



**HAL**  
open science

# ROVERS: Incentive-based Recruitment of Connected Vehicles for Urban Big Data Collection

Junaid A. Khan, Yacine Ghamri-Doudane

► **To cite this version:**

Junaid A. Khan, Yacine Ghamri-Doudane. ROVERS: Incentive-based Recruitment of Connected Vehicles for Urban Big Data Collection. IEEE Transactions on Vehicular Technology, In press. hal-01964240

**HAL Id: hal-01964240**

**<https://inria.hal.science/hal-01964240>**

Submitted on 21 Dec 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# ROVERS: Incentive-based Recruitment of Connected Vehicles for Urban Big Data Collection

Junaid Ahmed Khan\* and Yacine Ghamri-Doudane†

\*University of Memphis, Memphis, TN, USA

†L3i Lab, University of La Rochelle, France

junaid.khan@memphis.edu, yacine.ghamri@univ-lr.fr

**Abstract**—The growth in mobile devices results in constant generation and consumption of large amount of data by mobile users on the go which is unbearable by the current mobile networks in terms of cost and bandwidth. At the same time, the technological advancements in modern vehicles allow us to harness their computing, caching and communication capabilities to support various smart city applications. It is now possible to recruit set of connected vehicles to collect, store and share heterogeneous information regarding urban streets and facilitate citizens with different location-aware services. However, for a user to find and retrieve relevant content among the fleet of hundreds of vehicles on urban roads is challenging due to high mobility and intermittent connectivity. To address this, we propose ROVERS for a service provider to recruit the best set of vehicles to facilitate users on urban streets with different location-based applications. To identify the best vehicles, we first use a distributed ranking scheme *CarRank*, where the vehicle autonomously classifies itself as important with respect to urban users’ interest. Then, we present a centralized recruitment scheme, exploiting game-theory for the service provider to fairly and optimally select the best vehicles under desired coverage, redundancy and quality requirements. Comparative analysis after in-depth simulations using realistic mobility traces of 2,986 vehicles shows that the set of vehicles selected using ROVERS yield better results compared to other selection approaches.

**Index Terms**—Data Collection, Urban Sensing, Information-Centric Vehicular Networking, Connected Vehicles, Nash Equilibrium

## I. INTRODUCTION

Smart phones, tablets and sensor-equipped vehicles today entail a constant generation and consumption of massive internet traffic to provide diverse (Location Based Services) LBS applications. Mobile network operators are currently struggling to cope with the monetary and bandwidth requirements of such large scale ubiquitous data collection.

Mobile crowdsensing based solutions although assist different LBS application by participating to various data collection task the urban environment, but are constrained by their energy, storage and availability [1]. A promising solution is, therefore, to utilize smart vehicles with their relatively high processing, storing and communicating capabilities to supplement the infrastructure network and facilitate citizens with different location-aware services [2]. However, vehicles on the road constantly generating and consuming a tremendous amount of data (including multimedia [3]) 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 urban environment. 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.

The major problem is for a city administration or a service provider to efficiently find and provide incentives to the most adequate vehicles with respect to the user relevant data from the fleet of vehicles along with the underlying challenge of intermittent connectivity. Pre-advertising or broadcasting all the data from each vehicle would result in a massive advertising overhead and a redundant information storm within the network. Thus, the challenge is to find the set of appropriate vehicles at the right time and place and fairly recruit them for efficient urban sensory data collection, storage and distribution towards a cost-effective city-wide coverage under a given redundancy and quality requirements [4] [5]. We target the following questions:

- How to classify a vehicle as an eligible candidate for different urban sensory data collection, storage and distribution?
- How to fairly recruit the best vehicles among the eligible candidates under a given city-wide coverage, redundancy and quality requirement?

To address these questions, we believe only a set of vehicles can be considered eligible based on their daily commute (mobility patterns) and the user-relevant popularity of their frequently visited neighborhoods. Network analysis use different variants of well known centrality measures such as Degree, Closeness, Betweenness, Eigenvector centrality to classify important nodes in a graph (network). Such schemes work well in social networks analysis to analyze connectivity between static entities but are however, inapplicable to dynamic network topologies and are incapable to cope with intermittent connectivity and high mobility of vehicles.

Therefore, in this paper, we present ROVERS (**R**ecruitment of **O**ptimal Vehicles for **E**fficient **R**oad **S**ensing) with the aim to find and recruit important vehicles for different LBS applications. To do so, we first use our proposed vehicle ranking scheme *CarRank* to decide the vehicle eligibility [6]. *CarRank* is an autonomous ranking scheme for a vehicle to classify itself based on novel factors that are not affected by the intermittent connectivity challenge such as the popularity of information with respect to the users, the vehicle’s spatio-temporal availability and its urban neighborhood.

Once we are able to identify eligible vehicles, we devise a centralized algorithm for a service provider or urban municipality to fairly and optimally recruit a subset of these important vehicles for the efficient gathering, storing and publishing of location-aware content for a given coverage, redundancy and quality requirements (i.e. wide coverage, low redundancy). Moreover, we cater the existence of untrustworthy vehicles by devising a game between the service provider and vehicle and present a Pareto efficient solution for the money flow. Particularly, we derive the utilities for the service provider as well as the vehicle that ensure an adequate monetary reward is paid to the recruited vehicles proportional to the content they delivered to the users instead of their CarRank score.

Information-Centric Networking (ICN) [7] [8] is a networking architecture that decouples the service from the host and remove content association to any physical location, thus, practical in high mobility and intermittent connectivity scenarios involving vehicles. It allows intermediate nodes to en-route cache content while forwarding it to a user. Thus, we borrow the named-data networking concept of the information-centric networking paradigm to overcome the mobility and intermittent connectivity challenge in connected vehicles.

Results reveal that the proposed vehicle recruitment scheme ROVERS can efficiently find eligible vehicles for urban data collection with high user interest satisfaction.

The major contributions of this paper can be summarized as follows:

- A distributed vehicle ranking scheme *CarRank* is proposed, where each vehicle can autonomously classify itself in the network based on the importance of its collected/cached information, its spatio-temporal availability and network connectivity.
- A non-collaborative game is presented for vehicles receive monetary reward proportional to the delivered content instead of its CarRank score.
- A centralized recruitment algorithm to optimally and fairly select the best set of vehicles identified by CarRank for urban data collection while maintaining acceptable city-wide coverage with less redundant data.
- We perform in-depth simulations on 2,986 connected vehicles using realistic mobility traces to evaluate the proposed vehicle recruitment scheme.

The rest of the paper is organized as follows. The next section highlights the major related work. In Section III, we present the system model followed by the vehicle selection problem formulated in Section IV. The information-centric CarRank scheme to analytically find the vehicle eligibility is discussed in section V. Section VI discusses the numerical evaluation based on simulations. In Section VII we conclude the paper and discuss future directions.

## II. RELATED WORK

Recruitment of smart vehicles for data collection is studied recently in [9] where high reputation users are recruited to perform urban sensing for a desired coverage, given a limited budget. However, the authors do not provide a proper metric to classify and identify the participants eligibility. Similarly,

the authors in [10] propose a vehicle recruitment algorithm to select the minimum number of vehicles that can achieve the required coverage requirements for a road section. Another recruitment approach [11] proposed a heuristic to minimize the number of vehicles selected for crowdsensing under a coverage and budget constraints.

The work in [12] formulated vehicle recruitment as an NP-Complete problem and proposed a greedy approximation along a genetic algorithm to solve it in polynomial time under a given coverage and quality constraint. GoSense [13] is a platform targeting personalized crowdsensing tasks to achieve the maximum sensing coverage by selecting the minimum number of vehicles. The authors use the vehicle projected trajectory and speed in order to be recruited as a participant for urban sensing. The above works only considered coverage as the metric while ignoring the vehicle inherent abilities such as its availability and connectivity to facilitate data collection in an urban scenario.

Similarly, in [14], the vehicle selection problem is formulated as a knapsack problem and a greedy optimal selection strategy is proposed to select vehicles with appropriate sensing capabilities in a given area and time under budget restrictions. In another work [15], vehicles are selected as participant for urban crowdsensing where the focus is to assign multiple sensory task to vehicles. The vehicle recruitment problem is proved to be NP-Hard under a budget and coverage constraint and a greedy strategy is proposed to recruit participants vehicles based on their capability to cover more roads in the urban environment. The authors in [16] also consider multi-task based crowdsensing using vehicles, where the recruitment is proven to be NP-Hard and a greedy algorithm is proposed to maximize the joint coverage of the recruited vehicles under a given budget constraint. [17] also proved the vehicle recruitment problem as NP-Hard and proposed two heuristic algorithms to reduce the participants cost for a given trajectory. The above works used a simplistic strategy such as the vehicles sensing duration or their capability to cover more roads along budget as the constraint, and thus, does not reflect its relevance with respect to the users' interest. Moreover, the authors does not address the incentives provisioning to the recruited sensing vehicles.

The authors in [18] defined a duration-principle to limit the sensing operation to limited time epochs instead of having vehicles performing a continuous sensing. The duration-variable participant recruitment (DPR) problem is shown to be NP-Hard and a near-optimal approximation participant selection algorithm is presented where vehicles that maximizes the uncorrelated observation are selected. The approach does not provide the flexibility of achieving a user desired spatio-temporal coverage of vehicles in different urban neighborhoods under a budgetary constraint.

Game theory is exploited in [19] to cache 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. Another such work is proposed in [20] where the authors adapt a game theoretic approach in ICN to stimulate wireless access point owners

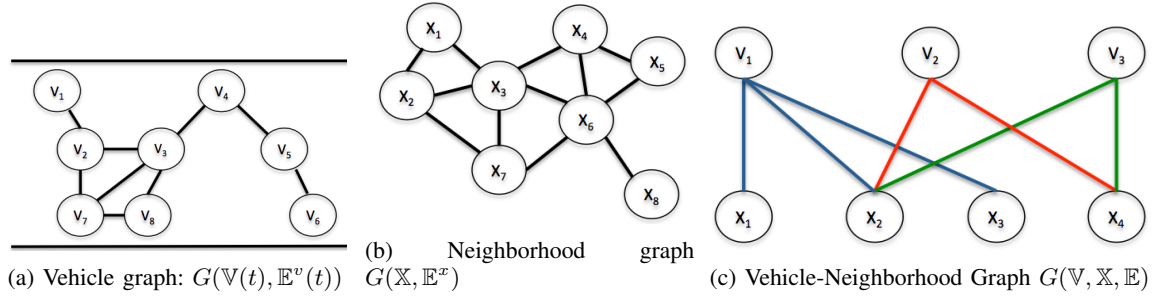


Figure 1: Network Model

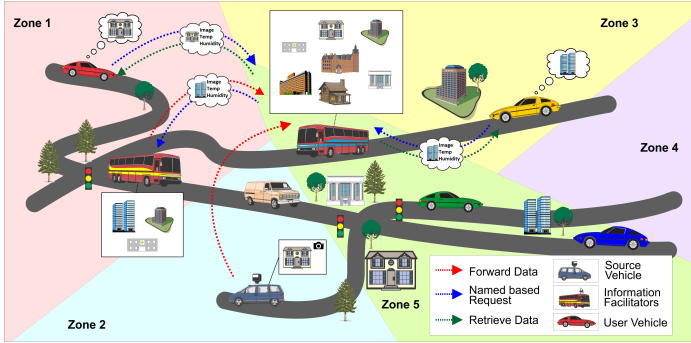


Figure 2: System Overview

to jointly lease their unused bandwidth and storage space to a content provider under partial coverage constraints. Both these works use game theory in the network of fixed nodes while we are the first ones to consider the case where vehicles can be used for data collection and storage in the network.

Different from all the above discussed works, ROVERS is a vehicle recruitment method for a large scale city-wide coverage which exploits game theoretic concepts to fairly recruit the best set of vehicles towards an efficient and scalable urban data collection. ROVERS caters the intermittent connectivity and high mobility challenge of vehicles. Instead of using typical unstable measures such as the inter-vehicle contact interval or their contact duration, the vehicle reputation is measured based on mobility independent metrics such as the importance of the information the vehicle is providing and its availability in the neighborhood.

### III. SYSTEM MODEL

The system overview is outlined in the Figure 2. We begin by an example of a vehicle equipped with an On-Board Unit (OBU) interconnecting the different sensors and cameras to sense and monitor urban roads. 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 provide below the network connectivity model and then

discuss the possible business model in a typical data collection in a location-aware application.

#### A. Network Model

The network is comprised of vehicles as nodes in an undirected graph  $G(\mathbb{V}(t), \mathbb{E}^v(t))$ ,  $\mathbb{V}(t) = \{v\}$  as the set of vehicles on urban roads at time  $t$  and  $\mathbb{E}^v(t) = \{e_{jk}(t) \mid v_j, v_k \in \mathbb{V}, j \neq k\}$  as the edges representing the connectivity between vehicles  $v_j$  and  $v_k$  at time  $t$  based on each their respective communication range. We define the urban neighborhood as an undirected graph  $G(\mathbb{X}, \mathbb{E}^x)$  where the set  $\mathbb{X} = \{x\}$  represents an urban neighborhood  $x$  and  $\mathbb{E}^x = \{e_{pq} \mid x_p, x_q \in \mathbb{X}, p \neq q\}$  is the edges set connecting them through the roads network.

Since we are dealing with location-aware information collected or stored by a vehicle in the urban neighborhood, we define *Information association* as a bipartite graph  $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$ . Here  $\mathbb{V}$  is the set of vehicles in  $G(\mathbb{V}(t), \mathbb{E}^v(t))$  and  $\mathbb{X}$  is the set of neighborhoods in  $G(\mathbb{X}, \mathbb{E}^x)$  as shown in Figure 1c. The edge  $\mathbb{E} = \{e_{ij} \mid v_i \in \mathbb{V}, x_j \in \mathbb{X}\}$  associates each vehicle to neighborhoods  $X_v \subset \mathbb{X}$  for which it collected or cached information.

Information association allows us to use the ICN hierarchical naming convention “/neighborhood/road-section/information-type”. The *information-type* can be content related to different LBS applications for congestion information [21], safety warnings or Infotainment on a *road-section* in a *neighborhood*. For example, a vehicle can prioritize safety information over infotainment, however, to consider a generic content encompassing all types of information, in the remaining of the article we consider the set  $X = \{x\}$  to represent location/neighborhood or information/content.

We consider a spatio-temporal vehicular network taking into account a dynamic time-varying topology. Therefore, we divide the time into consecutive time-slots  $T = (\bar{t}_1, \bar{t}_2, \dots)$ , where  $\bar{t}_k = [t_k, t_{k+1})$  is the  $k^{th}$  time-slot with the time instant  $t_k$  as the beginning of the time-slot. We represent  $d(x, x_k)$  as the distance between the associated content location  $x$  and  $x_k$  and the vehicle’s current position at time instant  $t_{k+1}$ , where  $x, x_k \in X_v$ .

We assume vehicles are aware of the neighborhood graph  $G(\mathbb{X}, \mathbb{E}^x)$  as the city map. However, it is unlikely that it collected and thus stored data of all the neighborhoods. Therefore, the content a vehicle collects and store can be limited depending on its storage capacity and the its daily commute

between different neighborhoods. The notations used in the paper are summarized in Table I.

### B. Business Model

The basic business model for a location-aware application can have a service provider and a set of users subscribed for content regarding different urban neighborhoods. Three players can be considered: (i) the service provider with a monetary gain proportional to the content delivered to the users. (ii) Vehicles as ROVERS with relatively large resources (caching, computing and communication) capable to facilitate nearby mobile users with the information regarding urban streets. We assume such vehicles incur a facilitating cost proportional to their provided resources, which is less than the cost of using the infrastructure network, and therefore, receive a reward greater than the consumed resources from the service provider. (iii) The mobile subscribers intended to consume the content and pay a cost proportional to the received content from the service provider where the cost to download content from a nearby rover in an ad-hoc network is less than using the infrastructure network.

Using vehicles as ROVERS economically benefit all the three players: (a) The service provider cost and bandwidth consumption is reduced by avoiding the infrastructure network to deliver content. (b) ROVERS receive a reward which is more than their incurred cost, and, (c) The mobile users retrieve content from nearby ROVERS using a low-cost connectivity in an ad-hoc network.

## IV. PROBLEM FORMULATION

In this section we formally describe the ROVERS selection problem by first defining the objective function followed by the optimization problem taking into account the underlying constraints.

### A. Objective Function

The service provider can select a limited set of vehicle as ROVERS to achieve a desired coverage in an urban neighborhood. The ROVERS selection by a service provider can therefore be formulated as an optimization problem where we are interested in finding the optimal number of ROVERS  $V \in \mathbb{V}$  that provides an acceptable coverage in different neighborhoods. The objective function  $F: \mathcal{V} \rightarrow \mathbb{R}_+$  aims to select the best set of vehicles  $\mathcal{V} = \{V : V \subset \mathbb{V}\}$ , that satisfy a given coverage requirements.

### B. Optimization problem

Urban sensing requires sufficient coverage of every street, however the term ‘‘sufficient’’ can be relative to different applications. An application can have different coverage requirement regarding the number of vehicles, the accuracy/precision and timeliness of the data collection and most importantly the amount of user interests for a data related to a neighborhood. Such coverage requirements can differ for different neighborhoods. For instance, based on the application requirement, a densely populated neighborhood might require more vehicles

Table I: List of Notations

Notation	Description
$\mathbb{V}$	Vehicles set
$\mathbb{X}$	Neighborhoods set
$\mathbb{E}^x / \mathbb{E}^v$	Edges set between neighborhoods/vehicles
$\mathbb{E}$	Edge between vehicles and neighborhoods
$\bar{t}_k$	$k^{th}$ time-slot
$t_k / t_{k+1}$	time instant current/next
$X_v$	Neighborhoods cached at vehicle $v$
$d(x, x_k)$	Distance from current location $x_k$ to $x$
$I_x^v$	Interests satisfaction frequency for $x$
$r_x$	Content $x$ responded in previous slot
$R_x$	Total responds for content $x$
$t_x^f$	Time of previous respond for $x$
$\bar{t}_d$	interest deadline (average)
$t_x$	Interest validity bound for $x$
$\tau$	Information validity
$\delta$	Information validity tuning parameter
$\lambda$	Tune importance based on distance from $x$
$s_x^v$	Vehicle itself source for $x$
$f_x^v$	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 $x_k$
$MI_x^v(t_k; x_k)$	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)
$k_\Gamma^v$	Vehicle average neighbor degree
$\Gamma_v$	Set of neighbors for vehicle $v$
$C_\Gamma^v$	Neighbor vehicle centrality
$f_\Gamma^v$	Vehicle neighborhood importance function
$C_v$	Vehicle centrality
$\alpha / \beta / \gamma$	Tuning parameters for each function
$\theta$	Smoothing factor for vehicle centrality
$A_x^{min} / A_x^{max}$	Minimum/maximum coverage requirement
$A_x$	Neighborhood coverage vector
$A_v$	Vehicle coverage vector
$G_T$	Stage game for $T$ time-slots
$u_v$	Vehicle utility
$\succ$	Preference relation between content

to cover given the number of content demands. The contrary is also possible where only one vehicle is sufficient to satisfy the users’ need in the same densely populated neighborhood. It is also to note that the users’ interests might differ at different times of the day in different neighborhoods. Therefore, to address the coverage requirements, we define below the spatio-temporal coverage requirement for a content at the selected ROVERS:

**Definition 1:** (Spatio-temporal coverage) We define the vector  $A_x(\bar{t}_k)$  as the number of vehicles in the set  $V \in \mathbb{V}$  required to provide content/coverage for the neighborhood  $x \in \mathbb{X}$  at the time-slot  $\bar{t}_k$  based on the desired redundancy, cost and quality requirements of a location-aware application.

The number of vehicles to sense or cover a particular neighborhood can be reduced to avoid redundant data collection and vehicle recruitment costs. Similarly the quality of the collected data can be improved by designating more ROVERS to provide coverage in a neighborhood. Thus, we require at least  $A_x^{min}$  and at most  $A_x^{max}$  vehicles at a time-slot  $\bar{t}_k$  to

satisfy the redundancy, cost and quality requirements. The term  $A_x^{min}$  is the lower bound for the spatio-temporal coverage while  $A_x^{max}$  is the upper bound representing the maximum number of vehicles needed under a quality requirement. For instance, setting  $A_x^{min}(\bar{t}_k) = A_x^{max}(\bar{t}_k) = 1$  defines the minimum redundancy requirement by allowing exactly one vehicle covering each neighborhood at a time-slot  $\bar{t}_k$ . The ROVERS selection problem can be formulated as follows:

$$\begin{aligned} & \underset{V}{\text{minimize}} && F(V) \\ & \text{subject to} && A_x^{min}(\bar{t}_k) \leq A_x(\bar{t}_k) \leq A_x^{max}(\bar{t}_k), \forall x, \forall \bar{t}_k, \end{aligned}$$

The objective function allows a service provider to recruit the most suitable vehicles as the ROVERS set  $V$  where the constraint states that the recruited vehicles should satisfy the spatio-temporal coverage requirements for a neighborhood  $x$  at a time-slot  $\bar{t}$ .

In the next section, we present CarRank as the novel vehicle centrality scheme for vehicles to autonomously classify themselves as suitable candidates to become ROVERS in the network.

## V. CARRANK FOR ROVERS CLASSIFICATION

We witness thousands of vehicles on urban roads, however, a service provider can recruit only a limited set of vehicles for urban data collection, given its budgetary limitations. Thus, it is important to efficiently identify vehicles suitable for data collection. The issue is analogous to finding important nodes in the network using centrality measures though, typical centrality measures require a network wide information where the high mobility of vehicles and the dynamic network topology make it difficult to compute a vehicle's centrality. It cannot be done by the service provider as regular centrality computation would require continuous data exchange between the service provider and the vehicle, resulting in high bandwidth consumption. To cater this, we present CarRank as the criteria which classifies vehicles better suitable for ROVERS in a distributed fashion.

CarRank is based on three novel albeit important parameters, (i) the stored **information importance** as the vehicle's ability to classify itself based on the importance of the cached information with respect to the urban users (ii) the vehicle's **spatio-temporal availability** as its availability at the right time and place for collecting, storing or distributing information based on its mobility pattern and (iii) its **neighborhood importance** referred to the network topological connectivity as the ability to better diffuse/spread information in the network.

### A. Information Importance

Users in an urban environment interested in different location-aware content can generate interests towards vehicles where a content with an increasing number of interests can be considered popular. We leverage the ability of a vehicle to respond with content for the user interest it receives to periodically compute the importance of the cached information/content. It can do this by just observing the amount and frequency of user interests it received and subsequently

responded back with a content. Such observation can be made at regular time intervals where the frequency and granularity is subject to the application requirement. However, to generalize, the frequency of satisfied interests or content delivered by the vehicle in the time-slot  $\bar{t}_k$  can be defined as  $I_x^v(\bar{t}_k) = \frac{r_x(\bar{t}_k)}{R_x}$ ,  $A_x(\bar{t}_k) \neq 0$ . The term  $r_x(\bar{t}_k)$  represents the number of interests successfully responded with content in the previous time-slot and  $R_x$  represents the total responds to the user interests for the content  $x \in X_v$  by the vehicle  $v \in \mathbb{V}$ .

The user interests for a location-aware information might vary, i.e. information regarding a congestion on a road change with respect to time, and is no longer valid for users after the congestion is cleared. CarRank ensures that the importance of the information reflect such behavior by defining an exponential decay function as the user relevant information validity:

**Definition 2:** (User Relevant Information Validity) The information validity with respect to the users  $\tau$  at the time-slot

$$t_{k+1} \text{ is defined as: } \tau_x(t_{k+1}) = \begin{cases} 1 & t_{k+1} \leq t_x^f + \bar{t}_d \\ e^{-\delta \bar{t}_d} & t_{k+1} > t_x^f + \bar{t}_d \end{cases}$$

where,  $t_x^f$  is the time instant since the content  $x$  was last responded for a user interest. The average interest deadline specified by the users can be given as  $\bar{t}_d = \frac{1}{n} \sum_x^n t_x$ . Here  $n$  is number of interests responded by the vehicle in the previous time-slot and  $t_x$  is the deadline specified by a user interest for the content  $x$ . The tuning parameter  $\delta \in [0, 1]$  addresses the decay rapidity which depends on the application requirement. (i.e. maximum 1 hour can be the validity of an information regarding an accident).

In case of no active interests received by the vehicle for a content in the previous time-slot and the average interest validity deadline is expired, the information importance should be decayed exponentially as it is no longer useful to the urban users. Similarly  $\tau_x$  can be set to unity in case the application requires the information be always available for the users.

The vehicle computes the stored information importance for the time instant  $t_{k+1}$  as:

$$\begin{aligned} f_I^v(t_{k+1}) &= f_I^v(t_k) + \\ & \frac{1}{|X_v|} \sum_{x \in X_v} (\tau_x(t_{k+1}) I_x^v(\bar{t}_k) (1 + d(x, x_k))^{-\lambda} + s_x^v(t_{k+1})) \end{aligned} \quad (1)$$

where,  $f_I^v(t_k)$  is the importance of the cached content computed in the previous time-slot (time instant  $t_k$ ). The term  $|X_v|$  is the cardinality of the sub-graph  $X_v \subset \mathbb{X}$ , i.e. set of all neighborhoods for which content is collected/cached by the vehicle  $v$ .

In case the vehicle does not satisfy interest in the previous time-slot, then  $I_x^v(\bar{t}_k) = 0$  to avoid the vehicle increasing the information importance.

Here,  $\tau_x$  is the content  $x$  validity metric,  $d(x, x_k)$  is the distance between the content location  $x$  and  $x_k$  and the vehicle's current position, where  $\lambda \in [0, 1]$  is a tuning parameter for the impact of distance from the content location towards the information importance computation.

We use  $s_x^v \in [0, 1]$  as the percentage of time the vehicle  $v$  was the original content  $x$  source (i.e. collected information itself). One possible use of  $s_x^v$  is to allow interests for a

particular content later in time route to the original source vehicle which collected the data in the network as a content distribution network.

### B. Spatio-temporal Availability

Spatio-temporal availability of the vehicle reflects the driver social behavior. It considers the vehicle physical availability in a neighborhood 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 to predict its urban mobility pattern. 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  $R_x^v(t_k, x_k)$  is the interests satisfied for content  $x$  at the current time and location in the past and  $R_T$  is 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. 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:

$$MI_x^v(t_k; x_k) = \sum_{\forall t_k \in T} \sum_{\forall x \in X_v} p_x^v(t_k, x_k) \log \left( \frac{p_x^v(t_k, x_k)}{p_x^v(t_k) p_x^v(x_k)} \right), \quad (2)$$

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 neighborhoods:

$$f_{T,X}^v(t_{k+1}) = \frac{1}{|X_v|} \sum_{x \in X_v} MI_x^v(t_k; x_k) \quad (3)$$

The function in (3) 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 neighborhood.

### C. Neighborhood Importance

The neighborhood of the vehicle in a distributed system is important for efficient content distribution and storage. We

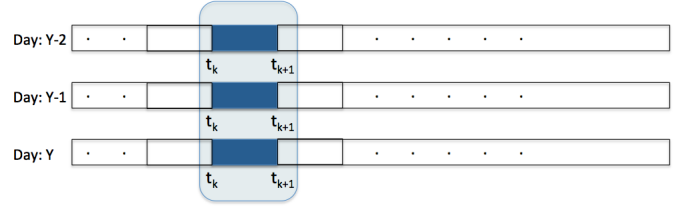


Figure 3: Spatio-temporal availability in the same time-slot

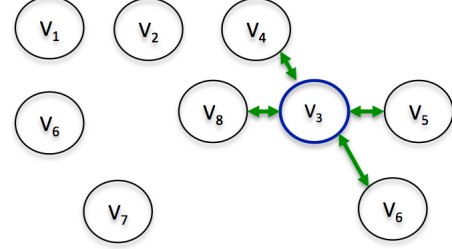


Figure 4: Neighborhood Centrality Exchange

incorporate the neighborhood information by allowing other ROVERS in the vehicle's transmission range to share their respective importance as well as their connectivity information. The idea is to consider better connected vehicles with better information collection and diffusion capabilities in its vicinity. This instantiates the use of the vehicles physical topological information. For this purpose, we consider the vehicle's assortativity as its average neighbor degree  $k_\Gamma^v$ . Besides topological connectivity, each neighbor centrality  $C_\Gamma^v$  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:

$$f_\Gamma^v(t_{k+1}) = \frac{1}{k_v} \sum_{\Gamma_v \in \mathbb{V}} C_\Gamma^v(t_k) \cdot k_\Gamma^v \quad (4)$$

where  $k_v$  is the vehicle degree at time  $t$  in  $G(\mathbb{V}(t), \mathbb{E}^v(t))$ .

Since centrality measures are unfeasible to use due to the lack of the knowledge of the entire network information to the vehicle at the time of importance computation. Therefore, the function  $f_\Gamma^v(t_{k+1})$  in (4) considers more information than just the degree of the vehicle, at the same time, maintaining a local scope and learn only from local information within the vehicle range as shown in the Figure 4.

**Vehicle Centrality:** The Exponential Weighted Moving Average (EWMA) vehicle centrality at the time instant  $t_{k+1}$  is computed as where the functions in Equations (1), (3) and (4) contribute to compute CarRank:

$$C_v(t_{k+1}) = \theta C_v(t_k) + (1 - \theta)[\alpha f_\Gamma^v(t_{k+1}) + \beta f_{T,X}^v(t_{k+1}) + \gamma f_T^v(t_{k+1})] \quad (5)$$

To consider the impact of each function with respect to an application requirement, we use  $\alpha, \beta$  and  $\gamma$ , and  $\theta$  as the tuning parameters, where  $\alpha + \beta + \gamma = 1$ , and  $\theta \in [0, 1]$ . For example, the vehicle can better collect or spread content if it is in a densely populated (i.e. highly connected) neighborhood.

**Algorithm 1** CarRank

---

```

1: INPUT: Graph  $G(\mathbb{V}, \mathbb{X}, \mathbb{E})$  :
2: OUTPUT: CarRank  $C_v(t_{k+1})$ 
3: for cached content  $x \in X_v$  in vehicle  $v \in \mathbb{V}$  do
4:   Compute its importance (1)
5:   Compute mutual information with respect to the content
     (2)
6: end for
7: for neighbor vehicle  $\Gamma_v \in \mathbb{V}$  do
8:   Exchange assortativity  $k_{\Gamma}^v$ 
9:   Exchange  $C_{\Gamma}^v(t_k) \leftarrow$  neighbor centrality
10: end for
11: Compute spatio-temporal availability (3)
12: Update neighborhood importance (4)
13: Update CarRank (5)
14: return  $C_v(t_{k+1})$ 

```

---

In such a case, the vehicles neighborhood importance can be weighted more compared to say, spatio-temporal availability.

Algorithm 1 summarizes the periodic CarRank computation by the vehicle  $v$  after every  $\bar{t}$  duration. For each cached content regarding a neighborhood  $x$ , it computes the information importance using Equations (1) (Line 5). Further, to find the spatio-temporal availability in Equation (3), the vehicle computes the respective mutual information for each of the cached content  $x$  using Equation (2) (Line 6). Similarly, it computes the average neighbor degree using Equation (4) and exchanges it with the neighboring vehicles in range. The information importance function, spatio-temporal availability and neighborhood function all-together contribute to the vehicle CarRank score for the next time-slot  $t_{k+1}$  using Equation (5). Each vehicle shares its computed CarRank score with the service provider and neighboring vehicles every  $\bar{t}$  duration where the choice of a suitable  $\bar{t}$  depends on the application requirements.

*Computational Complexity:* Algorithm 1 considers a single vehicle  $v$  leveraging local information regarding its cached information and its neighborhood towards computing CarRank. The vehicle time consumption for the first loop in Lines 3 – 6 is in the order of  $|T| |\mathbb{X}|$ , the maximum information (i.e. the set  $\mathbb{X}$ ) possible to cache. Further, its consumption for the second loop (Lines 7 – 10) is in the order of  $|T| |\Gamma_v|$ , i.e. its neighbors at the time of CarRank computation scaled by all time-slots. The tasks following (Lines 11-14) are simply  $\mathcal{O}(1)$  statements, Algorithm 1 run-time complexity can thereby be given as  $\mathcal{O}(|T| (|X_v| + |\Gamma_v|))$ . Such linear complexity scheme of CarRank is affordable and therefore can be scalably computed in polynomial time by modern vehicles with sufficient computational resources.

## VI. GAME FOR ROVERS SELECTION

CarRank allows vehicles to rank itself to become eligible for urban data collection, however, it is possible CarRank computation can be counterfeited by the vehicles where rational or untrusty vehicles can increase their respective CarRank score. Therefore, once the best vehicles eligible to become

ROVERS can be identified using CarRank, we use game-theoretic concepts to achieve a fair recruitment of the best set of ROVERS  $V$  where instead of the CarRank scores the vehicle receives a high payoff proportional to the amount of content it delivers  $x$  to the users.

In this section, we first provide the corresponding game formulation, define the utility function as the payoff to consider by both the vehicle and the service provider, followed by the Nash Equilibrium condition towards the fair allocation of both the service provider payoff as well as the vehicle's reward.

### A. Game Formulation

We formulate the non-cooperative game for the fair recruitment of ROVERS as the tuple  $\langle V, A, u, G^T \rangle$  with the vehicles set  $V \in \mathbb{V}$  as players to play the stage game  $G$  for  $T$  periods as  $G^T$  with the service provider, thus, two players at a time. At each state (time-slot), we consider a sub-game where the action set (coverage) of  $m$  players is represented by  $A = (A_1, \dots, A_m)$ . Thus, for each player,  $A_v = \{A_{x_1}, A_{x_2}, \dots, A_{x_n}\}$  is the (finite) set of possible actions or strategies as the set of neighborhoods for which it can provide coverage based on its cached content. Similarly,  $u = (u_1, \dots, u_m)$  is the utility function for  $m$  players, where  $u_m$  is the utility/reward of  $m^{\text{th}}$  player proportional to the content it delivers to the users.

The **Pure Strategy** for a vehicle  $v$  to provide coverage regarding a neighborhood  $x$  in the time-slot  $\bar{t}_k$  is the binary action set  $A_x(\bar{t}_k) = \{0, 1\}$ , i.e. to provide coverage for  $x$  or not. Similarly, the **Mixed Strategy** for the vehicle  $v$  is the action set  $A_v = \{A_{x_1}, A_{x_2}, \dots, A_{x_n}\}$  where the actions are set to unity ( $A_{x_i} = 1$ ) for neighborhoods  $X_v$  the vehicle provides coverage in the set of all known neighborhoods i.e.  $X_v \in \mathbb{X}$ .

### B. Solution Concept

Given the pure strategy and mixed strategy of the vehicle, we define below the best response for a vehicle to prioritize its most associated neighborhood (collected/cached related data) by either frequent visits or more availability in the respective neighborhood.

**Definition 3:** (Best Response) For a vehicle  $v$ , the best response is a preference relation  $\succ$  defined as a probability distribution over the set of content cached  $X_v$ , where, for two contents,  $x_1$  and  $x_2$ ,  $x_1 \succ x_2$  indicates that node  $v$  prefers to provide coverage for  $x_1$  over  $x_2$ .

**Utility function:** We will derive below the appropriate utility function for both the service provider and the vehicle in order to achieve a fair payoff/reward. The vehicle utility should be proportional to the amount of content delivered to the users. It should also be greater than the vehicle cost of delivering content which can be proportional mainly to its fuel consumption along possibly other supplementary costs. Similarly the service provider utility should consider the amount of content delivered to the users where its costs of using vehicles/ROVERS should be less than using the infrastructure network. Thus, in order to use a common utility function agreeable between both players, we define the payoff or utility function for both the vehicle  $v$  and the service provider as:



$$u_v(t_k) = \sum_x I_x^v(t_k) \cdot c_x, \quad A_x(t_k) \neq 0, \forall x \in X_v$$

where,  $I_x^v(t_k)$  are the amount of interests satisfied/content delivered for neighborhood  $x$  by the vehicle  $v$  and  $c_x$  is the delivery cost per content  $x$  paid to the rover. Since the payoff is proportional to the amount of content delivered, any attempt to increase or alter the CarRank score will not impact the amount it receives from the service provider.

Let us recall the preference relation between neighborhoods for the vehicle. It will collect more data or deliver more content for a neighborhood it is more related to by either visiting it or caching its content. Thus, the following Nash equilibrium condition should be respected:

**Nash Equilibrium:** The Nash equilibrium over the mixed strategies for a vehicle states that  $v$  chooses a strategy profile  $A_v^*$  to gain more payoff over another strategy vector  $A_v$  such that:

$$u_v(A_v^*) \geq u_v(A_v), \forall \bar{t}_k,$$

There will be multiple Nash equilibria since the time is divided into time-slots where different vehicles are recruited as ROVERS for a neighborhood at different time of the day. We observe a perfect game since both the service provider and the vehicle are aware of the utility function, i.e. the amount of content delivered and the cost per unit content paid by the service provider. Therefore, we devise a sub-game perfect equilibrium condition below to further refine the solution concept:

**Definition 4:** (Sub-game Perfect Equilibrium) In a sub-game perfect equilibrium, a vehicle chooses a strategy profile  $A_v^*$  at the time-slot  $\bar{t}_k$  in the stage game  $G_T$  to gain more payoff over another strategy vector  $A_v$  at the same time-slot such that:

$$u_v(A_v^*, \bar{t}_k) \geq u_v(A_v, \bar{t}_k),$$

*Proof:* (Equilibrium Existence) Naturally, based on the daily commute, a vehicle tends to provide more coverage for the neighborhoods it visited and will respond with the related content more frequently, thus, its best response suggest the existence of at least one Nash equilibrium.

*Proof:* (Equilibrium Uniqueness) Since the time-slots are a consecutive sequence of the time  $T = (\bar{t}_1, \bar{t}_2, \dots, \bar{t}_N)$ , and do not super impose. By backward induction, we can have a unique Nash equilibrium for each vehicle, where at each individual time-slot  $t_k$  (i.e. each sub-game), it follows its respective best response.

The above sub-game equilibrium condition is for the vehicle strategy to cover a set of neighborhoods  $X_v$  at each time-slots, where within the time-slot it will have preference relations for satisfying user interests for content related to some neighborhoods more than others as defined by its best response. Therefore, to further refine on per content basis, we define below the pareto optimality where in a sub-game (time-slot  $t_k$ ), its utility function is proportional only to the amount of content it delivered.

---

### Algorithm 2 ROVERS selection

---

```

1: INPUT: All vehicles  $\mathbb{V}$ 
2: OUTPUT: ROVERS  $V \subset \mathbb{V}$ 
3: while ( $A_x^{min}(\bar{t}_k) \leq A_x(\bar{t}_k) \leq A_x^{max}(\bar{t}_k)$ )  $\forall x, \forall \bar{t}_k$  do
4:   Find  $v' = \arg \max_{v \in \mathbb{V}} C_v(t_k)$ 
5:    $V \leftarrow V \cup v'$ 
6:    $A_V(\bar{t}_k) \leftarrow A_V(\bar{t}_k) + A_{v'}(\bar{t}_k)$ 
7:    $u_V \leftarrow u_V + u_{v'}$ 
8: end while
9: return ROVERS  $V$ 

```

---

**Definition 5:** (Pareto Optimality) A vehicle payoff  $u_v(A_x^*, \bar{t}_k)$  for a neighborhood  $x$  based on the strategy  $A_x^*$  is **pareto optimal** in the stage game  $G_T$  if  $u_v(A_x^*, \bar{t}_k) \geq u_v(A_x, \bar{t}_k)$ , where it is **strongly pareto optimal** if  $u_v(A_x^*, \bar{t}_k) > u_v(A_x, \bar{t}_k)$  for a given time-slot  $t_k$ .

The above Definitions 4 and 5 provide the condition required to achieve the unique Nash equilibrium condition and result in a fair vehicle recruitment while taking into account its preferences. We present below a simple selecting procedure for the service provider to recruit based on the desired spatio-temporal coverage constraint and the vehicle individual rationality.

#### C. Selecting Procedure

We present here a centralized heuristic for the service provide to recruit the best vehicles identified by CarRank to achieve a desired coverage, given its redundancy and quality requirements at different locations and different time-slots. The best ROVERS selection procedure is shown in Algorithm 2. The spatio-temporal coverage constraints are stated in Line 3, where for each time-slot, the service provider ensures the desired coverage and redundancy requirements for each neighborhood content  $x$  are respected. The minimum and maximum required vehicles  $A_x^{min}(\bar{t}_k)$  and  $A_x^{max}(\bar{t}_k)$  are stated for all time-slots  $t_k$  and for all set of contents  $\mathbb{X}$ . The service provider finds the best ROVERS  $v' \in \mathbb{V}$  identified using CarRank in each neighborhood  $x$  and time-slot  $t_k$  and add them to the set of ROVERS  $V \subset \mathbb{V}$  in order to recruit them (Line 4).

It is to note that each vehicle interested to be recruited as rover shares its CarRank score  $C_v(t_k)$  with the service provider. The respective coverage provided by the vehicle  $A_{v'}(t_k)$  is added to the total coverage provided by the set of ROVERS  $A_V(t_k)$  recruited for the time-slot  $t_k$ . Subsequently, its utility  $u_{v'}(t_k)$  is added to the total utility  $u_V(t_k)$  to be paid to the set of recruited ROVERS for the time-slot  $t_k$ . Finally, the algorithm returns the selected set of vehicles as rover along with their respective coverage and utility to be paid by the service provider.

#### D. Convergence and Complexity

The ROVERS selection process in Algorithm 2 converges in the following condition: (i) Sufficient amount of vehicles  $V$  are found to be recruited for the desired coverage requirements for each neighborhood within a given budget limits or (ii) Coverage requirements in Line 3 are meet for an unconditionally

Table II: Simulation Parameters

Parameter	Value
Simulator	NS3
Nodes	2,986
Duration	1hr
Time granularity	1 sec
Mobility trace	Köln, Germany
Area	$6 \times 6 km^2$ city center
Packet	1KB
Range	100 -150m
Fading model	Nakagami
Path loss model	Log distance
Simulations	5

unlimited budget and (iii) Not enough vehicles are available to achieve required coverage for a neighborhood and time-slot.

The complexity of the recruitment process is proportional to the amount of neighborhoods in the city, the time granularity of the recruitment process repetitions and the amount of vehicles selected as ROVERS. The statement  $\arg \max$  in Line 4 is running on a sorted set of vehicles and therefore has a time complexity of  $\mathcal{O}(1)$ . The algorithm's time complexity can be expressed as  $\mathcal{O}(|T| |\mathbb{X}| |A_x^{max}(\bar{t}_k)|)$ , the complexity of the order of three nested loops.

In the worst case, the maximum coverage requirements  $A_x^{max} = \mathbb{V}$  for a content  $x$  can be for all vehicles  $\mathbb{V}$  to be available to provide the content in all neighborhoods  $\mathbb{X}$  in the urban environment. This results in a time complexity of  $\mathcal{O}(|T| |\mathbb{X}| |\mathbb{V}|)$ . On the other hand, in the best case, the maximum coverage requirements  $A_x^{max} = 1$  can be chosen such that, at maximum, a single vehicle provides coverage for the neighborhood  $x$ , thus the temporal complexity of the algorithm can be expressed as  $\mathcal{O}(|T| |\mathbb{X}|)$ . It is to note that it is unfavorable for the service provider to have such maximum coverage requirement  $A_x^{max} = \mathbb{V}$ , i.e. recruiting all vehicle to provide coverage at all neighborhoods due to the increase in its cost in paying a reward/utility to all vehicles. Therefore, a careful choice of  $A_x^{max}$  can substantially reduces the time complexity of the recruitment process. Nevertheless, ROVERS is an efficient approach ensuring the service provider recruit the best vehicles in polynomial time under any possible coverage and quality requirements.

## VII. PERFORMANCE EVALUATION

We evaluate the proposed game-based model using ndnSIM module of NS-3 [22] where the named-data networking NDN architecture is implemented with a discrete time granularity of 1 second. The vehicles mobility is extracted from a realistic model for large scale vehicular mobility in Cologne, Germany [23]. We choose vehicles as potential candidates to become ROVERS leveraging the relatively high computing, caching and communication capabilities. For the impact of fading and shadowing due to obstacles such as buildings, we use the combination of Nakagami propagation and Log distance path loss model.

A total of 2,986 vehicles are used to validate the city-wide scalability of our vehicle identification and selection scheme. The analysis is performed by dividing an area of  $6 \times 6 km^2$

into 36 uniform neighborhoods each of  $1 \times 1 km^2$ , where,  $|\mathbb{X}| = 36$ ,  $|\mathbb{V}| = 2,986$  and  $\bar{t}_k = 10$  minutes within total simulation time of  $T = 60$  minutes.

### A. Simulation Scenario

The simulation scenario implements two types of vehicles, (i) the rover or source vehicle which collects and cache content and (ii) the consumer vehicles interested in the content. Consumer vehicles generate interests for a pre-known content sequence with a randomly set average validity deadline  $\bar{t}_d$ .

We consider 36 contents representing the set of neighborhoods with each unique chunks of size 1KB following a Zipf distribution (skewness parameter=1) as the content popularity profile, i.e. frequent interests for more popular content. We also allow ROVERS to perform en-route caching where uniform buffer sizes of 10MB are assigned to each vehicle. Thus, any caching vehicle already containing the content responds to the consumer interest. 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 (interests) in the first 10 minutes.

For each vehicle, the CarRank score  $C_v$  is computed using Equation 5 periodically after each  $\bar{t} = 10$  minutes. The tuning parameters defined earlier are application dependent, however, to maintain generality in our experiments and to perform an unbiased experimentation, the tuning parameters are set such that,  $\alpha = \beta = \gamma = 0.33$  and  $\theta = \lambda = \delta = 0.5$ . The service provider coverage quality requirements is set to have at least one vehicle  $A_x^{min}(\bar{t}_k) = 1$  and a maximum of  $A_x^{max}(\bar{t}_k) = V$  vehicles for the neighborhood  $x$  and time-slot  $\bar{t}$ . The utility  $u_v$  is derived for each vehicle and  $u_V$  is derived for the set of vehicles selected to provide coverage for each neighborhood under the Nash Equilibrium conditions for each vehicle based on its pareto optimality.

Algorithm 2 is implemented to find the best ROVERS  $V$  from the set of all vehicles  $\mathbb{V}$  in the network. Moreover, we implement STRIVE [24] as a social aware routing protocol in vehicular networks to forward the user interests to retrieve content from ROVERS.

We perform each simulation up to five times by analyzing different set of nodes as ROVERS in order to compute up to 95% confidence intervals.

The set of ROVERS as the best set of vehicles identified by CarRank are compared with the following schemes:

- Centrality-based selection: Benchmark social-aware approaches use Degree, Closeness, Betweenness, Eigenvector centrality, thus we implement each centrality scheme to select the best set of vehicles as ROVERS using Algorithm 2.
- Only Rank: approach where the best vehicles are selected based on their reputation/rank irrespective of the cost, coverage and quality requirements [9].
- Only Budget: as a cost minimizing approach where the service provider selects the vehicles which cost less irrespective of their reputation, quality and coverage requirements [14] [16] [17].
- Greedy Variant: an approach where the maximum coverage providing vehicles are selected in a greedy manner

Table III: CarRank in different set of Simulations

Simulation	1		2		3		4		5		Mean	Confidence Interval
Car Rank	ID	Score	ID	Score	ID	Score	ID	Score	ID	Score		
1	764	1	210	1	1179	1	271	1	36	1	1	0
2	1356	0.8182	178	0.6978	907	0.9566	295	0.9657	395	0.6212	0.8119	$\pm 0.13$
3	294	0.8177	298	0.6770	444	0.8568	595	0.8329	1902	0.5166	0.7402	$\pm 0.13$
4	46	0.7831	424	0.6012	2276	0.8511	1179	0.8116	1926	0.4410	0.6976	$\pm 0.15$
5	1454	0.7361	428	0.5701	682	0.8384	1926	0.7878	1147	0.4166	0.6698	$\pm 0.15$
6	169	0.7289	132	0.5420	2325	0.8285	2300	0.7709	46	0.4127	0.6566	$\pm 0.15$
7	969	0.7287	444	0.5361	653	0.8255	2064	0.7642	1384	0.3991	0.6523	$\pm 0.16$
8	1384	0.7185	270	0.5149	2527	0.8210	2436	0.7613	895	0.3903	0.6412	$\pm 0.16$
9	949	0.7174	39	0.4952	1581	0.7907	46	0.7375	2251	0.3559	0.6193	$\pm 0.16$
10	1115	0.7157	169	0.4934	399	0.7532	1386	0.6734	477	0.3367	0.5945	$\pm 0.15$

Table IV: User Interest Satisfied (UIS) by the best vehicles classified using CarRank compared with other centralities

Simulation	CarRank		Degree		Closeness		Betweenness		Eigenvect	
CarRank	ID	CSI	ID	UIS	ID	UIS	ID	UIS	ID	UI
10min	764	346196	764	346196	230	349	225	1115	764	34
20min	764	346196	230	349	469	311	947	345500	470	37
30min	764	346196	59	345051	469	311	1939	353698	282	37
40min	764	346196	145	351023	427	392	947	345500	469	31
50min	764	346196	969	354646	1750	348481	464	536	969	35
60min	764	346196	386	257	1233	343453	38	356	95	16

by the service provider while not taking into account the reputation, quality and cost of the recruited vehicles [10] [11] [12] [13] [15].

The vehicle utility  $u_v$  and the corresponding utility for the selected vehicles  $u_V$  is derived based on the vehicle centrality  $C_v$  obtained by each of the above scheme. The following performance metrics are used for evaluations:

- User Interests Satisfied: The user interest satisfied or content delivered is the key metric to evaluate the best identified and selected set of ROVERS.
- Throughput: Another important metric is the per node throughput achieved by the best set of identified and selected ROVERS.
- Cache hit rate: To evaluate the content collecting and storage capability of the best set of identified and selected vehicles, we use the cache hit rate of the selected ROVERS.

### B. Results: ROVERS Classification

Results showing the CarRank score of the best vehicles from five simulation runs using the realistic mobility traces from Cologne are shown in Table III. The number of best set of vehicles can differ with respect to different applications, though as an example, we highlight the top 10 vehicles after each simulation run with their respective CarRank score. The last two columns show the average score at each rank and the confidence interval for a confidence level of 95%. Trace results for all the vehicles are obtained from 5 simulation runs but we are interested to analyze only the best ranked vehicles. We normalize the CarRank score with respect to the best identified vehicle, i.e. the top vehicle will have unity score followed by the relative score of other vehicles.

We notice that the vehicle with ID 764 is identified to have the best ranked vehicle in the network. One reason is that it responded more frequently to the incoming interests and thereby

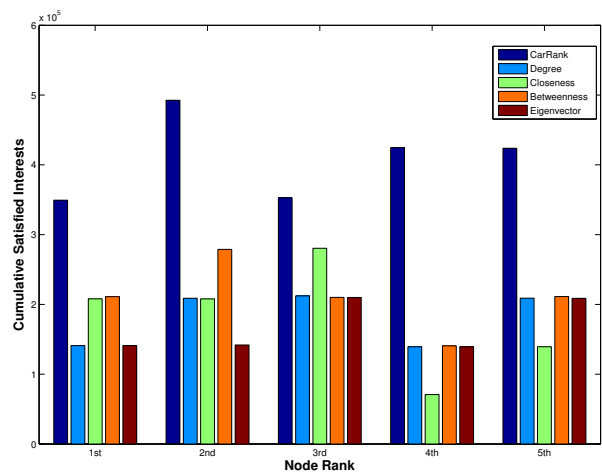


Figure 5: User Interests Satisfied by the five best vehicles classified using CarRank compared with other centrality schemes (Degree, Closeness, Betweenness, Eigenvector)

delivered more content for its associated neighborhood. At the same time, its spatio-temporal availability with respect to the content location and the neighborhood connectivity also contribute towards its CarRank score.

1) *User Interests Satisfied (UIS)*: We compare the user interest satisfied cumulatively by the best ranked vehicles classified using CarRank with the state of the art centrality schemes. Figure 5 shows such comparison for the best five nodes classified using each centrality scheme. Existing centrality schemes only considers the physical network topology towards computing the vehicle importance in the network while ignoring the user interests satisfied. Nevertheless, the best vehicles identified by CarRank delivered more content to users by satisfied more interests compared to other schemes in all the five set of simulations. It is because CarRank considers user interest satisfaction and important information as key factors towards finding the best ROVERS 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. The Table IV shows the User Interests Satisfied (UIS) 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

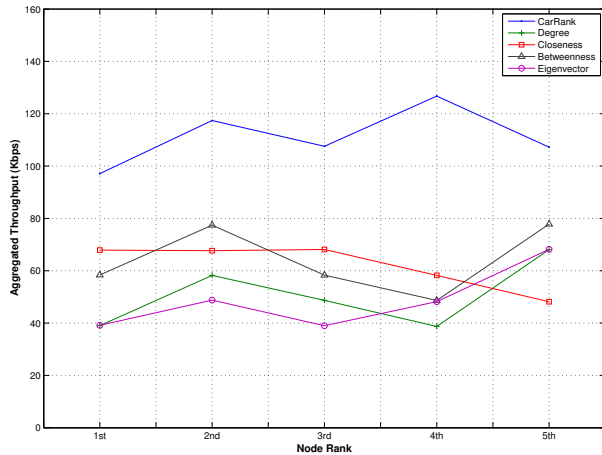


Figure 6: Throughput achieved by the best vehicles classified using CarRank compared with other centrality schemes

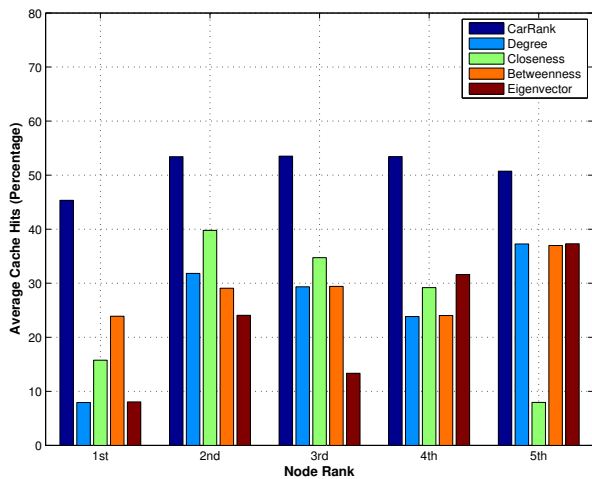


Figure 7: Cache hit-rate achieved by the best vehicles classified using CarRank compared with other centrality schemes

in the simulation. We observe some nodes with a higher UIS score than the node identified by CarRank. For instance, the vehicle with ID 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 III also reflects the importance of vehicle with ID 969 as it is classified as the 7th top vehicle by CarRank as well. The vehicle with ID 764 turn out to be an outlier as it is identified as 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 topological connectivity and spatio-temporal availability in the network to respond from the beginning till the end of the simulation.

2) *Throughput*: CarRank is also evaluated by studying the per-node throughput achieved by the best vehicles in

the network. Figure 6 compares the per-node throughput of the best vehicles classified by each centrality scheme. We compute the average throughput (kbps) where the best nodes classified using CarRank yield 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 followed by Degree, Betweenness and Eigenvector centrality. Closeness centrality on the other hand shows a decrease along the vehicles ranking order. CarRank outperformed all schemes due to the additional factors towards vehicle importance computation such as the content user relevant importance and the spatio-temporal availability, while other schemes rely only on topological measures (node degree or shortest paths) towards vehicle importance computation.

3) *Cache Hit-rate*: We evaluate the best identified vehicles for the en-route caching feature. We computed the cache hit rate at the five best vehicles classified using each scheme. A cache hit is measured when a vehicle successfully respond to a user interest for a content it already cached. The cache hit-rate for best classified vehicles identified by each scheme is shown in the Figure 7. The vehicles classified using CarRank yielded a relatively high cache hit-rate compared to the vehicles identified by the other centrality schemes. One reason is that with CarRank, the vehicle which cached en-route important information responds more frequently with the corresponding content to the incoming users' interests compared to other vehicles.

Moreover, the best node classified using CarRank tend to store more popular content as they belong to popular and better connected neighborhood and are spatio-temporal available for user interests in the network. This proves that en-route caching feature provided by ICN along CarRank efficiently caters the mobility and intermittent connectivity constraints posed by vehicular networks towards efficient urban data collection.

### C. Results: ROVERS Selection

In this section we discuss the results from applying the game along the Algorithm 2 to recruit the best set of vehicles as ROVERS. We first find the user interest satisfied by such ROVERS followed by their achieved throughput.

1) *User Interests Satisfied*: Figure 8 shows the user interest satisfied by the recruited set of best ROVERS in the network observed for a duration of one hour. In the figure to the left, we compare the set of best ROVERS selected using Algorithm 2 using CarRank as well as other centrality schemes. It is evident that CarRank outperformed other centralities with a user interest satisfaction of up to 75% around 30 minutes in the simulation. Similarly, it yields no less than 30% interests satisfaction during the entire simulation duration. It is because, we consider the user interests for the CarRank computation criteria besides the topological connectivity.

On the other hand, the famous Eigenvector centrality fails to select appropriate vehicles as set of ROVERS with respect to the user interests as it resulted in less than 20% satisfied interests. It is because it consider the most connected com-

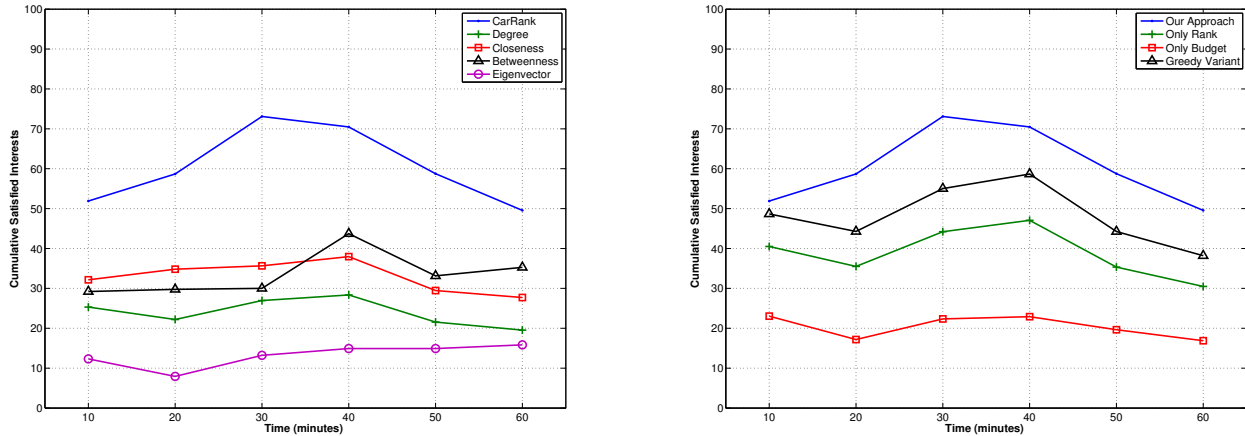


Figure 8: User Interest Satisfied by the best ROVERS compared by centrality scheme (left) and selection procedure (right)

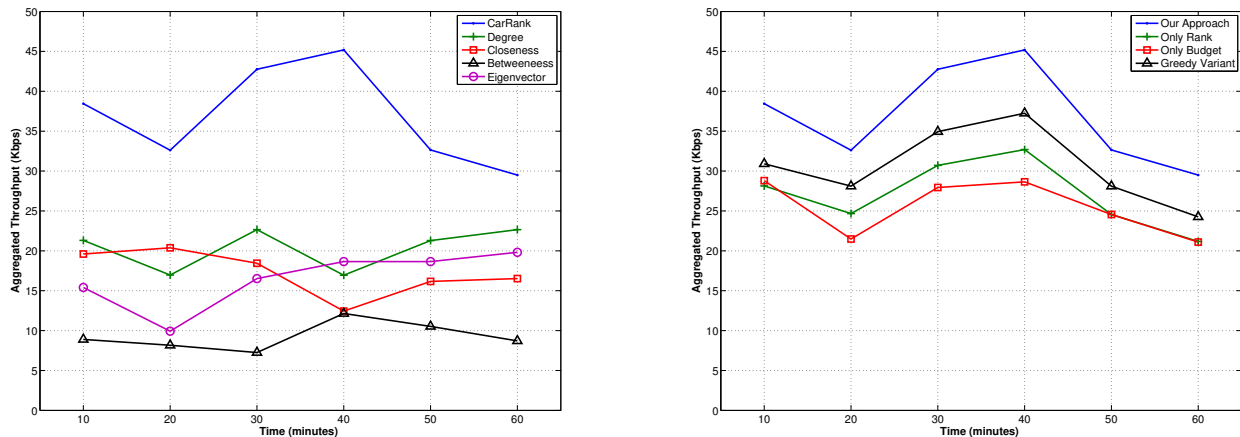


Figure 9: Throughput achieved by the best ROVERS compared by centrality scheme (left) and selection procedure (right)

ponents in the network as the most central nodes, thus only relying on the topological connectivity of the network.

From the comparison with centrality schemes, we can infer that CarRank is best suitable centrality metric to classify appropriate ROVERS with respect to the user interests.

Figure 8 (right) also shows the results by comparing our approach of selecting ROVERS with three other methods, i.e. solely rank based, budget-oriented, and a greedy variant preferring vehicles providing high coverage. We observe that our approach of selecting ROVERS yielded the best results (upto 75% at 30 minutes) with respect to the satisfied user interests. It is because besides the fact that the selected ROVERS delivered more content to the users, they were well connected and more related to their respective neighborhoods based on their commute or mobility pattern.

Our approach is followed by the greedy variant with maximum 60% interests satisfied at 40 minutes in the simulation where selecting the low cost vehicles as ROVERS resulted in the worst performance in terms of the user satisfied interests.

It is not necessary that the low cost vehicles can serve the users better. Thus, CarRank along the non-collaborative game is a feasible solution for selecting the best ROVERS in the network of vehicles. It is due to the consideration of user interests, connectivity, vehicle availability along catering the selfishness.

2) *Throughput*: We also compute the aggregated throughput of the best set of ROVERS selected using our approach compared with existing centrality schemes as well as other selection approaches. Figure 9 show such result where the figure to the left shows the comparison of the throughput of selected ROVERS classified using CarRank with those classified by other centrality schemes. We see that the temporal behavior of CarRank clearly outperforms other centrality schemes. It resulted in around 2x throughput with up to 45 kbps at 40 minutes and no less than 30kbps. It is because such vehicles can be relied upon as relays to provide information with better throughput in the network.

Degree centrality is second to CarRank with 20 – 25 kbps

throughput, where, contrary to common beliefs, Betweenness centrality resulted in a poor performance with less than 15 kbps throughput. It is due to the dynamic nature of the vehicular networks, betweenness centrality fails to identify a stable set of nodes to rely on for urban data collection.

Figure 9 (right) reveals the comparison results of our selection game with other possible approaches. We notice that our approach yielded overall the best throughput results with no less than 30kbps. The poorest performance is shown by using budget based selection where even the maximum throughput is less than 30kbps.

Thus, we can infer from the overall throughput results that our proposed selection game achieves the best performance. It is mainly due to the consideration of vehicles which delivered more content instead of considering only the network properties such as connectivity to identify and recruit the best ROVERS in the network.

### VIII. CONCLUSIONS AND FUTURE DIRECTIONS

This paper targets the identification and recruitment of the optimal set of socially important vehicles for large scale “user relevant” urban data collection. To do so, borrowing game-theoretic concepts, we present ROVERS for an incentive-based recruitment of the best vehicles required for urban data collection under a desired coverage, redundancy and quality requirements. Further, we allow smart vehicles to autonomously rank themselves based on user relative importance where a vehicle considers the importance of location-aware information, its neighborhood topology along its spatio-temporal availability (mobility pattern) to find its importance in the network. Results by comparing with different recruitment approaches revealed that ROVERS is best suited to efficiently identify and select important vehicles that satisfies up to 75% user interest compared to less than 50% for those recruited by other selection approaches. Similarly, the throughput achieved is 2x more than that of the compared approaches.

ROVERS enables us to recruit vehicles to collect large scale data from urban neighborhoods to assist different smart city applications. However, there is still a need to efficiently cache the gathered data at different city-wide information hubs to make it available for the users in an urban environments. Therefore, the focus of our future work is to propose distributed content caching schemes leveraging the recruited vehicles to efficiently store the collected data locally within the city.

### REFERENCES

- [1] E. Wang, Y. Yang, and K. Lou, “User recruitment for optimizing requester’s profit in self-organized mobile crowdsensing,” *IEEE Access*, vol. 6, pp. 17 518–17 526, 2018.
- [2] J. A. Khan and Y. Ghamri-Doudane, “Saving: socially aware vehicular information-centric networking,” *IEEE Communications Magazine*, vol. 54, no. 8, pp. 100–107, 2016.
- [3] R. Bruno and M. Nurchis, “Efficient data collection in multimedia vehicular sensing platforms,” *Pervasive and Mobile Computing*, 2014.
- [4] L. Kong, K. Ren, M. K. Khan, Q. Li, A. Rayes, M. Debbah, and Y. Nakamura, “Sustainable incentive mechanisms for mobile crowdsensing: Part 1,” *IEEE Communications Magazine*, vol. 55, no. 3, pp. 60–61, 2017.
- [5] X. Chen, M. Liu, Y. Zhou, Z. Li, S. Chen, and X. He, “A truthful incentive mechanism for online recruitment in mobile crowd sensing system,” *Sensors*, vol. 17, no. 1, p. 79, 2017.
- [6] J. A. Khan and Y. 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*.
- [7] B. Ahlgren, C. Dannowitz, C. Imbrenda, D. Kutscher, and B. Ohlman, “A survey of information-centric networking,” *Communications Magazine, IEEE*, vol. 50, no. 7, pp. 26–36, 2012.
- [8] K. V. Katsaros, V. Sourlas, I. Psaras, S. Rene, and G. Pavlou, “Information-centric connectivity,” *IEEE Communications Magazine*, vol. 54, no. 9, pp. 50–57, 2016.
- [9] S. Abdelhamid, H. S. Hassanein, and G. Takahara, “Reputation-aware, trajectory-based recruitment of smart vehicles for public sensing,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1387–1400, 2018.
- [10] D. Zhao, H. Ma, L. Liu, and X.-Y. Li, “Opportunistic coverage for urban vehicular sensing,” *Computer Communications*, vol. 60, pp. 71–85, 2015.
- [11] K. Han, C. Chen, Q. Zhao, and X. Guan, “Trajectory-based node selection scheme in vehicular crowdsensing,” in *Communications in China (ICCC), 2015 IEEE/CIC International Conference on*. IEEE, 2015, pp. 1–6.
- [12] Z. He, J. Cao, and X. Liu, “High quality participant recruitment in vehicle-based crowdsourcing using predictable mobility,” in *Computer Communications (INFOCOM), 2015 IEEE Conference on*. IEEE, 2015, pp. 2542–2550.
- [13] T.-Y. Yu, X. Zhu, and H. Chen, “Gosense: Efficient vehicle selection for user defined vehicular crowdsensing,” in *Vehicular Technology Conference (VTC Spring), 2017 IEEE 85th*. IEEE, 2017, pp. 1–5.
- [14] Y. Liu, J. Niu, and X. Liu, “Comprehensive tempo-spatial data collection in crowd sensing using a heterogeneous sensing vehicle selection method,” *Personal and Ubiquitous Computing*, vol. 20, no. 3, pp. 397–411, 2016.
- [15] W. Zong, Z. Liu, S. Yang, Q. Yuan, and F. Yang, “Multi-task oriented participant recruitment for vehicular crowdsensing,” in *International Conference on Internet of Vehicles*. Springer, 2017, pp. 92–104.
- [16] X. Zhang, Z. Yang, and Y. Liu, “Vehicle-based bi-objective crowdsourcing,” *IEEE Transactions on Intelligent Transportation Systems*, 2018.
- [17] X. Wang, W. Wu, and D. Qi, “Mobility-aware participant recruitment for vehicle-based mobile crowdsensing,” *IEEE Transactions on Vehicular Technology*, 2017.
- [18] M. Hu, Z. Zhong, Y. Niu, and M. Ni, “Duration-variable participant recruitment for urban crowdsourcing with indeterministic trajectories,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 271–10 282, 2017.
- [19] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, “Pricing and resource allocation via game theory for a small-cell video caching system,” *IEEE Journal on Selected Areas in Communications*, 2016.
- [20] M. Mangili, F. Martignon, S. Paris, and A. Capone, “Bandwidth and cache leasing in wireless information centric networks: a game theoretic study,” *IEEE Transactions on Vehicular Technology*, 2016.
- [21] Q. Zhang, H. Zheng, J. Lan, J. An, and H. Peng, “An autonomous information collection and dissemination model for large-scale urban road networks,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1085–1095, 2016.
- [22] S. Mastorakis, A. Afanasyev, and L. Zhang, “On the evolution of ndnSIM: an open-source simulator for NDN experimentation,” *ACM Computer Communication Review*, Jul. 2017.
- [23] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas, “Generation and analysis of a large-scale urban vehicular mobility dataset,” *Mobile Computing, IEEE Transactions on*, vol. 13, no. 5, pp. 1061–1075, 2014.
- [24] J. A. Khan and Y. Ghamri-Doudane, “Strive: Socially-aware three-tier routing in information-centric vehicular environment,” in *Global Communications Conference (GLOBECOM), 2016 IEEE*. IEEE, 2016, pp. 1–7.



**Junaid Ahmed Khan** (junaidakhan88@gmail.com) is currently a research fellow at the FedEx Institute of Technology and the Computer Science department, University of Memphis, TN, USA. Prior to this, he was a post-doctoral researcher at Inria Agora team, CITI lab at the National Institute of Applied Sciences, Lyon, France from October 2016 to January 2018. he has been a visiting research fellow at the Computer Engineering Department, University of California, Santa Cruz, June 2016 to October 2016. He has a Ph.D in computer science

at the Gaspard-Monge Computer Science Laboratory (LIGM-UMR 8049) University of Paris-Est, Marne-la-Vallee, France since August 2013. He has also been a visiting researcher at the University of La Rochelle, France, from December 2013 till September 2016. He completed his B.S. degree in telecommunication engineering from the National University of Computer and Emerging Sciences, Islamabad, Pakistan, in June 2009. In May 2013, he received a Master's degree in Electrical Engineering from the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan. His research interests are information-centric networks, Autonomous and Connected Vehicles, Fog Networking, Internet of Things (IoT), and complex/social networks analysis.



**Yacine Ghamri-Doudane** is currently a full professor at the University of La Rochelle and director of its Laboratory of Informatics, Image and Interaction. Before that, he held an assistant/associate professor position at ENSIE, a major French post-graduate school located in Evry, France, and was a member of LIGM-UMR 8049. From February 2011 to July 2012, he regularly visited the Performance Engineering Laboratory of University College Dublin, Ireland. He received an engineering degree in computer science from the National Institute of Computer

Science, Algiers, Algeria, in 1998, an M.S. degree in signal, image, and speech processing from the National Institute of Applied Sciences, Lyon, France, in 1999, a Ph.D. degree in computer networks from University Pierre & Marie Curie, Paris 6, France, in 2003, and a Habilitation to Direct Research in computer science from Université Paris-Est in 2010. His current research interests lie in the area of wireless networking and mobile computing, with a current emphasis on topics related to IoT, wireless sensor networks, and vehicular networks. He holds three international patents, and has authored or co-authored seven book chapters, about 25 peer-reviewed international journal articles, and more than 80 peer-reviewed conference papers. Since 1999, he has participated in several national and European-wide research projects in his areas of interest, among them two regional research projects, two national research projects, nine European-wide research projects as well as three EU COST Actions. As part of his professional activities linked to the computer networking research community, he also acted as Chair the IEEE ComSoc Technical Committee on Information Infrastructure & Networking from January 2010 to December 2013, and is currently chairing the IEEE ComSoc Humanitarian Communications Technologies Ad Hoc Committee (HCTC). He is an Editorial Board member of the IEEE IoT Journal, Springer AoT, Elsevier JNCA, and Wiley WCMC, is co-Editor-in-Chief of Springer/KICS CT Express, as well as the founding Editor-in-Chief of the IEEE ComSoc Ad Hoc and Sensor Network Technical Committee Newsletter. Among other conference involvements, he has acted as the TPC Chair of IEEE CCNC 2015, and Symposium Co-Chair of IEEE ICC 2009, 2010, 2012 and 2018 as well as IEEE GLOBECOM 2012 and 2015.