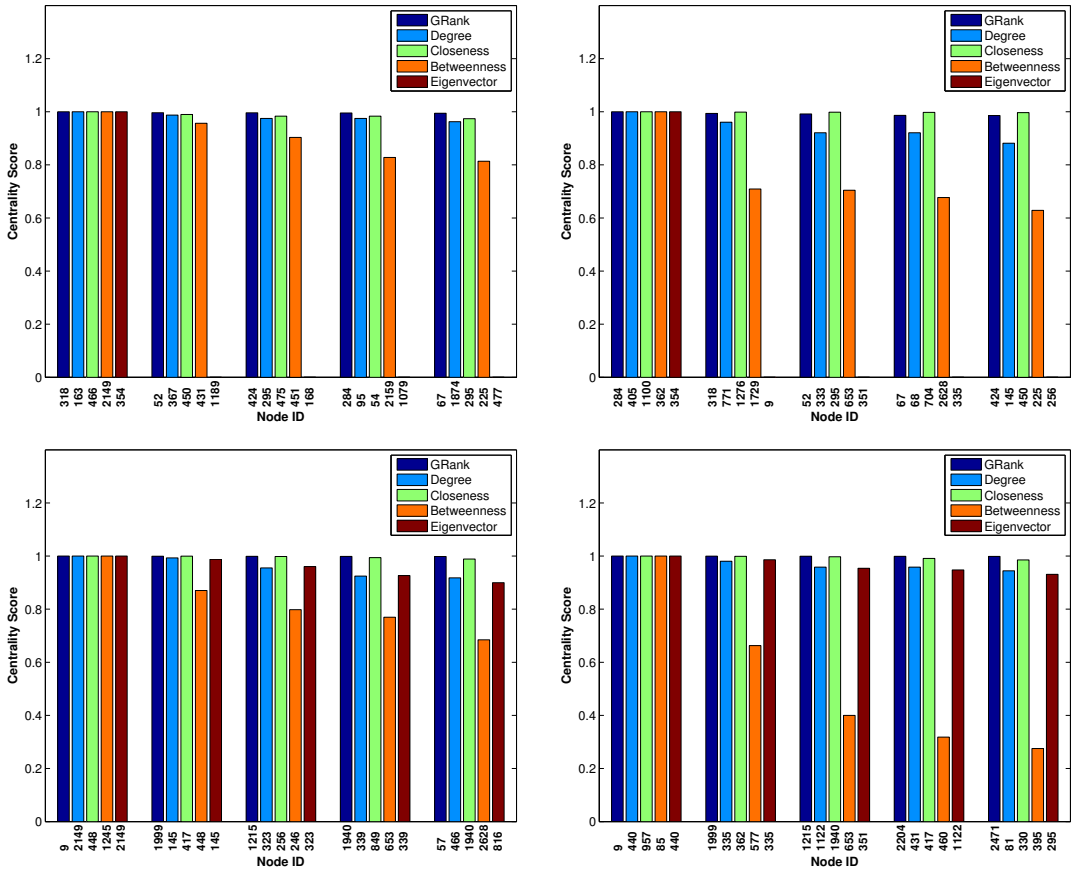


Figure 5.3: Temporal Snapshots after each 15 minutes comparing top identified nodes by each schemes

set of consumers and producers. Typical ranking schemes only takes into account physical topology towards computing a node importance in the network, ignoring the satisfied user interests. Nevertheless, GRank satisfied around 50% more user interests than other schemes in all the ten set of simulations due to the consideration of vehicle interest satisfaction profile along with its reachability and availability as key factors towards the vehicle importance.

5.5.2.2 Temporal Network Behavior Analysis

It is important to efficiently analyze the time varying behavior of our algorithm under the dynamic VANET topology. The time varying behavior of the relative score of the top five nodes identified by all schemes are shown by periodic network snapshots after each 15 minutes interval in Figure 5.3. We consider the top node identified by each scheme as benchmark by assigning it a unity score followed by the other vehicles in the ranking order.

At the beginning, vehicle 318 is ranked as the top vehicle by GRank, thought the other schemes underrated it. This is because we consider stable metric such as fixed locations associated to the vehicle instead of using ephemeral topological information. Vehicles also change places along the ranking order. For example, around 15 to 30

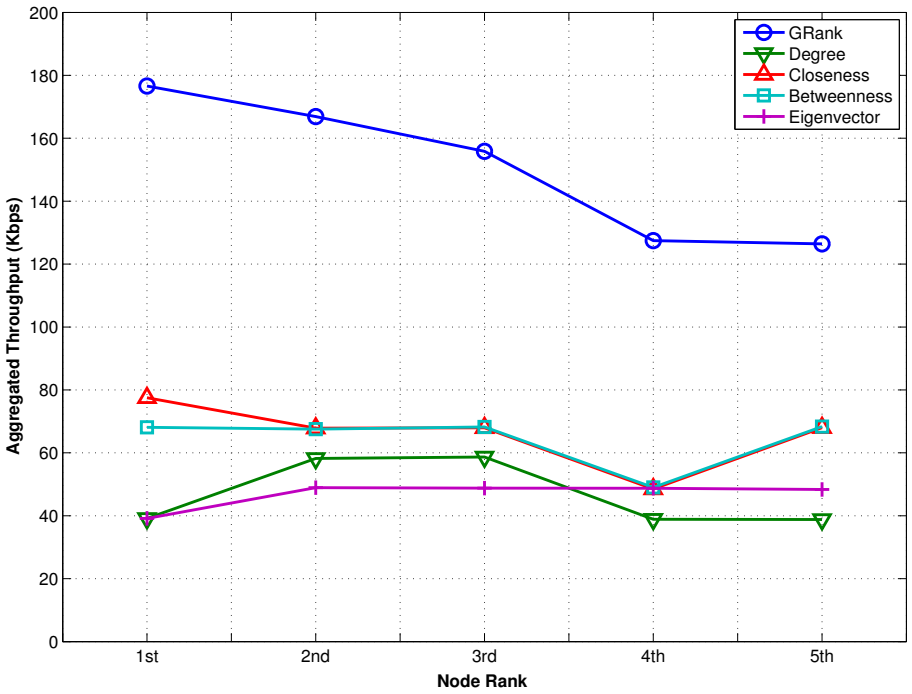


Figure 5.4: Average aggregated throughput by the top identified nodes using each scheme in ten simulations

minutes in the simulation, the top ranked vehicle 318 is replaced by 284, which is then replaced by the vehicle 9 at around 30 minutes till the end of the simulation duration.

An interesting results was observed in the initial 30 minutes (Figure 5.3 a,b): Only one node yields a high Eigenvector centrality score followed by other nodes with a negligible Eigenvector centrality score. We investigate this behavior and found that the principle eigenvalue yields the top nodes where the eignvector is shifted towards the principle component. Thus, resulting in a single vehicle as the top vehicle. This proves that the famous Eigenvector centrality fails to assign significant score to a large fraction of nodes in a larger network, while GRank does not show such behavior. Typical centrality schemes result in different set of top vehicles at each snapshot. It is because such schemes only consider the instantaneous shortest paths between vehicles towards ranking the vehicles at a particular time instant as well as require the complete network topological information. However, such complete network information is not available to an individual vehicle in highly unstable VANETs. GRank considers the information walks between locations in the city ensuring more stable set of top nodes, thus, not affected by the network dynamics which is not the case for other schemes.

5.5.2.3 Aggregated Per Node Throughput

We evaluate the proposed ranking scheme by analyzing the throughput at important nodes in the network. Figure 5.4 shows the aggregated per node throughput of the

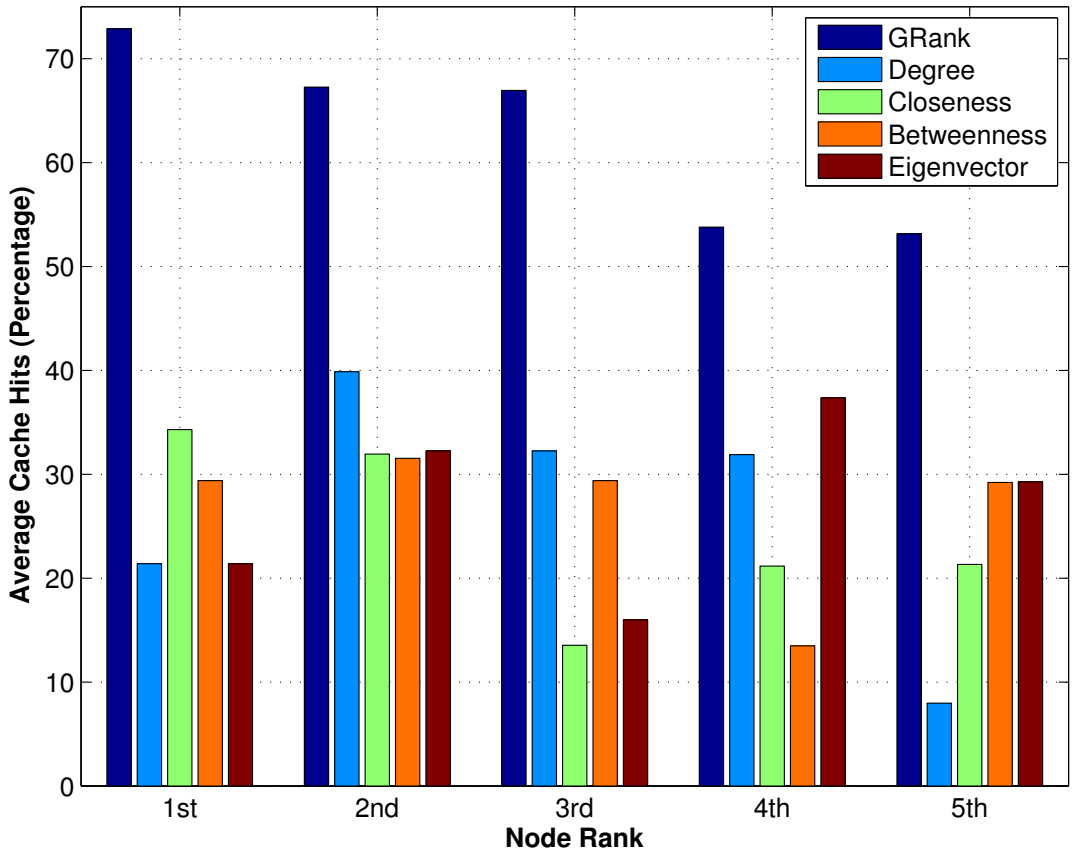


Figure 5.5: Average cumulative cache hit rate by the top identified nodes using each scheme in ten simulations

top nodes identified by each scheme. The average aggregated throughput (Kbps) is computed over the entire simulation duration for ten set of simulations. The top nodes identified by GRank yields three times more throughput compared to centrality schemes since they satisfied more frequently interests regarding popular locations. For GRank, the per node throughput of the top ranked vehicles shows a decrease between the ranking order while a variation is seen for Degree, Betweenness and Eigenvector centrality. It is clear that GRank outperformed all schemes as it incorporates information importance in order to rank vehicle, while other schemes entirely rely on topological measures such as the node degree or shortest paths towards finding vehicle importance. Moreover, higher per nod throughput also shows that efficiency of top ranked vehicles in various network operations.

5.5.2.4 In-network Cache Hit-rate

We evaluate GRank for the CCN built-in feature of In-Network caching at the intermediate nodes. For this purpose, we computed the cache hit rate at the top five nodes identified by each scheme. A second successful response by a vehicle for the same content is considered a cache hit. Figure 5.5 shows the cumulative cache hit rate for

the top nodes identified by each scheme for the entire simulation duration for ten set of simulations. The top vehicles identified by GRank yield a higher hit rate than those identified by the other schemes as vehicle containing important information relevant to the users responds and subsequently cache more frequently leveraging their mobility pattern compared to other vehicles. This proves that In-Network caching offered by CCN in GRank implementation overcomes the mobility and intermittent connectivity constraints of VANETs.

5.5.3 Results: Selected Set of Vehicles

In this section, we present the evaluation results of the proposed social welfare game. The Algorithm 7 optimally selects the set of best ranked vehicles under city-wide content caching requirements. Besides the social welfare game, we implement the following benchmark selection schemes for comparison:

- Reputation Aware (RA) recruitment ([13] and [52]) where the set of vehicles are selected based on their rank/reputation as a key factor.
- Minimize Budget (MB) approach where the only vehicles with the minimum cost are selected for content caching services till the consumption of dedicated budget.
- Maximize Coverage (MC) aims to select the set of vehicles which maximize the city-wide spatio-temporal coverage requirements ([15] and [55]).
- Budgeted Maximum Coverage (BMC) where the target is to maximize coverage within the given budget limitations ([23]).

The set of vehicles selected by each scheme are validated by the following performance metrics:

5.5.3.1 Success Rate

Success rate refers to the percentage of the generated consumer interests successfully satisfied over the entire simulation duration. The overall success rate of the selected set of vehicles by our approach is compared with the benchmark approaches. Figure 5.6 shows the percentage of consumer interests for different locations successfully satisfied by the set of recruited vehicles by each approach. The proposed SWO results in a success rate between 55% to 67% during the simulation for the incoming user interests. Only the famous BMC follows by achieving a maximum of 54% at the end of the simulation. We observe that selection based on only cost (MB) yields the worst performance while opting only for rank (RA) and coverage (MC) results in a maximum of 42% success rate. It is because SWO consider the vehicle reputation (GRank) as well as its individual-rationality to satisfy more interests in its associated urban locations.

5.5.3.2 Aggregated Throughputs

We find the throughput achieved by the set of vehicles selected using each approach for a city-wide information hubs. Figure 5.7 depicts the aggregated throughput achieved over the simulation duration. The proposed SWO results in the throughput of about

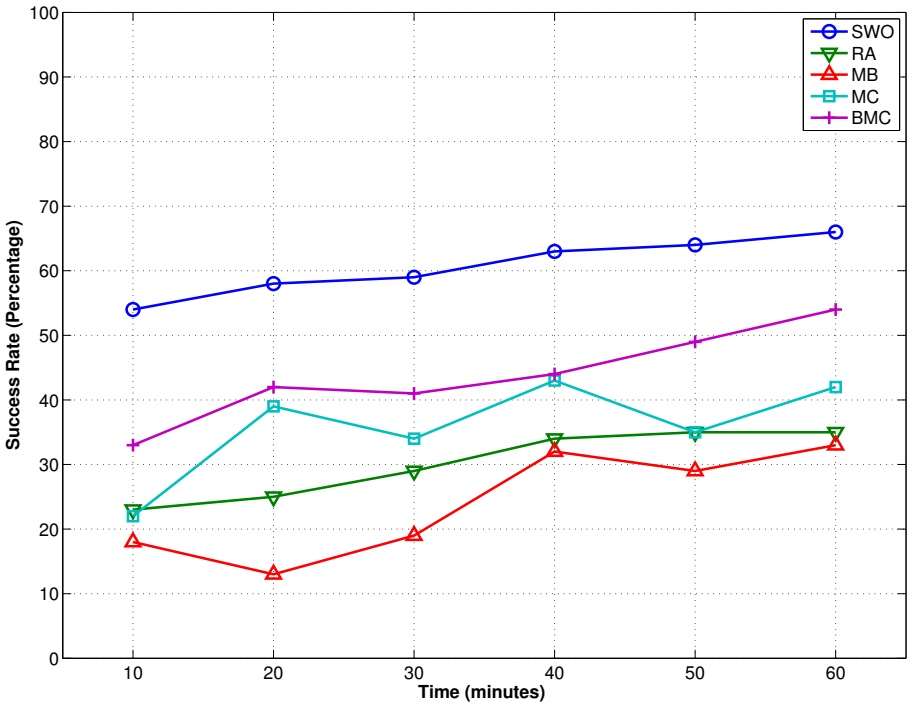


Figure 5.6: Average success rate achieved by the selected set of vehicles (SWO: Social Welfare Optimization, RA: Reputation Aware, MB: Minimize Budget, MC: Maximize Coverage, BMC: Budgeted Maximum Coverage)

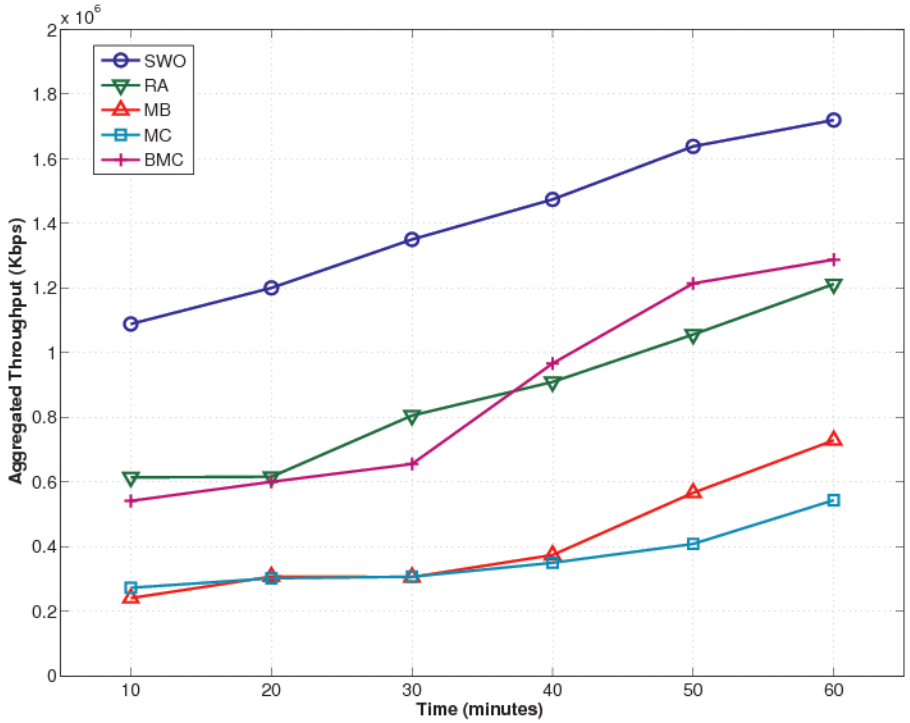


Figure 5.7: Average aggregated throughput achieved by the selected set of vehicles (SWO: Social Welfare Optimization, RA: Reputation Aware, MB: Minimize Budget, MC: Maximize Coverage, BMC: Budgeted Maximum Coverage)

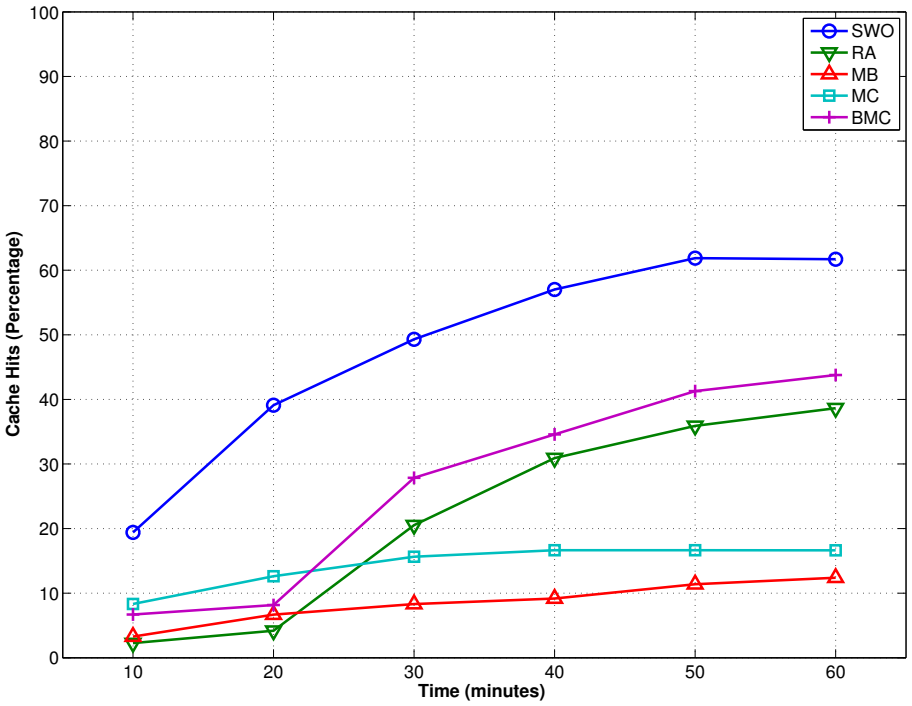


Figure 5.8: Average cache hit rate achieved by the selected set of vehicles (SWO: Social Welfare Optimization, RA: Reputation Aware, MB: Minimize Budget, MC: Maximize Coverage, BMC: Budgeted Maximum Coverage)

twice the amount achieved by the vehicles selected using RA and BMC and three times that of MB and MC approach. This infers that the selected set of vehicles identified by GRank along SWO are spatio-temporally available to actively participate in network activities. We also noticed an increase in throughput is observed due to the consideration of cumulative throughput achieved at the individual vehicles during the simulation. This reveals that such vehicle can be recruited as reliable information hubs for content storage at different urban neighborhoods.

5.5.3.3 In-network Cache Hit-Rate

A key criteria to evaluate content caching capability is to compute the content cache hit ratio achieved by the selected set of vehicles. Similar to individual vehicle cache hit-rate, we compute the cumulative cache hit-rate of the vehicles selected using each approach as shown in Figure 5.8. Despite the high mobility and intermittent connectivity, the vehicles selected as information hubs using our proposed SWO yield up to 60% cache hit rate compared to other approaches. RA and BMC follows with a maximum cache hit-rate of 42% while MB and MC results in below 20% cache hit rate. The major reason for the selected vehicles to be capable to respond for a subsequent interests for content is their availability at the right time and the right location to

satisfy user interests. It is also noteworthy that the preference ordering defined by the social choice function and social welfare function (Definition 3 and 4) helped the increase in cache hits for content of relatively important locations.

5.6 Conclusions

It is challenging for the current cellular networks to cope with the increasing content demand from urban mobile users. In this chapter, we target the recruitment of vehicles as distributed content caches to maximize content availability close to mobile users in an urban environment. We modeled the identification of such vehicles as node identification problem in a network where important nodes are identified as important information hubs. To do so, we first introduced an innovative vehicle ranking algorithm, "GRank", exploiting the vehicle autonomous decision making ability, allowing it to classify its eligibility as potential content cache. GRank uses an information-centric approach to identify user relevant important city-wide locations and then rank the vehicle accordingly. Unlike our previously proposed schemes, InfoRank and CarRank which were proposed for local data collection, GRank goes beyond local scope as it considered a global reachability and availability of cached content at vehicles with respect to user relevant interests. At the same time, GRank is a network aware centrality metric depending on relatively stable metrics compared to the unstable benchmark centrality schemes.

Of all the vehicles on the road, it is challenging to select and recruit the best set of vehicles that are adequate for content caching in an urban environment under given budget and coverage requirements. To do so, we modeled the problem as a social welfare optimization game to optimally select the best ranked candidates to be recruited as content caches taking into account the spatio-temporal coverage and budget requirements. To ensure fairness proportional to the vehicle cost, we derive the vehicle utility satisfying a social welfare without perturbing its daily commute. Simulations are performed using realistic mobility traces with up to 2986 vehicles. Results suggest that the recruited set of vehicles yield twice the amount of success rate, throughput and cache hit ratio compared to other schemes. In this chapter, we successfully identified and recruited the potential vehicles as content caches in an urban environment, however, there is still a need for caching decisions to find out which of the candidate vehicles identified by GRank should cache which content, thus requiring an efficient collaborative content cache management system. Therefore, the next chapter discusses fair content cache management between spatio-temporally vehicles allowing them to self-organize in order to cater selfish vehicles and optimizing content caching with respect to limited storage and monetary resources.

Chapter 6

SAVING as Data Storage: A Collaborative Caching Game at Mobile Fogs

6.1 Introduction

Daily lots of mobile users in an urban environment access content on the internet from different locations. It is challenging for the connection centric nature of current service providers to cope with the increasing demand from large number of spatio-temporally collocated users on the move. We presented the identification and selection of urban caches in the previous chapter which alleviates the issue with in-network caching feature to offload content close to users [30], though still efficient cache management is required to find who should cache what, when and where in an urban environment, given limited resources.

In order to answer these questions, we propose in this chapter to model the content cache management problem as a coalition game where the set of vehicles self-organize and collaborate to perform distributed content caching. Cache management is studied in different literature, though there is no adequate model which considers the spatio-temporal content characterization regarding its availability and popularity in the network. Existing schemes only considers content popularity along policies such as First-in First out (FIFO), Least recently Used (LRU) and Least Frequently Used (LFU) where such schemes lack policies for optimal content retrieval in the network.

To address this, we first define a novel relation between urban content popularity and availability in the network. Based on which we then find the node eligibility to cache content based on its urban reachability and centrality. To maximize the distributed urban cache and content availability, we propose a coalition game for resource pooling allowing users to self-organize into mobile fogs to cater rational users assuming a monetary reward is paid proportionally to the cost for caching. Nodes autonomously choose to merge into different spatio-temporal coalitions to offer maximum storage buffer to the service provider. Results show that our content caching approach achieved an average cache hit rate of 70% compared to an average of 20% using existing cache management policies.

The chapter organization is the following. The next section discuss the context and motivation for the need of mobile fogs for urban data storage. Section 6.3 models the content popularity and availability relation for efficient cache management. The Section 6.4 presents the coalition formation for the mobile fogs to enable collaborative caching in an urban environment. Performance evaluation and results are discussed in Section 6.5 and Section 6.6 concludes the chapter.

6.2 Context and Motivation

The increase in smart mobile devices results in growth of content demand by lots of consumers in closer urban proximity each with multiple portable devices. For example, large number of users on the move in an urban environment are interested to watch the video of a latest episode of a hot TV show/drama or a sports highlights. Provisioning of such popular content to each user requires lots of redundant connections between users and the service provider, given that the content is requested by lots of spatio-temporally co-located users with similar social interests.

It is now challenging for the current "connection centric" network infrastructure to facilitate content availability for such large number of mobile users in close proximity in an urban environment while offering attractive tariff plans supporting unlimited bandwidth. Here again, we advocate to use the recently proposed Content-Centric Networking (CCN) [2] paradigm which address the issue by decoupling the content provider-consumer and support in-network caching at intermediate nodes.

Content caching at individual nodes exploit different content replacement strategies such as first-in first-out (FIFO), Least Recently Used (LRU) and Least Frequently Used (LFU) though none of such policies is able to model the content profile with respect to the economic and social interest of large number of users. On the other hand, collaborative caching at set of spatio-temporally co-located nodes can allow them to provide complete caching services to the service provider and receive reward from the service provider proportionally to the offered storage. In order to provide such complete services, the nodes need friends to collaborate where multiple nodes together can efficiently offer the necessary storage services.

To do so, in this chapter, we transform the content caching concept towards an innovative approach where mobile nodes in an urban environment can subscribe to offer distributed caches to the service provider. We model the problem as a coalition game where we define *Mobile fogs* as sets of co-located mobiles nodes that can self-organize to form a "fog" to offer distributed resources (computing, communications, caching) closer to urban users. Thus, the aim is to maximize content availability and minimize cost by caching content closer to many consumers.

This however invokes the following questions; first, given the massive content constantly generated and consumed by mobile devices, which content is important to cache with respect to the social interests of geographically co-located users at different times of the day? Second, given limited resources, which nodes can be considered suitable candidates for urban caching given limited resources, both for the service provider and the mobile nodes? Third, given the existence of lots of such subscribed nodes including rational nodes, how to ensure fair selection of mobile nodes to cache in different urban neighborhoods, specially, in populated areas (spatial) or during peak traffic hours

(temporal) to maximize content availability?

Thus, there is a need to address questions regarding "Who should cache what, when and where?". This requires an understanding of the content as well as the node's spatio-temporal profiles. To address this, we first define a novel relation considering simultaneously the content spatio-temporal popularity and availability to characterize its importance. Based on the content profile, we apply social networks metrics (CarRank and GRank) for a node to autonomously find its caching capability by computing its local and global centrality as a metric of its connectivity and reachability in the network.

We then present a coalition game between nodes for resource pooling to offer maximum storage where spatio-temporally co-located nodes self-organize into mobile fogs to offload content from the infrastructure. Each node merges into coalitions formed at high centrality nodes according to its preferences relations over coalitions. Since the distributed caching problem is NP-Hard [12], an optimization problem is thus formulated along with an algorithm to maximize the offered distributed storage to the service provider with required content availability, storage and budgetary constraints for the individual node as well as for the service provider. The service provider finally selects the optimal coalitions required to cache in an urban environment with the goal of maximizing content availability. Hence, the contribution of this chapter can be summarized as follows:

- A novel content spatio-temporal availability-popularity relation is proposed to decide a content importance in order to be cached at a node.
- A new set of node eligibility metrics presented for social-aware content caching. These metrics aim at classifying a node's caching capability proportionally to its connectivity, cost and reachability in the network.
- A coalition game is proposed to form mobile fogs for resource pooling at spatio-temporally co-located nodes where service provider optimally selects coalitions of nodes as urban caches to maximize content availability.

The objective here is to show that CarRank and GRank enabled vehicles to self organize into social aware coalitions with high success rate and cache hit rate for the generated user interests compared to existing algorithms.

6.3 Who should cache What?

The first step towards efficient content caching is to understand the eligibility of the content to be cached based on a relevant amount of user demands relatively to the content as well as its availability in the network. Before discussing the content eligibility, we first present the network model to consider for the rest of the chapter. Similarly for a node, we define a novel approach to model the relation between the content availability and its popularity, while considering its validity. Moreover, it is important to consider the storage space available and the cost associated for caching content at the node along its reachability in an urban environment. In the following sections we will model a relation considering all such aspects.

6.3.1 Network Model

The network connectivity is modeled as an undirected graph $G(\mathbb{V}(t), \mathbb{E}^v(t))$, where $\mathbb{V} = \{v\}$ is the set of nodes and where the time instant t integrates the temporal network characteristics. $\mathbb{E}^v(t) = \{e_{jk}(t) \mid v_j, v_k \in \mathbb{V}, j \neq k\}$ is the set of edges $e_{jk}(t)$ modeling the existence of a communication link between nodes j and k at time instant t. To consider the dynamic nature of the network topology, we assume the time $T = (\overline{t_1}, \overline{t_2}, ...)$ as a sequence of regular time-slots, where the k^{th} time-slot is represented as $\overline{t_k} = [t_k, t_{k+1})$. The urban environment is represented by the undirected graph $G(\mathbb{L}, \mathbb{E}^l)$, the set of vertices $\mathbb{L} = \{l\}$ represents different locations l and the set of edges $\mathbb{E}^l = \{e_{pq} \mid l_p, l_q \in \mathbb{L}, p \neq q\}$ are the respective boundaries that connects different neighborhoods.

We define the set of uniformly sized content as $X = \{x\}$ where x is an indivisible content chunk in the network. For each content, it is important to understand its availability profile as well as its popularity profile. Similarly, we need to take into account the spatio-temporal relation of its availability versus its popularity at a location $l \in L$ and a time slot $\bar{t}$. Therefore, in the remaining of this section we model and map the content availability and its popularity with respect to location and time.

6.3.2 Spatio-temporal Content Profile

It is important to consider both, the spatial and temporal characteristics of the content available as well as demanded in the network. To do this, we let the node extrapolate from the past information for the cached content availability and popularity to decide regarding its future profile. We define below the proposed content availability and popularity respectively:

Definition 1

(Content Availability) We formally define the spatio-temporal availability for a content piece $x \in \mathbb{X}$ cached at a node $v \in \mathbb{V}$ as

$$A_v^x(l_a, \bar{t}_b) = \begin{cases} 1, & \text{x cached at v} \\ 0, & \text{otherwise} \end{cases}$$

where, $l_a \in \mathbb{L}$ denotes a location and $\bar{t}_b$ is the time slot at which the content is available at the node. Thus, the content availability is zero for all the content $x \in \mathbb{X}$ not cached at the node at a particular location and time slot.

To model the storage requirements for a particular content x, we denote Q_x as the number of chunks/pieces cached at a node, assuming the possibility of having multiple content chunks. The file size for the content can be defined as $F_x = Q_x \cdot x$. Thus, the file size of all the content cached at a node v is given as $F_v(l_a, \bar{t}_b) = \sum_x A_v^x(l_a, \bar{t}_b) \cdot F_x$, i.e. the sum of file sizes for all the possible content at location l_a and time slot $\bar{t}_b$.

The content popularity profile is modeled considering users with common social interests therefore we define below for a node, the spatio-temporal content popularity with respect to a location and time for the cached content.

Table 6.1: SAVING II - List of notations

Notation	Description
$\mathbb{V}$	Set of nodes
$\mathbb{L}$	Set of locations
$\mathbb{X}$	Set of content
$\mathbb{E}^l/\mathbb{E}^v(t)$	Edges set between locations/nodes at time t
$\mathbb{E}$	Edges set between nodes and locations
$T/t/\bar{t}$	Total time / time instant / Time slot
$t_v^{x_f}$	Time instant: last interest satisfied
$\overline{t_v^{x_{av}}}$	Time-slot: average content validity
$\bar{t}_k$	k^{th} time-slot under consideration
λ_v^x	Number of received interests for content x
Λ_v^x	Number of received interests at v for all contents
M	Total number of nodes
N	Total number of content chunks
A_v^x	Content x availability at node v
Q_x	Number of chucks for content x
F_x	File size for content x
F_v	File size for all content at v
P_x	Content popularity matrix
p_x	User interest probability for content x
τ_v^x	Content validity metric
δ	Content replacement tuning parameter
I_v^x	Interest satisfaction frequency
r_v^x/R_v^x	Successful/Overaall interests satisfied
Π_v^x	Content popularity/availability relation
LC/GC	Node Local/Global centrality
f_I^v/α	Content importance function/tuning parameter
$f_{L,T}^v/\beta$	Sp-temporal availability function/tuning parameter
f_Γ^v/γ	Neighborhood connectivity tuning parameter
θ	LC/GC tuning parameter
ς	Node centrality tuning parameter
C_v	Node centrality
B_v/B_v^t	Available/total buffer at node v
f_v^C/f_v^R	Node cost/reward function
S_x	Percentage of space occupied by x
C'	Additional node costs
R_B	Reward offered to node
U_v/U_s	Node/coalition utility function
S	Set of coalitions
$\succ$	preference relations among coalitions
y_v	node payoff in imputation
w_s	weight for coalition s

Definition 2

(Content Popularity) The content popularity in the network is represented by the probability matrix

$$P_x(l,\bar{t}) = \begin{pmatrix} p_x(l_1,\bar{t}_1)\ p_x(l_1,\bar{t}_2) & \dots & p_x(l_1,\bar{t}_N) \\ \vdots & \ddots & \vdots \\ p_x(l_N,\bar{t}_1)\ p_x(l_N,\bar{t}_2) & \cdots & p_x(l_N,\bar{t}_N) \end{pmatrix}$$

where $p_x(l_a,\bar{t}_b) = \frac{\lambda_v^x(l_a,\bar{t}_b)}{\Lambda_v^x}$ is the probability of user interests for content x at location l_a and time $\bar{t}_b$. Here, $\lambda_v^x(l_a,\bar{t}_b)$ represents the number of interest for the content x at location l_a and time $\bar{t}_b$ and $\Lambda_v^x = \sum_x \lambda_v^x(l_a,\bar{t}_b)$ is the total interests for all contents cached at the node v at the required location l_a and time slot $\bar{t}_b$. This metric considers content already cached (available) in the node as a result of a user interest while considering all contents fairly.

To consider social awareness, we assume nodes constantly receives interests for content regarding different information. Thus, the node can consider an information as popular if it observes an increase in the number and frequency of user interests for the associated content in a respective location and time-slot. It is important to find the node relation to the cached content with respect to the received user interests in order to efficiently cache content of interest for the intended users in the network. Therefore we define below a user interest satisfaction based metric which complies with such criteria:

Definition 3

(Interest Satisfaction Frequency) We define $I_v^x(l_a,\bar{t}_b) = \frac{r_v^x(l_a,\overline{t_b})}{R_v^x}$ as the frequency of user interests satisfied by the node for content x at location l_a and time-slot $\overline{t_b}$, where $r_v^x(l_a,\overline{t_b})$ are the number of successful responses in the previous time slot and $R_v^x = \sum_x r_v^x(l_a,\bar{t}_b)$ are the overall successful responses for all the contents, $\forall x \in X_v$ cached at the node $v \in \mathbb{V}$ at the respective location and time slot.

We assume each user interest specifies a hint (such as Time To Live - TTL) as the content validity for the intended user. Since such TTLs vary for interests received from different users at the node v. We let the node compute (record) the average interest validity hint extracted from all the received user interests. It is defined as the average time slot specified by user interest for the cached content validity averaged over all the interests received in the past for the content. It enables the node to decide whether the cached content is still demanded by users in the network or not. Therefore, we allow each node to devise a content replacement strategy based on such user specified content validity hint. Let $t_v^{x_f}$ be the time instant when the interest for the content x was previously satisfied by node v, where x_f denote the final instant satisfaction for content x and the average interest validity time slot is represented as $\overline{t_v^{x_{av}}}$ for the content.

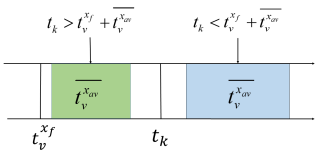


Figure 6.1: Content validity hint indicated by user interests

Definition 4

(Content Validity) Leveraging the user interest validity hint, we define $\tau_v^x(l_a, \overline{t_b})$ as the measure of the content validity scope for the next time instant t_{k+1} as:

$$\tau_v^x(l_a, \overline{t_b}) = \begin{cases} 1 & t_k \leq t_v^{xf} + \overline{t_v^{xav}} \\ e^{-\delta \overline{t_v^{xav}}} & t_k > t_v^{xf} + \overline{t_v^{xav}} \end{cases}$$

where $\delta \in [0, 1]$ is the tuning parameter to adjust the exponential decay. The content validity metric τ can be better described in Figure 6.1 showing examples on both conditions for the current time instant t_k less/greater than the sum of the time instant the content was last responded t_v^{xf} and the time slot $\overline{t_v^{xav}}$ specifying the average user interest validity.

Using this decay, a node decides how quickly to replace the content which is not needed. This decay can vary between applications requiring different content replacement policies, thus we left it as a tunable metric. The purpose is for a node to take into account how long there exist active interests for a cached content. In case the node receives no active interest in the previous time slot and the average interest validity is expired, the content validity follows an exponential decay since the information is of less importance in the network. On the other hand, τ is set to unity for the content required to be constantly cached at the node.

Definition 5

(Content Availability vs Popularity Relation) We present a metric on the content eligibility for a node to model together a cached content's "availability" while considering its "popularity" with respect to the received user interests. The content popularity metrics, τ and p_x, model the user relevant content validity and popularity while A_v^x and I_v^x represent the content x availability at node v along the frequency of interests satisfied for users in an urban environment. The content availability at a location l_a and time $\overline{t_b}$ can be mapped together with its popularity in the network by the relation:

$$\Pi_v^x = \epsilon \cdot p_x \tau_v^x + (1 - \epsilon) \cdot A_v^x I_v^x \tag{6.1}$$

where A_v^x is a binary variable modeling the content x availability at a node v and ϵ is the tuning parameter to balance between the content available at the node and content not available at the node but still popular indicated by p_x. Therefore, the first term in Equation 6.1 considers the popularity of content either cached or not cached at the node and the second term models the availability of content already cached at the node.

The above relation allows us to equally consider both the content popularity and availability towards finding the eligibility. The term τ_v^x decides the content x validity with respect to the node v, p_x is the interest probability for the content and I_v^x is the frequency of satisfied interest for the content by the node. Thus, the eligibility of the content cached a node is proportional to the node-content relation/association with respect to the user interest. The relation Π_v^x allows a node v to consider simultaneously the content x availability and popularity profile with respect to user interests towards its caching decisions.

It is to note that Π_v^x is a single metric that incorporates a hybrid of content recency and popularity (τ_v^x and p_x) as the content eligibility metric at a node. It is also to note that the second term is assigned zero for content $x \in \mathbb{X}$ not cached at the node v at a particular location l_a and time slot $\overline{t_b}$ to avoid unnecessary content eligibility computation.

6.3.3 Node Eligibility

We believe content caching should be made at social-aware mobile nodes identified in the network using centrality measures. The idea is more precisely to classify a node for content caching in an urban environment. However, it is not possible to compute typical centrality measures (Degree, Closenesss, Betweenness, Eigenvector) in dynamic network topology for urban mobile nodes. Furthermore, existing schemes follow a network-centric approach in order to analyze the network, thus, ignoring the content importance with respect to the user relevance.

Therefore, we use novel set of centrality schemes allowing a node to decide its local and global caching capability. Local importance measures a node's eligibility for short-term content distribution based on the cached content importance as a measure of its eligibility, its connectivity and spatio-temporal availability in the network. On the other hand, global importance enables the node to facilitate long-term content availability in the network based on its overall reachability and mobility pattern in the urban environment. We describe below the two centrality metrics used in order to decide a node's eligibility to be selected as information hub in the network.

6.3.3.1 Local Centrality

Using the nodes contact frequency and duration to decide its capability to cache in the network is a challenging task because of the rapid changes in the time evolving network topology. To overcome this, we propose the use of a novel social aware metric, which simultaneously considers three essential parameters, the content user relevant importance, the node spatio-temporal availability and its network connectivity [18]. The user's interest satisfaction frequency discussed before for the cached content is a key metric for a node's eligibility as it regularly responds to user interests. The interests are assumed to be generated and received from the neighboring nodes using multi-hop interest forwarding. We consider the following local parameters known to the node for analytically finding its importance:

- Content importance: content importance measures the node relevance to users for a particular cached content, i.e. the interest-response frequency is a vital

factor to classify a content's importance.

- Spatio-temporal availability: It reflects the social-behavior based on the node's habitual routes. Spatial availability reflects the node's recursive presence in a location, while temporal availability refers to its relevance in time for a location.
- Neighborhood importance: Neighborhood importance shows the node topological connectivity in order to be capable to offer cache accessible to nodes. An easily reachable and well connected node in a network topology can act as an efficient information hub in an urban environment.

The nodes first classify the cached content taking into consideration its relevance to the users interest using the relation Π_v^x in Equation 6.1. It then considers the associated content eligibility to find its relative importance in the network using the local centrality. Thus, it determines a node eligibility as a local information hub responsible for the efficient content caching:

$$\begin{aligned} LC_v(t_{k+1}) = \theta \times LC_v(t_k) + (1-\theta) \times [\alpha f_I^v(t_{k+1}) \\ +\beta f_{L,T}^v(t_{k+1}) + \gamma f_\Gamma^v(t_{k+1})] \end{aligned} \tag{6.2}$$

where f_I^v, $f_{L,T}^v$ and f_Γ^v are the importance functions for the cached content, node's spatio-temporal availability and its neighborhood, respectively. Each function's contribution is normalized by the terms α, β and γ, where $\alpha + \beta + \gamma = 1$, where $\theta \in [0, 1]$ allows the node to increase its centrality more or less rapidly with respect to the previous time-slot. The impact of each parameter differs with respect to different applications. For example, if the node is located in a better connected neighborhood, it can easily spread information. Therefore, the corresponding node weights the content user-relevant importance along the neighborhood more than the spatio-temporal availability function.

6.3.3.2 Global Centrality

Inspired from the concept of communicability in complex networks [10], GRank [21], a global centrality scheme allows a node to use a new stable metric named "Information communicability" to rank different content in an urban environment and rank itself accordingly. Using GRank, the node finds each content reachability and popularity at different locations taking into consideration the user interest satisfaction. It also considers its mobility pattern between different locations in the city along its availability in each location. Nodes available in popular locations in the city qualify as important information hubs with higher global centrality score in the network.

For a node v, the global centrality $GC_v(t_{k+1})$ for the next time instant t_{k+1} is updated as the Exponential Weighted Moving Average (EWMA) function of the current and previous global centrality as shown in the relation below:

$$GC_v(t_{k+1}) = \theta \times GC_v(t_k) + (1-\theta) \times f_G^v(t_{k+1}), \tag{6.3}$$

where $\theta \in [0, 1]$ is the tuning parameter which allows the node to adjust its importance with respect to the previous time-slot, $GC_v(t_k)$ is the node's global centrality at the beginning of the current time-slot and $f_G^v(t_{k+1})$ is the node's global centrality at the end of the current time slot t_k.

Combined Local and Global Centrality

In this section we propose to combine the local and global centrality. Thus, the combined node centrality C_v for the next time slot can be represented as:

$$C_v(t_{k+1}) = \varsigma \times LC_v(t_{k+1}) + (1 - \varsigma) \times GC_v(t_{k+1}),$$

where, LC_v and GC_v expressed in Equation 6.2 and 6.3 are the node's local and global centrality, respectively. The rationale behind considering both centralities can be explained by two facts. First, we need to classify the node short term eligibility for spatio-temporally caching and accessing relevant content locally, closer to users. Second, we need to assess the node reachability with respect to different urban neighborhoods taking into account the content availability and popularity in an urban environment.

Node Utility

The node utility is composed of its centrality, storage capacity size and the associated cost proportional to the storage space provided to the service provider. The corresponding buffer space available at each node is represented by $B_v = B_v^t - F_v$, where B_v^t is the total storage space at the node and F_v is the occupied space. Since we are interested in the buffer size each individual node can spare for caching, we consider, $B_v(l_a, \bar{t}_b)$ as the amount of storage buffer offered by the node at location l_a and time slot $\bar{t}_b$. It is to note that caching more popular content yields more cost since such content are to be constantly placed in cache for potential user interests on longer timespan ($\tau = 1$). We define the node cost function as:

$$f_v^C(l_a, \bar{t}_b) = \frac{1}{|X|} \sum_x \Pi_v^x(l_a, \bar{t}_b) \times S_x + C'$$

where $\Pi_v^x(l_a, \bar{t}_b)$ models the content profile by considering both its spatio-temporal availability and popularity, S_x is the percentage of storage space occupied by the content and C' comprises additional costs such as mobility, energy consumption etc.

We assume a monetary reward is paid to the node contributing its cache besides the natural incentive for the node to be itself interested in the content it cached. The function $f_v^R(l_a, \bar{t}_b) = B_v(l_a, \bar{t}_b) \times R_B$ rewards the node proportionally to the offered buffer space B_v at location l_a and time slot $\bar{t}_b$, where R_B is the reward paid per storage unit dedicated by the service provider for each node. We derive the node's utility as:

$$U_v(l_a, \bar{t}_b) = C_v(t_k) \times f_v^R(l_a, \bar{t}_b) - f_v^C(l_a, \bar{t}_b)$$

where C_v is the node's centrality score at the latest time slot, $f_v^R(l_a, \bar{t}_b)$ and $f_v^C(l_a, \bar{t}_b)$ are the node's reward and cost functions, respectively. The node utility is proportional to centrality in order to consider the best suitable candidates to manage spatio-temporally content caching in the network.

6.4 Distributed Fogs Formation as Coalition Game

As explained before, the nodes collaborate to offer their buffer to the service provider in order for it to cache content. This is achieved by forming mobile fogs for distributed

resources (such as storage in our case) offered by set of co-located nodes with similar interests aiming to maximize their utility. To do this, we use game theoretic concept of coalition formation where the service provider optimally selects the best coalitions formed by nodes in all areas at all time slots. The coalition game $G(\mathbb{V}, U)$ for the set of nodes $\mathbb{V}$ targets to form the set of coalitions $S = \{s\}$, $s \subseteq \mathbb{V}$ as the respective mobile fogs. The real-valued function $U : 2^{|\mathbb{V}|} \rightarrow \mathbb{R}$ associates to each coalition s a value $U_s(l, t)$ as the total payoff available to players in coalition s at the location l and time slot $\bar{t}$ in an urban environment. We define below the preference relation for the nodes preferences to merge into a particular coalition yielding a larger amount of utility. The service provider use the same preference relation to select the best coalition of nodes offering the maximum utility among the set of coalitions formed in the network at different locations and time slots.

Definition 6

(Coalitions Preference Relation) For any node, v, a preference relation $\succ$ is defined as a complete, reflexive and transitive binary relation over the set of all coalitions that node v can possibly form $\{s \subseteq \mathbb{V} : v \in s\}$.

Each node prefer to merge to a coalition with larger utility, thus increasing its chance to be selected by the service provider. For example, given two coalitions s_1 and s_2 for the node v to join, $s_1, s_2 \subseteq \mathbb{V}$, such that $v \in s_1$ and $v \in s_2$, $s_1 \succ s_2$ indicates that node v prefers to merge with s_1 over s_2. The node adds its utility to the coalition it joins, thus increasing the coalition utility.

It is also to note that nodes prefer to merge into coalitions with larger utility formed by high centrality nodes in order to gain more payoff proportional to the overall storage offered by the coalition. Node utility is based on two factors, its individual payoff proportional to its centrality, offered storage and associated cost and its payoff in the coalition proportional to the overall centrality, collaborative storage, and cost offered by the set of nodes in the coalition. The optimization problem can thus be formulated as follows to choose the coalition (form mobile fogs) offering the maximum utility over preferences.

$$
\begin{aligned}
\underset{S}{\text{maximize}} \quad & \sum_{s \in S} U_s(l, \bar{t}) \\
\text{subject to} \quad & \sum_{v \in \mathbb{V}} f_v^C(l, \bar{t}) \leq f_v^{C_{th}}(l, \bar{t}), \forall l, \forall \bar{t} && (\mathrm{C}_1) \\
& \sum_{v \in s} A_v^x(l, \bar{t}) \geq A_v^{x_{\mathrm{th}}}(l, \bar{t}), \forall s, \forall l, \forall \bar{t} && (\mathrm{C}_2) \\
& f_v^R(l, \bar{t}) \geq f_v^{R_{\mathrm{th}}}(l, \bar{t}), \forall v && (\mathrm{C}_3) \\
& \max_{v \in \mathbb{V}} B_v(l, \bar{t}) \leq B_v^{\mathrm{t}}(l, \bar{t}), \forall v && (\mathrm{C}_4)
\end{aligned}
$$

The objective function maximizes the utility of coalitions motivating nodes to merge into a coalition with larger utility. The first constraint C_1 addresses the cost limitations where the cost paid to nodes as incentives is limited by a threshold cost a service provider can afford. The constraint C_2 ensures that the content availability by each coalition at each location and time should be more than a desired threshold specified by the service provider. It allows fairness by specifying a minimum content availability

requirements at all locations and timings. Constraints C_3 and C_4 deal with the node individual resources, where, for all nodes, (i) the reward offered to the node for providing its storage should be greater than the associated cost and (ii) the buffer offered by the node should not exceed the total buffer size available at the node.

6.4.1 Solution: The Core

The target is to motivate each node to participate towards forming the grand coalition of all nodes also known as the core of the game i.e. the node receives

$$U_v(l,\bar{t}) = \begin{cases} U_v(l,\bar{t}), & \text{if joins coalition} \\ 0, & \text{otherwise} \end{cases}$$

To do so, we first define for each location and time slot an imputation $y = (y_1, y_2..., y_M)$, $M = |\mathbb{V}|$ as the distribution of the total payoff/reward among nodes such that (i) $y_v \geq U(\{v\}), \forall v$, i.e., a node receives at least as much as it could obtain on its own, without cooperating with anyone else (individual rationality). (ii) $\sum\limits_{v} y_v = U(M)$ (efficiency). The core is the set of imputations y such that $y_s = \sum\limits_{v \in s} y_v \geq U_s, \forall s \subseteq \mathbb{V}$ and equality if $s = \mathbb{V}$.

Non-emptiness and Uniqueness of the Core

In order to derive the non-emptiness and uniqueness of the core, we define a set of weights w_s for coalitions, where $0 \leq w_s \leq 1, \forall s \subseteq \mathbb{V}$, is a balancing set of weights if $\forall v \in \mathbb{V}$, $\sum\limits_{s, v \in s} w_s = 1$.

Definition 7

(Balanced Coalition Game) The coalitional game $(\mathbb{V}, U)$ is balanced if and only if, for every balancing set of weights w, we have $\sum\limits_{‰0 \neq s \subseteq \mathbb{V}} w_s \cdot U_s \leq U_{\mathbb{V}}$. We consider the weights for each coalition as the percentage of total nodes in the coalition. Then, we check whether the game is balanced by finding the condition in Definition 7. Thus, in a balanced game, any arrangement feasible for the different coalitions is feasible for the grand coalition as well. We can now state the Bondareva-Shapley Theorem as:

Theorem 1

(Bondareva-Shapley Theorem) The coalitional game $(\mathbb{V}, U)$ has a non-empty core if and only if it is balanced.

Proof

The non-emptiness and uniqueness of the core can be proven using the content availability constraint C_2, where the service provider selects a single coalition per location and time to ensure spatio-temporal fairness i.e $A_v^{x_{\text{th}}}(l,\bar{t}) = 1, \forall l \forall \bar{t}$. For each location and time ${}^{\mathbb{V}}C_s$ combinations of nodes are possible in a coalition. The combinations

of coalitions for all locations at all timings can be given as $\prod_{l}\prod_{\bar{t}}{}^{\mathbb{V}}C_s(l,\bar{t})$. The service provider fairness requirements for a single coalition i.e. ${}^{\mathbb{V}}C_s = 1, \forall l, \forall \bar{t}$, results in $\prod_{l}\prod_{\bar{t}}{}^{\mathbb{V}}C_s(l,\bar{t}) = 1$. Thus, we are able to reduce 2^M possible combinations for coalitions to a single coalition, i.e. $1 << 2^M$, as the “grand" coalition.

6.4.2 Algorithm: Optimal Selection among Coalitions

Algorithm 7 summarize the coalition formation for spatio-temporally correlated nodes with similar social interests offering their storage resource as a mobile fog in an urban environment. For a given set of nodes $\mathbb{V}$, it returns the selected coalitions set $S \subseteq \mathbb{V}$.

First, we initialize the coalitions to empty set. For each location and time slot, we find the node $v' \in \mathbb{V}$ offering the best utility as the node with the highest centrality. The node is merged to the coalition subset $s \subset S$ as in Line 5 and 6, its associated utility and cost (Lines 7 and 11) is added to the corresponding coalition utility and cost, respectively. We also add the content availability at v' for each content x as its availability for the coalition s as stated in Lines $8-10$. Line 12 ensures the cost of nodes merged into coalition do not exceeds a threshold dedicated by the service provider. Moreover, the content availability for all content should also meet certain requirements for different areas and time slots. For the rest of the nodes (except most central one), if for a node $u' \in \mathbb{V}$, the coalition s results in higher utility than the node by itself without surpassing buffer limits, the node merges into the coalition and adds its utility to the coalition as in Lines $14-16$. For each content x cached at node u', the content availability (Lines $17-19$) and its associated cost (Line 20) is updated at the set s. Finally, the set of coalitions S in Line 25 returns the coalitions for all locations and time slots.

6.5 Performance Evaluation

We evaluate the proposed coalition game model using ns-3 where the named-data networking model of the CCN architecture is implemented using ndnSIM [1]. The urban vehicle mobility is extracted from a realistic model for large scale mobility in Cologne, Germany. A total of 2986 vehicles are used to validate the scalability of our urban caching approach at mobile fogs. The analysis is performed by dividing an area of $6X6km^2$ into 36 uniform neighborhoods each of $1X1km^2$ with a time granularity of 10 minutes. The simulation parameters are summarized in Table 6.2, followed by a description of the simulation scenarios used for the performance evaluation.

6.5.1 Simulation Scenario

The simulation scenario implements two types of vehicles, the content producers as the source vehicles for a content and consumer vehicles interested in the content. Consumer vehicles generate interests for a pre-known content sequence following a Zipf distribution to model the content popularity profile allowing frequent interests for more popular content. Any producer vehicle already cached the content responds to the consumer interest. We allow intermediate vehicles to perform in-network caching

INPUT: Set of Nodes $\mathbb{V}$
OUTPUT: Selected coalition set $S \subseteq \mathbb{V}$
Initialize $S = \phi$, $s = \phi$,
for each location $l \in \mathbb{L}$, time-slot $\bar{t} \in T$ **do**
  $v' = \arg\max_{v \in \mathbb{V}} U_v(t_k)$
  $s = s \cup v'$
  $U_s \leftarrow U_s + U_{v'}$
  for each content x in v' **do**
    $A_s^x \leftarrow A_s^x + A_{v'}^{x_{\text{th}}}$
  end for
  $f_s^C \leftarrow f_s^C + f_{v'}^C$
  while $(f_s^C \leq f_s^{C_{th}})$ and $(A_v^x \geq A_v^{x_{\text{th}}}), \forall x$ **do**
    for each node u' , $u' \neq v'$ **do**
      if $U_s > U_{u'}$ and $B_{u'} \leq B_{u'}^{\text{t}}$ **then**
        $s = s \cup u'$
        $U_s = U_s + U_{u'}$
        for each content x in u' **do**
          $A_s^x \leftarrow A_s^x + A_{u'}^{x_{\text{th}}}$
        end for
        $f_s^C \leftarrow f_s^C + f_{u'}^C$
      end if
    end for
  end while
end for
Return $S = S \cup s$

Algorithm 7: Self-organizing Coalition Formation

Table 6.2: SAVING II - Simulation Parameters

Parameter	Value
Simulation platform	NS-3
Number of vehicles	2986
Mobility trace	Cologne, Germany
Area	$6X6km^2$ city center
Duration	1 hour
Communication range	150m
Path loss model	Nakagami + Log distance
Packet size	1024 bytes
Time granularity	1 sec
Simulation Runs	10

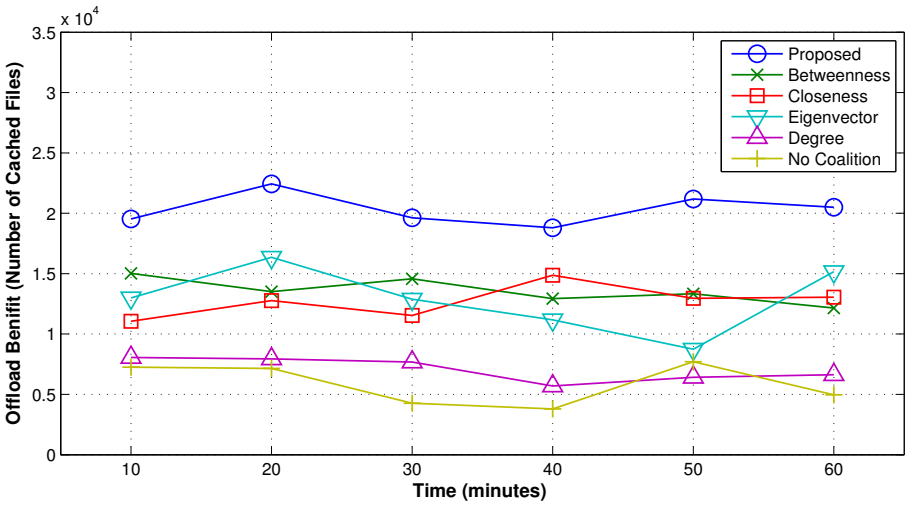


Figure 6.2: Total number of files cached using each scheme over an average of ten different simulations

where different buffer sizes are assigned to each vehicle. The content availability is modeled by randomly assigning it to each vehicle as its cached content. The content validity metric $\tau = [0, 1]$ is extracted from its popularity as well as availability profile tracing the generated interests.

For each vehicle, its centrality is computed using the proposed novel local and global centrality receptively. The tuning parameters α, β, γ are set to 0.33 and θ, ς are set to 0.5 in order to maintain generality in our evaluation, however, the significance of each parameter depends on the service provider requirements. Subsequently the vehicle utility is derived from its centrality, reward and cost extracted from its buffer space. Utilities for spatio-temporally co-located vehicles are computed towards the respective coalition formation. Algorithm 7 is implemented to find the best coalitions from the set of all coalitions formed in the network.

We perform each simulation up to ten times by analyzing different set of vehi-

cles as content caches in order to compute up to 95% confidence intervals.The fogs formed around the vehicles with utility for eligibility computed using our centrality are compared by implementing benchmark centrality schemes (Degree, Closeness, Betweenness, Eigenvector). We also evaluation the coalition formation by a comparison with an approach without coalition formation for the following performance metrics:

- Cache Hits: Percentage of time when content found in vehicle cache for all vehicles in the network
- Offload Benefit: The total amount of content files cached at vehicles to relieve infrastructure
- Spatio-temporal Coalitions: Graphical depiction of the mobile fogs formed in the network at different locations and timings

6.5.2 Simulation Results

6.5.2.1 Cache Hits

We evaluate the overall cache performance by finding the total number of cache hits by all the vehicles in the network. Figure 6.3 depict the cache hit rate in percentage by comparing different schemes for the vehicle utility. The analysis is performed for different time slot at 10 minutes granularity where the coalitions centered at the vehicles with utility calculated by our scheme outperforms existing schemes with around 60-85% cache hits. Moreover, the case of no coalition between vehicles resulted in the a poor performance with cache hits of less than 10% during the entire simulation. Existing social metrics such as typical centrality schemes yielded a cache hit rate of around 20-30%, thus, validating the efficiency of using the proposed centrality scheme.

We also observe an increase in the cache hits over time for our approach. It is because the coalition centered at the vehicles with high utility cache more content and subsequently satisfies more interests resulting in a higher cache hit rate. Such increase in not observed by the other schemes since each time slot at different location yields different set of vehicles, unreliable for content caching. This also validates the fact that coalition improves caching performance as all the cases where coalitions was formed resulted in higher cacher hits than the case where no coalition was formed.

6.5.2.2 Offload Benefit

The benefit of offloading at mobile fogs is validated by finding the total number of files cached in the mobile fogs. Figure 6.2 shows the offload benefit at different times in the simulation. It is clear that the total number of content files made available for the user by caching at vehicles by our approach is more than all existing schemes as well as the case of no coalition. The case of no coalition failed to cache files substantially in the network with five times lesser number of files , where other schemes cached half the amount of files compared to our approach. This proves that caching content at mobile fogs centered around high centrality vehicles identified by our approach maximize the overall content availability in the network. It is because such vehicles received, cached and thus satisfied more user relevant interests in the network.

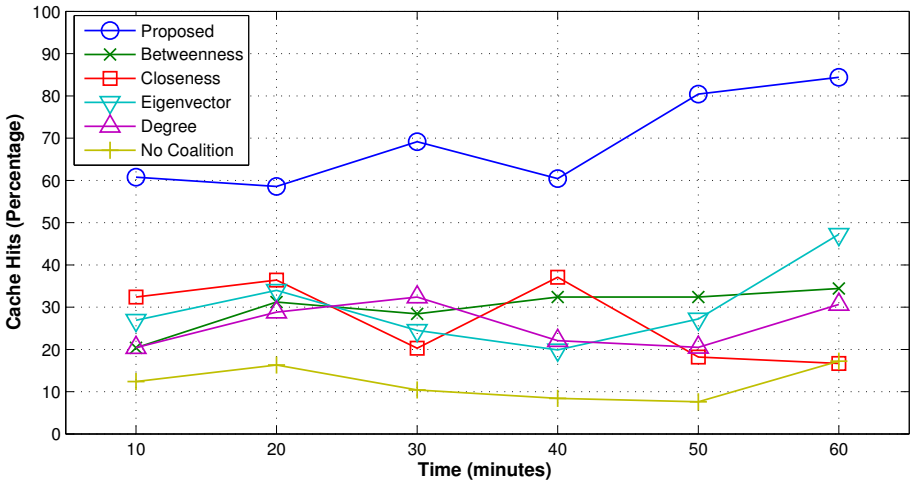
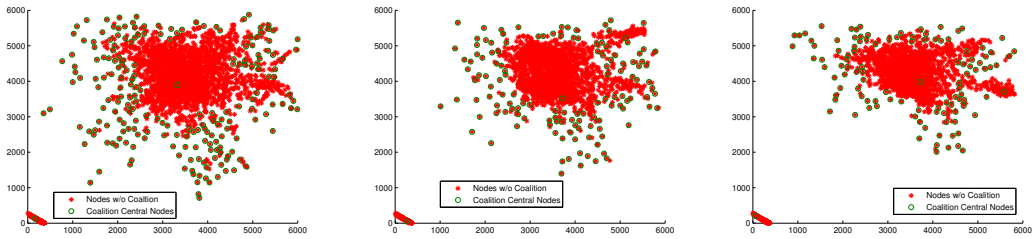


Figure 6.3: Average cumulative cache hit rate by all vehicles with utilities computed using each scheme in ten simulations



(a) Coalition formed at $t = 10$ minutes (b) Coalition formed at $t = 30$ minutes (c) Coalition formed at $t = 60$ minutes

Figure 6.4: Spatio-temporal Mobile Fogs for $6X6$ km area, central vehicles for each coalition identified in green

6.5.2.3 Spatio-temporal coalitions

It is important to evaluate the existence and uniqueness of the core in the proposed coalition game, therefore, Figure 6.4 shows the mobile fogs formed around 10 minutes, 30 minutes and 60 minutes into the simulation for the Cologne City center ($6km^2$). Individual vehicles without coalition are marked red '*' while coalitions are identified by the central vehicle of each coalition marked as green circle. First, we observe that all vehicles form coalition resulting in the existence of grand coalition as the core of the game. It is obvious from the fact that even the isolated vehicles Secondly, we noticed that most coalition central vehicles are located at the edge of the network. We also observe that a single central vehicle represents the coalition of the most denser part of the network. The network analysis at three different snapshots highlighted the benefit of considering spatio-temporal factor in our approach as we can see different set of coalitions are formed at different timings. Thus, we can clearly observe the union of smaller coalition forming the grand coalition in the network.

6.5.3 Summary of Findings

Results shown above depict that cache management in an urban environment can be efficiently managed by forming mobile fogs allowing nodes to form coalitions for resource pooling to offload content. It is better to dedicate “high centrality" nodes for distributed management of such coalitions in order to reduce overhead at the network infrastructure. As indicated in Figure 6.4, it is important for individual nodes to consider the spatio temporal content profile while considering both its availability and popularity towards caching decisions. We observe that in case there are no other nodes in the vicinity, the node can also form self-coalition. We also found that coalition central node reside at the edge of the network which simultaneously reduce network traffic and improve user quality of experience. Thus, our analysis based on results from scalable simulation scenarios clearly validates our proposition that offloading content at mobile fogs in the form of coalitions at the network edge (closer to urban consumer) substantially benefits the infrastructure network.

6.6 Conclusions

Efficient content caching at vehicles requires an optimal cache management policy. There exist different cache management policies for individual nodes, however there is no work to address optimal content caching using a group of collaborating vehicles. We modeled the collaborative content caching using vehicles as a coalition game. More precisely, this chapter proposed a coalition game for social-aware collaborative content caching at nodes forming a “Mobile Fog" in an urban environment.

First we modeled the content profile by defining its spatio-temporal popularity and availability relation allowing each node to find the content eligibility, then finding its capability to cache content taking into consideration the offered resources and cost with respect to user interest profiles. The best co-located candidates self-organize into a coalition to share storage to offload content close to mobile users of similar social interests. We performed scalable simulations for an interesting case of vehicles as urban mobile caches on a realistic trace. Results shown that the nodes forming coalition based on our criteria offload twice more content and achieved 60% cache hits compared to other approaches yielding around 25% cache hits in the network.

Caching at urban mobile nodes need further investigations specially exploiting vehicles as urban information hubs. Moreover, our future work includes the study of content and node profile towards efficient content distributed in highly dynamic networks with multiple content providers while avoiding duplicate content downloads in the network.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

We presented an alternate solution to offload network data using smart vehicles with their distributed caching, computing and communicating capabilities to facilitate content availability for an urban mobile user. To do this, we proposed VISIT for data collection and SAVING as a data storage system at vehicles in urban environment. Both VISIT and SAVING exploit the information centric networking architecture and are based on the social interest of large number of users. The Vehicular Information-centric Socially Inspired Telematics (VISIT) address the identification and selection of the minimum set of socially important vehicles for urban data collection using innovative centrality metrics, " InfoRank" and "CarRank", where smart vehicles autonomously rank themselves based on user relative importance. The vehicle considers the importance of location-aware information, its neighborhood topology along its spatio-temporal availability to find its importance in the network. The minimum number of best ranked vehicles required for urban data collection are then optimally selected to achieve a desired coverage while avoiding redundancy within a limited budget. From the obtained results, it is clear that the vehicles identified and selected by proposed novel centrality metrics InfoRank and CarRank along the vehicle selection algorithms using ROVERS satisfied more user interests and yielded more throughput than other centrality schemes and algorithms.

SAVING is a Socially-Aware Vehicular Information-centric Networking solution targeting content caching at vehicles in an urban environment closer to the mobile user. We first defined a content spatio-temporal availability-popularity relation to decide a content importance in order to be cached at a vehicle. An new vehicle centrality metric "GRank" is then proposed for a vehicle to classify its eligibility as an efficient urban content cache in the network by first ranking different urban neighborhoods and then rank itself based on its reachability in the network. Then, an incentive driven social welfare game is formulated to fairly select the best eligible candidates to cache content for different urban neighborhoods catering individual rationality and content availability requirements under a given budget. Finally, a coalition game is proposed to form mobile fogs for collaborative caching at spatio-temporally co-located vehicles where service provider selects coalitions as urban caches to maximize content availability. Results have depicted that the vehicles identified by GRank as information

hubs selected using the social welfare optimization game along the vehicles identified by combining CarRank and GRank using the proposed coalition game yielded the best cache hit rate compared to all other schemes. Thus, proving VISIT and SAVING together to be an efficient data collection and storage system.

Our work aimed to focus the research community interest towards the application of combining socially aware content distribution scheme with the information-centric networking paradigm. Our findings explored possible solutions to the fundamental problem of where, what and how to collect and cache content in urban mobile networks under an increasing growth of mobile traffic.

7.2 Future Work

Our work opens path towards the development of applications based on the novel idea of social-aware urban data collection and storage applications for smart cities. We are looking forward to future research designs where the set of vehicles with high centrality score can collaborate for the efficient collection, storage and distribution of content for urban mobile users. Such socially important vehicles can be also useful in novel location-based services.

The data collection can be enhanced by coupling with latest development with Augmented Reality (AR) based applications where video feeds of multiple vehicles can be aggregated to collect real time video of an urban environment. Such data aggregation can be achieve either collaboratively between vehicles in a distributed fashion or in a centralized approach.

Caching at urban mobile vehicles need further investigations specially exploiting vehicles as urban information hubs. Future improvements in content caching aims to propose intelligent algorithms for vehicles to make autonomous and real-time decisions regarding the cached content. There is a need to develop efficient distributed cache management schemes with collaborative content replacement strategies with redundancy avoidance.

Moreover, our future work includes the study of content and vehicle profile towards efficient content distribution in highly dynamic networks with multiple content providers while avoiding duplicate content downloads in the network.

Open content distribution for future research issues includes also the development of novel efficient social aware routing strategies as well as flexible and scalable naming scheme in order to allow introducing novel applications. An example of such applications is the support of the high bandwidth consuming video streaming applications in content-centric mobile networks. Thus, we propose to explore the new trends exploiting socially-aware network computing, caching and communication in a content-centric approach to overcome the limitations of the existing connection-centric approach in front of the above mentioned research.

List of Figures

1.1 System Overview 17

3.1 Network Model 29
3.2 An example of vehicle city-wide coverage scope as a metric of its mobility pattern 33
3.3 Spatio-temporal availaibility in the same time-slot 37
3.4 Neighborhood Centrality Exchange 38
3.5 Comparison of the Cumulative Satisfied Interests by the top identified vehicles using InfoRank and existing centrality schemes 44
3.6 Cumulative Satisfied Interests by top identified vehicles using CarRank and existing schemes over an average of ten different simulation scenarios 45
3.7 Temporal snapshots after each 10 minutes comparing top identified nodes by each schemes - InfoRank 46
3.8 Temporal Snapshots of comparing top nodes identified by all schemes (CarRank, Degree, Closeness, Betweenness, Eigenvector centrality) . . . 47
3.9 Comparison of the average aggregated per node throughput achieved by the top identified vehicles using each centrality scheme 49
3.10 Average aggregated throughput by the top identified nodes using each scheme in ten simulations 50
3.11 Comparison of the average cumulative cache hit rate at the top identified vehicles using each centrality scheme 51
3.12 Average cumulative cache hit rate by the top identified nodes using each scheme in ten simulations 52

4.1 Temporal analysis of the average CSI by the optimized set of IFVs selected using each algorithm 61
4.2 Temporal analysis of the average CSI by the optimized set of IFVs selected using each centrality scheme 62
4.3 Temporal analysis of the throughput achieved by the optimized set of IFVs selected using each algorithm 63
4.4 Temporal analysis of the throughput achieved by the optimized set of IFVs selected using each centrality scheme 63

5.1 Information Walk 70

5.2 Cumulative Satisfied Interests by top identified vehicles using GRank compared with exisiting centrality schemes over an average of ten different simulations . . . 80
5.3 Temporal Snapshots after each 15 minutes comparing top identified nodes by each schemes . . . 81
5.4 Average aggregated throughput by the top identified nodes using each scheme in ten simulations . . . 82
5.5 Average cumulative cache hit rate by the top identified nodes using each scheme in ten simulations . . . 83
5.6 Average success rate achieved by the selected set of vehicles (SWO: Social Welfare Optimization, RA: Reputation Aware, MB: Minimize Budget, MC: Maximize Coverage, BMC: Budgeted Maximum Coverage) 85
5.7 Average aggregated throughput achieved by the selected set of vehicles (SWO: Social Welfare Optimization, RA: Reputation Aware, MB: Minimize Budget, MC: Maximize Coverage, BMC: Budgeted Maximum Coverage) . . . 86
5.8 Average cache hit rate achieved by the selected set of vehicles (SWO: Social Welfare Optimization, RA: Reputation Aware, MB: Minimize Budget, MC: Maximize Coverage, BMC: Budgeted Maximum Coverage) 87

6.1 Content validity hint indicated by user interests . . . 95
6.2 Total number of files cached using each scheme over an average of ten different simulations . . . 103
6.3 Average cumulative cache hit rate by all vehicles with utilities computed using each scheme in ten simulations . . . 105
6.4 Spatio-temporal Mobile Fogs for $6X6$ km area, central vehicles for each coalition identified in green . . . 105

List of Tables

3.1 VISIT I - List of Notations 30
3.2 Simulation Parameters for VISIT 40
3.3 InfoRank in different set of Simulations 41
3.4 CarRank in different set of Simulations 41
3.5 Cumulative Satisfied Interests by top vehicles identified by each scheme 44

4.1 VISIT II - List of Notations 57

5.1 SAVING I - List of Notations 69
5.2 SAVING I - Simulation Parameters 78
5.3 GRank in different set of Simulations 79

6.1 SAVING II - List of notations 93
6.2 SAVING II - Simulation Parameters 103

Bibliography

[1] Alexander Afanasyev, Ilya Moiseenko, and Lixia Zhang. ndnSIM: NDN simulator for NS-3. Technical Report NDN-0005, NDN, October 2012.

[2] Bengt Ahlgren, Christian Dannewitz, Claudio Imbrenda, Dirk Kutscher, and Börje Ohlman. A survey of information-centric networking. *Communications Magazine, IEEE*, 50(7):26–36, 2012.

[3] Marica Amadeo, Claudia Campolo, Antonella Molinaro, and Giuseppe Ruggeri. Content-centric wireless networking: A survey. *Computer Networks*, 72:1–13, 2014.

[4] Marco Valerio Barbera, Aline Carneiro Viana, Marcelo Dias de Amorim, and Julinda Stefa. Data offloading in social mobile networks through vip delegation. *Ad Hoc Networks*, 19:92–110, 2014.

[5] César Bernardini, Thomas Silverston, and Olivier Festor. Socially-aware caching strategy for content centric networking. In *Networking Conference, 2014 IFIP*, pages 1–9. IEEE, 2014.

[6] César Bernardini, Thomas Silverston, and Olivier Festor. A comparison of caching strategies for content centric networking. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.

[7] Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[8] Raffaele Bruno and Maddalena Nurchis. Efficient data collection in multimedia vehicular sensing platforms. *Pervasive and Mobile Computing*, 2014.

[9] Wei Koong Chai, Diliang He, Ioannis Psaras, and George Pavlou. Cache "less for more" in information-centric networks. In *International Conference on Research in Networking*, pages 27–40. Springer, 2012.

[10] Ernesto Estrada and Naomichi Hatano. Communicability in complex networks. *Physical Review E*, 77(3):036111, 2008.

[11] Mario Gerla, Eun-Kyu Lee, Giovanni Pau, and Uichin Lee. Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds. In *Internet of Things (WF-IoT), World Forum on*, pages 241–246. IEEE, 2014.

[12] Negin Golrezaei, Karthikeyan Shanmugam, Alexandros G Dimakis, Andreas F Molisch, and Giuseppe Caire. Femtocaching: Wireless video content delivery through distributed caching helpers. In *INFOCOM, 2012 Proceedings IEEE*, pages 1107–1115. IEEE, 2012.

[13] Sherin Abdel Hamid, Hatem Abouzeid, Hossam S Hassanein, and Glen Takahara. Optimal recruitment of smart vehicles for reputation-aware public sensing. In *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*, pages 3160–3165. IEEE, 2014.

[14] Bo Han, Pan Hui, VS Anil Kumar, Madhav V Marathe, Jianhua Shao, and Aravind Srinivasan. Mobile data offloading through opportunistic communications and social participation. *IEEE Transactions on Mobile Computing*, 11(5):821–834, 2012.

[15] Zongjian He, Jiannong Cao, and Xuefeng Liu. High quality participant recruitment in vehicle-based crowdsourcing using predictable mobility. In *Computer Communications (INFOCOM), 2015 IEEE Conference on*, pages 2542–2550. IEEE, 2015.

[16] Pan Hui, Jon Crowcroft, and Eiko Yoneki. Bubble rap: Social-based forwarding in delay-tolerant networks. *Mobile Computing, IEEE Transactions on*, 10(11):1576–1589, 2011.

[17] Bret Hull, Vladimir Bychkovsky, Yang Zhang, Kevin Chen, Michel Goraczko, Allen Miu, Eugene Shih, Hari Balakrishnan, and Samuel Madden. Cartel: a distributed mobile sensor computing system. In *Proceedings of the 4th international conference on Embedded networked sensor systems*, pages 125–138. ACM, 2006.

[18] Junaid Ahmed Khan and Yacine Ghamri-Doudane. Carrank: An information-centric identification of important smart vehicles for urban sensing. In *14th International Symposium on Network Computing and Applications , NCA 2015 Proceedings IEEE.*

[19] Junaid Ahmed Khan and Yacine Ghamri-Doudane. Saving: Socially aware vehicular information-centric networking. *IEEE Communications Magazine*, 54(8), 2016.

[20] Junaid Ahmed Khan, Yacine Ghamri-Doudane, and Dmitri Botvich. Inforank: Information-centric autonomous identification of popular smart vehicles. In *Vehicular Technologoy Conference VTC Fall, 2015 Proceedings IEEE.*

[21] Junaid Ahmed Khan, Yacine Ghamri-Doudane, and Dmitri Botvich. Grank-an information-centric autonomous and distributed ranking of popular smart vehicles. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–7. IEEE, 2015.

[22] Samir Khuller, Anna Moss, and Joseph Seffi Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[23] Samir Khuller, Anna Moss, and Joseph Seffi Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[24] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.

[25] Swarun Kumar, Lixin Shi, Nabeel Ahmed, Stephanie Gil, Dina Katabi, and Daniela Rus. Carspeak: a content-centric network for autonomous driving. *ACM SIGCOMM Computer Communication Review*, 42(4):259–270, 2012.

[26] Tuan Le, You Lu, and Mario Gerla. Social caching and content retrieval in disruption tolerant networks (dtns). In *Computing, Networking and Communications (ICNC), 2015 International Conference on*, pages 905–910. IEEE, 2015.

[27] Uichin Lee, Biao Zhou, Mario Gerla, Eugenio Magistretti, Paolo Bellavista, and Antonio Corradi. Mobeyes: smart mobs for urban monitoring with a vehicular sensor network. *Wireless Communications, IEEE*, 13(5):52–57, 2006.

[28] Jun Li, He Chen, Youjia Chen, Zihuai Lin, Branka Vucetic, and Lajos Hanzo. Pricing and resource allocation via game theory for a small-cell video caching system. *IEEE Journal on Selected Areas in Communications*, 2016.

[29] Yong Li, Depeng Jin, Zhaocheng Wang, Lieguang Zeng, and Sheng Chen. Coding or not: Optimal mobile data offloading in opportunistic vehicular networks. *IEEE Transactions on Intelligent Transportation Systems*, 15(1):318–333, 2014.

[30] Yong Li, Mengjiong Qian, Depeng Jin, Pan Hui, Zhaocheng Wang, and Sheng Chen. Multiple mobile data offloading through disruption tolerant networks. *IEEE Transactions on Mobile Computing*, 13(7):1579–1596, 2014.

[31] Michele Mangili, Fabio Martignon, Stefano Paris, and Antonio Capone. Bandwidth and cache leasing in wireless information centric networks: a game theoretic study. *IEEE Transactions on Vehicular Technology*, 2016.

[32] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. *Spatial tessellations: concepts and applications of Voronoi diagrams*, volume 501. John Wiley & Sons, 2009.

[33] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[34] Filippo Rebecchi, Marcelo Dias De Amorim, Vania Conan, Andrea Passarella, Raffaele Bruno, and Marco Conti. Data offloading techniques in cellular networks: a survey. *IEEE Communications Surveys & Tutorials*, 17(2):580–603, 2015.

[35] Sasank Reddy, Deborah Estrin, and Mani Srivastava. Recruitment framework for participatory sensing data collections. In *Pervasive Computing*, pages 138–155. Springer, 2010.

[36] John Scott. *Social network analysis*. Sage, 2012.

[37] Pavlos Sermpezis and Thrasyvoulos Spyropoulos. Inferring content-centric traffic for opportunistic networking from geo-location social networks. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2015 IEEE 16th International Symposium on a*, pages 1–6. IEEE, 2015.

[38] A Socievole, E Yoneki, F De Rango, and J Crowcroft. Ml-sor: Message routing using multi-layer social networks in opportunistic communications. *Computer Networks*, 81:201–219, 2015.

[39] Vasilis Sourlas, Lazaros Gkatzikis, Paris Flegkas, and Leandros Tassiulas. Distributed cache management in information-centric networks. *IEEE Transactions on Network and Service Management*, 10(3):286–299, 2013.

[40] Peyman TalebiFard, Victor CM Leung, Marica Amadeo, Claudia Campolo, and Antonella Molinaro. Information-centric networking for vanets. In *Vehicular ad hoc Networks*, pages 503–524. Springer, 2015.

[41] Saran Tarnoi, Kalika Suksomboon, Wuttipong Kumwilaisak, and Yusheng Ji. Performance of probabilistic caching and cache replacement policies for content-centric networks. In *39th Annual IEEE Conference on Local Computer Networks*, pages 99–106. IEEE, 2014.

[42] Sandesh Uppoor, Oscar Trullols-Cruces, Marco Fiore, and Jose M Barcelo-Ordinas. Generation and analysis of a large-scale urban vehicular mobility dataset. *Mobile Computing, IEEE Transactions on*, 13(5):1061–1075, 2014.

[43] Anna Maria Vegni and Valeria Loscri. A survey on vehicular social networks. *Communications Surveys & Tutorials, IEEE*, 17(4):2397–2419, 2015.

[44] Hongjian Wang, Yanmin Zhu, and Qian Zhang. Compressive sensing based monitoring with vehicular networks. In *INFOCOM, 2013 Proceedings IEEE*, pages 2823–2831. IEEE, 2013.

[45] Ning Wang and Jie Wu. Opportunistic wifi offloading in a vehicular environment: Waiting or downloading now? In *Proc. of the 35th IEEE International Conference on Computer Communications (IEEE INFOCOM 2016)*, 2016.

[46] Yonggong Wang, Zhenyu Li, Gareth Tyson, Steve Uhlig, and Gaogang Xie. Design and evaluation of the optimal cache allocation for content-centric networking. *IEEE Transactions on Computers*, 65(1):95–107, 2016.

[47] Yonggong Wang, Zhenyu Li, Gareth Tyson, Steve Uhlig, and Gaogang Xie. Design and evaluation of the optimal cache allocation for content-centric networking. *IEEE Transactions on Computers*, 65(1):95–107, 2016.

[48] Kaimin Wei, Xiao Liang, and Ke Xu. A survey of social-aware routing protocols in delay tolerant networks: Applications, taxonomy and design-related issues. *Communications Surveys & Tutorials, IEEE*, 16(1):556–578, 2014.

[49] George Xylomenos, Xenofon Vasilakos, Christos Tsilopoulos, Vasilios A Siris, and George C Polyzos. Caching and mobility support in a publish-subscribe internet architecture. *IEEE Communications Magazine*, 50(7):52–58, 2012.

[50] Shanhe Yi, Cheng Li, and Qun Li. A survey of fog computing: concepts, applications and issues. In *Proceedings of the 2015 Workshop on Mobile Big Data*, pages 37–42. ACM, 2015.

[51] Yu-Ting Yu, Francesco Bronzino, Ruolin Fan, Cedric Westphal, and Mario Gerla. Congestion-aware edge caching for adaptive video streaming in information-centric networks. In *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pages 588–596. IEEE, 2015.

[52] Yuanyuan Zeng and Deshi Li. A self-adaptive behavior-aware recruitment scheme for participatory sensing. *Sensors*, 15(9):23361–23375, 2015.

[53] Xu Zhang, Ning Wang, Vassilios G Vassilakis, and Michael P Howarth. Decision-making for in-network caching of peer-to-peer content chunks-an analytical modelling study. In *Network of the Future (NOF), 2014 International Conference and Workshop on the*, pages 1–5. IEEE, 2014.

[54] Dong Zhao, Huadong Ma, Liang Liu, and Xiang-Yang Li. Opportunistic coverage for urban vehicular sensing. *Computer Communications*, 60:71–85, 2015.

[55] Dong Zhao, Huadong Ma, Liang Liu, and Xiang-Yang Li. Opportunistic coverage for urban vehicular sensing. *Computer Communications*, 60:71–85, 2015.

[56] Xuejun Zhuo, Qinghua Li, Guohong Cao, Yiqi Dai, Boleslaw Szymanski, and Tom La Porta. Social-based cooperative caching in dtns: A contact duration aware approach. In *2011 IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems*, pages 92–101. IEEE, 2011.