FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Wireless Protocols and Channel Estimation for Data Gathering with Mobile Nodes

Pedro Miguel Salgueiro dos Santos

Programa Doutoral em Engenharia Eletrotécnica e de Computadores

Supervisors: Professora Doutora Ana Aguiar and Professor Doutor João Barros
University of Porto, Portugal

May 30, 2017

Abstract

Data collection is a powerful application in scenarios where mobile and static agents work towards a common goal. Mobile and vehicular ad hoc networks (M/VANETs) are a steadfast platform over which such applications can be built. Static nodes interact with the M/VANET as data sinks that provide access to cloud-based services, or as sensor nodes that regard the mobile network as a dependable communication backhaul. The relevance of our target application – data collection applications over ad hoc networks with mobile and static nodes – motivates the development of network design solutions addressing scenario characterization, infrastructure planning and network operation. Our thesis aims to show that the design of such solutions can be improved by the use of measurements and datasets from the target scenario.

The development of wireless applications requires an accurate characterization of the electromagnetic signal propagation. Empirical channel models aim to capture the behaviour of propagation from measurements of received signal strength and distance between wireless terminals. We address the problem of path loss model parameter estimation in presence of erroneous distance measurements, in particular those obtained from the GPS positions. Our main conclusion is that the path loss model can be estimated with a reasonable accuracy from unreliable distances, provided that the measurements are taken at distances beyond a few standard deviations of the GPS positioning error. In case the maximum communication range does not allow such large distances, we provide a method to correct the erroneous channel model. Field experiments were undertaken to collect measurement data in order to validate our approach.

In a number of scenarios, static sensor nodes can harness vehicular backhauls for collecting data to a base station. Sensors and backhaul gateways can be interfaced by static communication hubs, and network designers can use mobility and connectivity datasets from the target scenario to place hubs ensuring service requirements and minimizing resources. We address the challenge of placing communication hubs over large areas (e.g. at city scale) and driven by infrastructure-to-vehicle (I2V) service requirements, alongside constraints of other nature. Our solution strategy involves an model of I2V transfers estimation over large areas that builds on an experimental characterization of throughput and data transfers at the target scenario. Our placement strategy attains less 20% hubs than sensor nodes, and estimates of our model of I2V data transfers fall within one order of magnitude of measurements collected on site.

The operation of protocols for data collection can harness base station-centric strategies such as beaconing. Protocols that set up structured routes (such as spanning trees) from beaconing are bound to suffer degraded performance as routing information at the nodes becomes outdated. We study an opportunistic design so that traffic does not become restrained to rigid routes. Given that link-level reliability becomes impractical, network coding is introduced to provide reliability. We set up a simulation framework over real-world connectivity traces and carry out extensive design-space exploration and benchmarking against a reference structured protocol. Our results support a number of design recommendations for a network coding-based protocol, and clarify the conditions in which our solution exhibits better resilience to routing information degradation.

Resumo

A coleção de dados é uma aplicação importante em cenários em que agentes móveis e estáticos trabalham para um objectivo comum. As redes sem fios ad hoc móveis e veiculares (M/VANETs) são uma plataforma robusta sobre a qual se podem construir tais aplicações. Os nós estáticos interagem com a M/VANET como recetores finais dos dados da rede ad hoc e encaminhando os mesmos para serviços na Internet, ou como nós de sensorização que vêem na rede ad hoc móvel uma plataforma de comunicação eficaz. A importância da nossa aplicação-alvo – coleção de dados em redes ad hoc com nós móveis e estáticos – motiva o desenvolvimento de soluções de planeamento de redes orientadas à caracterização do cenário-alvo, planeamento de infrastrutura e operação da rede. Esta tese procura mostrar que o desenho destas soluções pode ser melhorado através do uso de medições e dados do cenário-alvo.

O desenvolvimento de aplicações que operam sobre redes sem fios exige uma descrição precisa da propagação do sinal electromagnético. Os modelos de canal empíricos capturam o comportamento da propagação num dado cenário a partir de medições de potência recebida e de distância entre terminais. Nós abordamos o problema da estimação dos parâmetros do modelo de atenuação de propagação (*path loss*) na presença de medições de distâncias incorrectas, em particular distâncias obtidas a partir de estimativas de posição GPS. A nossa principal conclusão é que os parâmetros do modelo de atenuação podem ser estimados com precisão razoável a partir de distâncias incorretas, com a cautela de que as medições são obtidas a distâncias superiores a alguns desvios-padrões do erro de precisão do GPS. Para os casos em que o alcance máximo de comunicação não permite tais distâncias, providenciamos um método para corrigir o modelo de canal incorreto. A nossa abordagem foi validada com dados obtidos em experiências de campo.

Num conjunto de cenários, nós estáticos de sensorização poderão utilizar redes ad hoc veiculares para transportarem os dados recolhidos até uma estação-base. O interface entre os nós estáticos e veiculares pode ser assegurado por nós sem fios agregadores de comunicações (*hubs*), e os arquitectos da rede podem usar dados de mobilidade e conectividade do cenário-alvo para planear a localização desses nós agregadores observando requisitos de serviço e economizando recursos. Nós abordamos a tarefa de planear a localização desses nós agregadores estáticos em áreas amplas e guiada por requisitos de serviço nas ligações entre nós estáticos e veículos, a par de outros requisitos de diferente natureza. A nossa estratégia de solução envolve um modelo de estimação dos dados transferíveis em ligações entre nós estáticos e veículos em larga escala e que se baseia numa caracterização experimental da taxa de transmissão e volumes transmitidos obtida no cenário-alvo. O nosso procedimento para planeamento da localização utiliza menos 20% nós agregadores do que nós de sensorização, e as estimativas do nosso modelo de transferências de dados entre nós estáticos e veículos estão dentro de uma ordem de magnitude em relação a medições tiradas nos locais.

Na operação de um protocolo de coleção de dados de nós móveis e estáticos para uma estação-base, a sinalização iniciada pela estação-base é uma estratégia simples para indicar a direção para a referida estação. Protocolos que estabelecem rotas estáticas (p.e. árvores mínimas) a partir de

sinalizações periódicas poderão sofrer desatualização da informação de roteamento, resultando numa degradação de performance. Nós estudamos um desenho de protocolo baseado em roteamento oportunístico, para evitar a restrição a rotas rígidas. Visto que a fiabilidade de conexões entre nós vizinhos se torna impraticável, a codificação em rede é introduzida para providenciar fiabilidade entre nó de origem e estação-base. Nós montámos uma plataforma de simulação sobre dados de conectividade obtidos de uma plataforma veicular, e conduzimos uma exploração do espaço de desenho e comparação com um protocolo de referência. Os nossos resultados suportam uma variedade de recomendações práticas para o desenho de protocolos baseados em codificação em rede para cenários de mobilidade, e clarificam as condições em que a nossa solução exibe melhor resistência contra a degradadação de informação de roteamento.

Acknowledgements

First and foremost, I cannot help but thank my advisor Ana Aguiar. Prof. Aguiar helped me navigate through the tremendous endeavour that is the pursuit of the degree of Doctorate in Philosophy, pointing the way and providing the so-much-needed pragmatism and stoicism whenever skies were cloudier, and cheering and rooting whenever a few rays of sunshine and success filtered through the clouds. I'll be forever thankful of her permanent dedication and stimulating attitude, always daring me to tackle on new challenges and never to shy away from the pursuit of excellence. A word of appreciation is also due to my advisor João Barros, for supporting me with funding, helping me get a co-advisor, providing a good work environment and developing the testbeds on which much of the work in this thesis was performed.

Acknowledgements are due to the direct contributors of the work in this thesis, which I will go through in a chronological fashion. The first token of appreciation goes to Dr. Traian Abrudan, one of the first and happiest collaborations that I had the pleasure to participate in and that resulted in a high-profile publication (the first of my Ph.D.), second only to the long-lasting friendship that remained from our time as co-workers. (I won't forget those nice afternoons we spent in Viana do Castelo taking measurements in the forest!) Next on the list is Dr. Prof. Daniel Lucani, who provided the seminal design and subsequent inestimable contributions on the network coding-based data collection protocol. At the time of the initial discussions, the scenario was that of MANET composed by firefighters in a forest fire fighting situation, but the work was eventually transferred to the harbour vehicular network. In that regard, I must thank Rui Meireles for providing the GPS traces dataset that I used to develop the protocol and for the insightful discussions on vehicular routing. Finally, a big bundle of acknowledgements goes to the UrbanSense team. The relationship between the UrbanSense project and my thesis is one of those successful cases in which your input to the project aligns with your thesis work. Perhaps not evident at first, slowly the feeling grew that the points in common with the work I had been doing were far greater than the differences. Therefore, I am very glad that I had the opportunity to work with Tânia Calçada, the project manager that would not deter even when institutional coordination was at its most chaotic, and Prof. Dr. Susana Sargento from the University of Aveiro, who provided the necessary resources for the infrastructure-to-vehicle and delay-tolerant experiments from the vehicular network side. A big thank you goes to the UrbanSense team, starting with and in no particular order, Carlos Pérez-Penichet, now a Ph.D. student in Uppsala (good luck!), Yunior Rojo, now a Ph.D. student in Porto, Bruno Fernandes (the shrimp-fanatic man), Tiago Lourenço (the infamous-discussions guy), Daniel Moura, the M.Sc. students I had the pleasure to work with under the umbrella of UrbanSense, namely André Sá, Diogo Guimarães and Fábio Cunha, and finally to the people at VENIAM, particularly André Cardote, Tiago Condeixa, Diogo Carreira and José Julião. The quality of the work and benefits of the collaboration were more than validated when the team got a DTN demonstrator accepted in the ACM MobiCom 2015 conference and won a coveted Best Student Paper award at the IEEE flagship conference on smart cities (ISC2) in 2016.

Now we go to the people that were not direct contributors or co-authors, but who should not be mistakenly assumed to have had a minor role in this thesis' unfolding. First, I provide a one-size-fits-all thanks to everyone whose name is Rui, as pretty much every Rui I know from the academic community helped me with the NC data collection protocol. Besides the already-mentioned Rui Meireles, I must thank Rui Costa and Rui Prior for the very insightful discussions that we had on the subject and that provided great contribution. Many thanks also go to João Almeida and João Paulo Vilela, that provided incommensurable support when I posed them a question about a NP proof that arose during the DCU placement work. Special thanks go to Luís Pinto and Sérgio Crisóstomo who, along with Traian Abrudan, partnered with me in the supremely ambitious project of equipping every house with a dynamic escape route system, shortly after I entered the Ph.D.. Well, as we say in Portuguese, the dream died at the beach, or in other words, the ambition was way too large for the resources, but we got a nice demo going on that me and Luís had the opportunity to showcase in a number of occasions, and produced an article that had a tough birth and unfortunately did not fit this thesis. Many thanks also go (in no particular order) to the people at the Shannon Lab with whom I ganged up the most, chief amongst which João Rodrigues, Susana Cruz and Saurabh Shintre, who provided tremendous input to my thesis in our weekly or ad hoc meetings and got me through the times when the Ph.D. degree seemed the farthest. Alongside follow Tiago Vinhoza, Diogo Ferreira (the SoundBlaster), Maricica Nistor, Paulo Falcão, Alex Ligo, Emanuel Lima, Orangel Azuaje, Luísa Lima, António Rodrigues, Carlos Pereira, Hana Khamfroush, Mate Boban and his wife Sanja Sontor, Gerhard Maierbacher and Ian Marsh; the people from the NSG real-time lab, such as Prof. Dr. Luís Almeida, André Moreira and Luís Oliveira; from FEUP, Prof. Dr. Ricardo Morla, Prof. Dr. João Paulo Cunha, Prof. Dr. José Canas Ferreira, Prof. Dr. José Silva Matos, Prof. Dr. Maria Rosário Pinho, Inês Coimbra, Héber Sobreira and Bruno Ferreira; from IT, Sílvia Bettencourt, Cristiana Silva, Elimary Silva and Marta Meira; from the Department of Computer Science, Prof. Dr. Pedro Brandão, Pedro Gomes and Eduardo Soares; from the University of Aveiro and/or IEETA, Fábio Marques; from Cister, José Marinho and Ricardo Severino; and from the Carnegie Mellon University, where I spent close to three months, Prof. Anthony Rowe and his team at the time, Max, Patrick and Frank, and my roommate Miguel Araújo.

Finally, a great thank you to all the people who helped and supported me while I was hurtling down the highway of the Ph.D. journey, starting with my father José and my mother Clementina, that provided all the necessary conditions and neverending support for me to reach this point and tackle this challenge. My brother Tiago and my life companion Sara are also to be thanked, along with my grandfathers António and Fernanda, my grandmother Ana, that past away while I was still undertaking this journey, and the remaining family, of which I must highlight Conceição Santos, the aunt that helped me put everything into perspective. Great many thanks also to my friends, including but not limited to, Diogo Malheiro, Nuno Silva, Luís Carvalho, Gonçalo Rendeiro, Filipe Sousa, João(zão) Santos, Célia Soares, Helena Martins, and Cláudia Conceição.

The work in this thesis was partially funded by the Fundação para a Ciência e Tecnologia under grant SFRH/BD/67178/2009.

To my father.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We discuss in this chapter the motivation for the work of this thesis, its technical scope, and resulting contributions. The outline of the document concludes the chapter.

1.1 Motivation and Scenarios

In a multitude of human or robotic situations, a group of mobile and static agents works towards a common goal. A commanding agent or center can carry out coordination in real-time or process monitorization for later actuation. In both cases, it is paramount that the commanding agent acquires an accurate representation of the involved processes and agents. Dependable information systems, and in particular data collection mechanisms, play a vital role in this task. Data collection systems equip the commanding agent or center with the necessary information to issue orders according to the current situation and/or actuate over a scenario in a way that best serves the end goal.

Such information systems are facilitated, in mobile groups, by modern ad hoc communication technologies and protocols. A steady and relentless improvement in autonomy and processing power of embedded devices, and in channel and network capacity attainable by current wireless technologies, has been paving the way for mobile and vehicular ad hoc networks – M/VANETs. Given the real-world relevance of data collection and the emergence of platforms that support this functionality, it is of clear interest to advance the state-of-the-art and address open technical challenges in this realm.

The generic scenario we target is any setting where the following three types of nodes co-exist: (i) mobile nodes, such as vehicles, people, boats or UAVs, equipped with wireless transceivers that allow for device-to-device ad hoc communication; (ii) infrastructural base stations or sinks, deployed in the scenario and forwarding collected data to a local or remote commanding agent; (iii) static nodes with limited communication capabilities and that may or may not perform a communication-related task (e.g. sensing). We focus on the IEEE 802.11 standards as the base technology of the addressed ad hoc networks. We now illustrate three specific scenarios that were motivated by projects under which the work of this thesis was carried out. In the following

descriptions, we will identify potential applications that data collection can support, and identify the subjacent ad hoc network structure that data collection mechanisms can be built on.

Forest Firefighting

Firefighting has claimed the life of 61 people over the course of 10 years in Portugal [1]. Of these, 23 deaths were in the context of forest scenarios or heading to the site. From actual discussions with the professionals, we learned that firefighters deployed at fire fronts are organized in teams of five elements and a team commander, assigned to a fire truck. The commanding officer stays by the truck monitoring the situation, and the team members that move in to attack the fire assume one of two sets of functions: some elements carry the hose to the fire front, and other elements provide support or scout the surrounding area. Firefighters on the terrain might be subject to heavy smoke, hot temperatures, dense foliage and uneven terrain, eventually leading to some team members inadvertently straying from each other or running into high-stress situations.

Data collection tools can contribute for the team commander to learn the overall situation of the team members and issue orders accordingly. For example, gathering the location and vital signals in real-time from individual firefighters allows the commanding officer to identify stress situations and prevent accidents due to excessive fatigue. Additional information can come from sensor nodes dropped in the fire front during the last airtank fly-by. By learning the propagation vectors of the fire in real-time, the team commander can position the team and put down the fire as quickly as possible.

To collect this information, the firefighters carry mobile devices to record their vital signs and transmit it wirelessly. Equipments such as smartphones or embedded devices with WiFi (IEEE 802.11b/g/n) capabilities may double as graphical interface for body area network sensors, team members status or fire evolution vectors. Vehicles act as data sinks with powerful WiFi routers, providing the on-site commanding officer with the received information or relaying it to a command center via a more powerful communication technology (e.g. WiMax, satellite). The sensor nodes, deployed in an ad hoc or planned fashion and also featuring WiFi technology, can integrate with the firefighters into a single mesh network, or use the firefighters as data couriers in case communication to neighbouring sensors is erratic or inexistent. Figure 1.1 presents a depiction of this scenario, in which the firefighters carry embedded transceivers to reach the commander at the base station and the static nodes.

Urban Services

It is envisioned that the Internet of Things (IoT) paradigm will bring considerable efficient improvement to the operation of a city, with estimates of potential economic impact reaching US$ 1.7 trillion [2]. Big data and information systems will play a crucial role in supporting the envisioned services [3], many of which are mobility-related. There is a variety of applications and services leveraging on large-scale data collection in urban context. The characterization of traffic flow in the main arteries can be used by the municipality's traffic department to increase fluidity

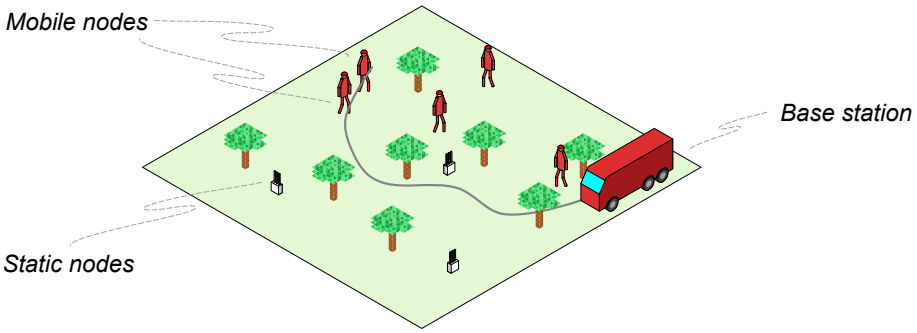


Figure 1.1: A forest firefighting scenario.

and prevent jams. The WiFi services in public transports can collect the origin and destination of users, and the public transportation authority can use these to assess travel demand with higher precision and reassign routes or bus frequency. Collecting the usage level of garbage bins allows the waste disposal department to recompute garbage collection routes and minimize truck wear.

Drawing on a particular instance, the IoT-based smart city platform *UrbanSense*, in Porto, includes a set of weather stations to record air quality (NO_2, O_3), meteorology (wind vane and speed, rain gauge) and life quality (noise, luminosity, UV radiation) metrics. These stations have been placed throughout the city with the goal of capturing episodes of anomalous variations in the observed metrics as well as creating long-term spatial and temporal descriptive models of the measured processes. Acquired data can be processed to obtained meaningful insights of the city's climate, alert the municipality's environment and life quality department when critical levels are reached, and motivate urban policy or design options to improve the dwellers' quality of life. A data collection strategy is necessary to support the collection of sensor data from disparate locations.

Some city services depend on dedicated fleets (e.g. garbage collection, public transportation) that can be equipped with vehicular communication capabilities to form a mobile network. On-board units (OBU) installed in the vehicles and road-side units (RSU) deployed throughout the city, both equipped with DSRC/IEEE 802.11p standard, enable vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication. WiFi access points (IEEE 802.11b/g/n) hosted by OBUs of some fleets (buses, in particular) benefit passengers while also being available to external clients. Road-side clients may connect opportunistically to the mobile access points to off-load relevant information, which is then routed to the corresponding command centers (e.g. public transportation authority, waste disposal and/or environment departments). If no real-time connection is available or the data is not time-sensitive enough, store-and-forward networking and services can be supported by the OBUs. An example scenario is shown in Figure 1.2, containing examples of data-producing equipments (garbage bins, weather stations, interactive ad boards), and wireless-enabled nodes such as outdoor access points and vehicular hotspots.

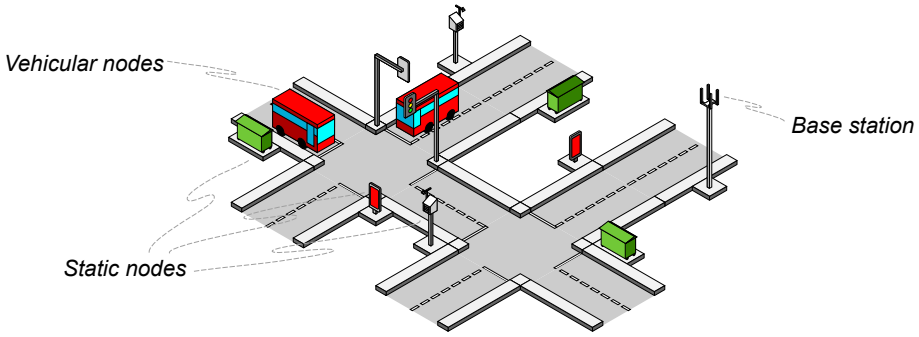


Figure 1.2: An urban scenario.

Port Operation

The revenue of container ports and ships is directly associated to its operational efficiency [4]. As a consequence, a considerable body of operations research addresses the minimization of vessel turnaround time at a port [5]. Vessel turnaround times depend on an efficient real-time resource allocation for container loading and unloading. In this perspective, it is necessary that the trucks that carry containers between loading areas of the premises (from and to ships, trains or road-legal trucks), know exactly where to head next for loading or unloading. This motivates the need for a command center and a communication network that supports a real-time information system.

Both container trucks and ships can be equipped with on-board units for V2V and V2I communication. With this setup, ships can transmit information to the trucks about the load they are carrying, expecting or prepared to load off onto the trucks. Road-side units installed throughout the premises link the vehicular network and the command center, which in turn can produce an assignment of trucks to handle the docked ships and issue those commands.

The relevant information from the DSRC-enabled mobile and static nodes can be routed within the multi-hop inter-vehicle communication towards the road-side unit, and then forwarded to the command center. This particular port scenario offers as much opportunities as hindrances to the operation of a data collection protocol. Wide areas of water allow for unobstructed communication, whereas the ever-changing walls of metal containers through which the trucks move incur in high multipath. A protocol that aims to collect data from trucks to the road-side units must cope with the variety of paths available at any given time. Refer to Figure 1.3 for visualization of this scenario, in which data from vehicular and static nodes must reach the base station.

1.2 Thesis Scope and Claim

We can formalize the object of analysis of this thesis in the following manner:

Data collection over ad hoc networks with mobile and static nodes

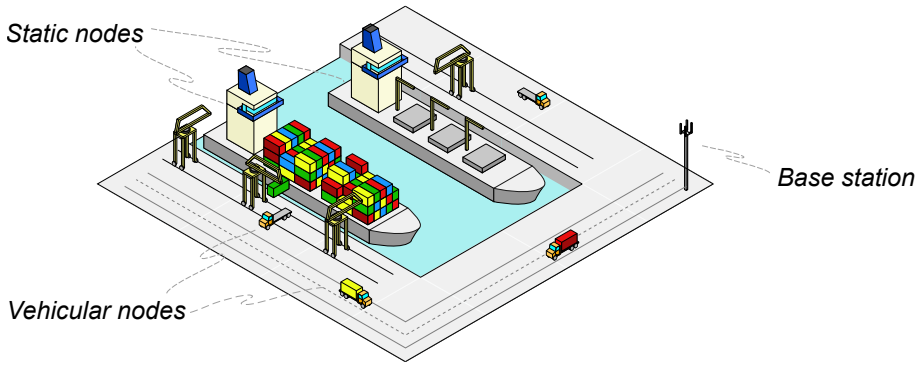


Figure 1.3: A container port scenario.

After having motivated the relevance of such application in the previous section, we now discuss potential technical advances on network design and operation and focus on the particular challenges this thesis addresses. We restrict ourselves to existing physical and data link layers, particularly IEEE 802.11b/g/n. Our target application can be broken down into three building blocks, which in turn raise challenges in three aspects of network design and planning:

Application		**Network Design**
Ad Hoc Networks	$\rightarrow$	Scenario Characterization
Mobile and Static Nodes	$\rightarrow$	Infrastructure Planning
Data Collection	$\rightarrow$	Network Operation

We discuss the nature and challenges of each facet of network design mentioned above, and identify a specific research topic that sets the scope of the work and contribution in this thesis.

1. Ad Hoc Networks $\rightarrow$ Scenario Characterization

An ad hoc network presumes the existence of a networking application on top of ad hoc links – links that can be initiated between wireless devices without intervention of a support infrastructure. At the physical layer, these data layer links are substantiated by electromagnetic propagation over wireless device-to-device channels. The response of such channels may vary from scenario to scenario due to various factors such as the existence or not of obstacles, reflective surfaces and sources of electromagnetic noise.

Propagation over wireless channels can be described by empirical models generated from measurement data. Such models are relevant as they inform the application designer of the communication range and propagation behaviour in a particular scenario, therefore constituting an important aspect of network design and planning. Empirical channel models typically break down propagation into two distinct phenomena: large-scale attenuation over distance (e.g. the path loss,

dual-slope and two-ray ground models) and small-scale fading (described as random processes following Rayleigh, Gaussian or Weibull distributions). Given that the model parameters are estimated from measurement data, two different settings may result in different parameter values for the same model.

Therefore, we identify the following research topic:

Propagation models for device-to-device channels

2. Mobile and Static Nodes → Infrastructure Planning

In the scenarios being addressed, we assume the co-existence of mobile and static nodes that are wireless-enabled. The static nodes produce data that can be collected by vehicular nodes, delivered at a road-side unit and forwarded to a backend server. Assuming freedom to place the static nodes, an important aspect of network design hinges on judicious infrastructure planning, as service requirements must be guaranteed while resources may be saved with efficient placement strategies.

Strategies for static node placement driven by I2V service rely on models of I2V data transfers. In turn, such models must build on a characterization of transfer rates and volumes, with respect to distance and speed, between terminals drawn from measurement data taken at the target scenario. The models thus created can be used to predict and/or estimate I2V service over a large area and support the static node placement application.

The second research topic in this thesis is:

I2V service characterization and static node placement

3. Data Collection Application → Network Operation

The data collection application itself is carried out by upper-layer protocols, as a collection protocol encompasses functionalities from the routing and transport layers. In fast changing topologies, such protocols have to cope with mobility-induced challenges such as additional packet loses due to volatile links and packet mis-routings.

Protocol development by means of simulation is often based on artificial mobility traces or models and generic propagation models. The use of propagation, mobility and connectivity data from the target scenario may improve considerably the evaluation and efficiency of the protocol when applied in the real-world.

The final research topic we address in this thesis is:

Data collection protocols over dynamic topologies

Thesis Claim

In all three aspects of network design, we observe that characterizing the scenario by means of measurement data is crucial to improve the operation and planning of the network and infrastructure to support the target application. In summary, we can claim that:

- **Scenario Characterization**: An accurate characterization of the wireless propagation at the target scenario allows to build better performing protocols.
- **Infrastructure Planning**: The use of datasets and/or measurements from the target scenario is instrumental towards a planning solution that is not underperforming nor over-dimensioned.
- **Network Operation**: Measurement data from the scenario, such as mobility or connectivity traces, may support the development of protocols to cope with fast-changing topologies.

The statement of this thesis draws from the transversal line of reasoning delineated throughout the previous claims:

> The design and planning of network operation and infrastructure
> can be improved through the use of measurement data from the scenario.

1.3 Research Challenges and Thesis Contributions

The research topics identified in the previous section determined the initial lines of work pursued in this thesis. As each line of work advanced, the seminal objectives evolved into concrete research challenges that posed innovative and pertinent problems in the light of the state-of-the-art of the corresponding field. We now go through the scope and rationale of each research challenge that we identified and tackled in the course of this thesis, and the contributions that resulted from each line of work.

Propagation Models for D2D Channels – Impact of Erroneous Positioning

We conducted field measurements to estimate model parameters of a device-to-device channel in a forest setting, for the purpose of describing wireless propagation in such setting and support protocol development. The necessary dataset is created by pairing contemporaneous position estimates and RSSI samples from one or both terminals. The position estimates were acquired with the Global Positioning System (GPS), which has steadily become the standard tool of the wireless modelling community to obtain location estimates due to its simplicity of use and wide-spread availability. However, we note that the Global Positioning System does not provide absolutely correct estimates at all times, and a crucial aspect driving the accuracy of GPS position estimates is the equipment quality. High-precision devices, with errors in the range of centimeters, are expected

to cost thousands of Euros, whereas lower-end GPS chipsets and software, such as those found in most consumer-electronics products (laptops, smart phones and lower-end GPS receivers), can be expected to have errors in the range of tens of meters. The impact of this uncertainty has been often disregarded in the measurement methodology of experiments to obtain datasets for model parameter estimation.

We set out to understand exactly how position errors impact the model extracted from the measurement pairs that use GPS, in what conditions can the largest errors be expected, and what can be done to prevent or mitigate *a posteriori* these errors. Overall, our contributions on device-to-device channel estimation using GPS are the following:

- a model of the impact of GPS positioning errors on ranging and path loss model estimation;
- a method to improve path loss estimation using only GPS distances;
- guidelines for designing measurement campaigns for path loss model estimation that reduce the impact of GPS errors.

This work, discussed in Chapter 3, has been published at IEEE Transactions on Wireless Communications [6].

I2V Service Characterization – Static Node Placement driven by I2V Service

Mobile and vehicular ad hoc networks are bound to interact with static elements, as depicted in the scenarios of Section 1.1. These may be infrastructural (road-side) or temporarily-deployed elements, and may serve or not a communication purpose. With the ever-growing number of communication-capable vehicles and wide-spread use of personal devices, the use case of static units that regard mobile nodes and networks as dependable communication platform to reach the Internet is becoming a reality. Vehicular fleets in which nodes take pre-defined routes with known frequency further increase their dependability and reliability as a communication backbone. During the development of a vehicular-based collection solution for urban road-side sensors, we characterized infrastructure-to-vehicle (I2V) communication by evaluating throughput and data volume performance of wireless links between mobile and static terminals, and studied how speed and distance impact their performance. We soon realized that the performance of I2V links and service by the vehicular network depends greatly of a careful and thorough placement planning of the road-side client nodes, and that this task implies innovative challenges. In the academic community, the RSU placement problem is mostly motivated in the perspective of the RSU as gateway to a large number of vehicular nodes. Our scenario addresses a slightly different take on this, as the RSU must maximize the I2V data volumes transferred to *any* vehicular node, which may not necessarily require strategies to reach large sets of nodes.

We developed a support framework that carries out road-side client node placement driven by, among other requirements, quality of service by a vehicular network. For this particular purpose, we created a procedure to estimate the data volumes that can be transferred in I2V links at potential

deployment locations. Our contributions on vehicular-infrastructure wireless interaction are the following:

- characterization of I2V connectivity between a road-side static wireless client and a fleet of access point-equipped public buses;
- formulation of a minimal placement problem for road-side client nodes and proposal of a solution strategy applied to a real-world scenario;
- a method to obtain a city-scale estimation of transferrable data volumes in I2V connections;
- validation of the placement solution against a real-world deployment in Porto.

This work is described in Chapter 4. The initial characterization experiments at a prototype DCU were published at ACM MobiCom 2015 Workshop on Challenged Networks [7], and the subsequent placement strategy has been through a first round of reviews in a submission to ACM Transactions on Sensor Networks [8].

Data Collection in Dynamic Topologies – Design-Space of a Network Coding Protocol

The last line of work in this thesis tackles the network operation perspective to support our target application. In a scenario of base station-driven collection from mobile/vehicular nodes, we hypothesize that an opportunistic protocol is a better fit for data collection than a protocol with fixed routes. We explore the use of wireless broadcast and probabilistic forwarding to propel the data packets towards the base station, and of network coding as a end-to-end reliability mechanism in the absence of link-wise reliability. We expect that, in the absence of link-level retransmissions, the additional packets produced by the wireless broadcast/opportunistic forwarding mechanisms will contribute to the requirement that enough coded packets reach the base station.

We developed a framework protocol for data collection in mobility scenarios that allows to carry out extensive design-space exploration of the forwarding and network coding mechanisms, shedding light over what works and what does not, and in what conditions. Our contributions are the following:

- An identification of the design and parameter aspects that a network coding protocol implies, and an analysis of the existing literature on those design aspects;
- Performance evaluation of the protocol under alternative implementations, by means of simulation over real-world connectivity traces;
- Performance comparison against a benchmark structured protocol, CTP.

This work is presented in Chapter 5 and a publication is under preparation.

We conclude this section by discussing the relationship between the contributions of each structural line of this thesis and the design of networks for data collection over ad hoc networks with mobile and static nodes. As visible in Figure 1.4, the first two parts of thesis (*Propagation*

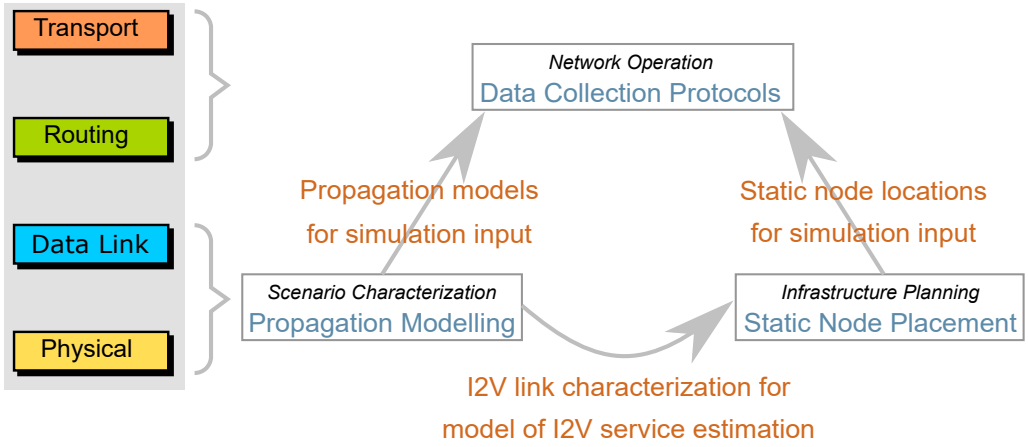


Figure 1.4: Role and relationship of contributions and components towards network design

Modelling and *Static Node Placement*) improve relevant aspects of the scenario description that can later be fed to simulation tools to develop tailored protocols.

1.4 Thesis Structure

The remainder of this document is as follows. In Chapter 2, we review the relevant literature for the work presented in this thesis. In Chapter 3, we analyze the impact of errors inflicted by the Global Positioning System in position estimates used for parameter estimation of device-to-device channel models. We start Chapter 4 by characterizing I2V links in an urban setting, and evolve to a discussion about the requirements that the placement of a platform of static nodes, wishing to rely on vehicular fleets for data transport to the cloud, must observe. In Chapter 5, we describe a data collection protocol that pairs opportunistic routing and network coding, and benchmark it against state-of-the-art protocols over real-world mobility traces. Finally, in Chapter 6, we draw some final remarks and sketch future lines of work.

Chapter 2

Related Work

This chapter is dedicated to presenting the state-of-the-art on the research topics discussed earlier. In Section 2.1, we review the existing channel models of electromagnetic propagation. The literature on characterizing infrastructure-to-vehicle (I2V) links and service is reviewed in Section 2.2. In Section 2.3, we provide a revision of the literature on routing and data collection techniques in M/VANETs and WSNs. Some final remarks are drawn in Section 2.4.

2.1 Path Loss Models for Wireless Propagation

We start by reviewing the existing literature concerning empirical models for wireless communication channels and their estimation from experimental data. Afterwards, the operation of GPS and its use in such campaigns is discussed.

2.1.1 Channel Models and Measurement Campaigns

The log-distance path loss model [9] is one of the most widely used models to describe the behavior of radio wave propagation. A relevant improvement on the simple path loss model is the Two-Ray Ground model [10], that takes into account signal reflection in the ground. Small-scale fading is often modelled by means of well-known distributions, such as Gaussian [11], Weibull [12] and Nakagami [13]. The associated path loss exponent and fading distribution parameters are estimated from empirical measurements. Many of the works on propagation and channel modelling address outdoor communication between a mobile user and a base station, which is usually tall or positioned in a high location. This is the case in [14], which presents path loss, scattering and multipath delay statistics for digital cellular telephony, measured in urban context with distances in the range of 1.5 to 6.5 kilometers. The channel models in [15] are estimated for the 5.3 GHz range in urban mobile communications, with distances up to a few hundred meters.

There is a wide body of work regarding vehicular channel model proposals and measurement campaigns in vehicular environments. The work of [16] provides a good overview on existing models for vehicular channel and propagation models. In [13], the authors report V2V narrowband channel measurements in the 5.9GHz DSRC band, in suburban environment. The authors

fitted the collected measurements with single and dual slope log-distance path loss models. Also at the 5.9 GHz band, the authors of [17] conduct car-to-car measurements with wideband channels (20MHz) in rural, highway and urban environments. The authors use a regular log-normal path loss model for the urban scenario, whereas for the rural and highway a two-ray ground model is used. The authors of [18] carry out measurements to obtain the impulse response of the channel and derive time-varying components. In [19], the authors carry out vehicle-to-vehicle measurements in four scenarios: highway, rural, urban and suburban scenarios. The authors arrive to the similar observation that the urban data is best characterized by a simple power law, whereas the rural follows a two-ray ground model. The authors of [20] present the interesting result, support by experimental data, that the application of simplified two-ray ground path loss models to vehicular simulations yields no significant increased value with respect to the free-space model, in most cases. A more sophisticated model, the two-ray interference model, is proposed for more accurate description of V2V communication.

A number of works address specifically the impact of obstructions in vehicular communication, in which a common strategy consists in proposing separate models for the line-of-sigh (LOS) and non-line-of-sight (NLOS) conditions. The authors of [21] propose a 5.9 GHz NLOS path loss and fading model specifically for intersections and estimated from field measurements. Communication under LOS is modelled by a log-distance model, whereas for NLOS it is modelled by a geometry-based model that takes into account the distances. Vehicular obstructions have also been shown to impact significantly V2V communications [22] in a manner that had not been captured previously in channel models. The authors of [23] propose a geometry-based model to incorporate the impact of vehicular obstructions. The model is developed using realistic datasets and validated against experimental measurements.

2.1.2 Overview and Application of GPS to Measurement Campaigns

The Global Positioning System (GPS) [24] has become one of the most widely used technologies for obtaining positioning information in outdoor scenarios. The system itself is composed by a Master Ground Control Station and a constellation of satellites revolving around the Earth. The satellites continuously send a signal that user equipment on the Earth surface can receive. The signal contains a code, called ephemeris, that carries the identification, clock, orbit and position in orbit of the satellite. The distance to the satellite, called pseudo-range, is proportional to the phase shift between the received signal and an internally-generated replica. If ephemeris and distance information are available for four or more satellites, the user location can be triangulated.

The accuracy of GPS is affected mainly by two factors: the geometry of the satellites visible to the user, and the quality of the pseudo-range estimates. Concerning the geometry of the satellites, a good distribution of the satellites in the sky will minimize the space of possible user locations. Regarding the pseudo-range estimates, they are affected by a variety of errors that are caused by the user equipment and the GPS system. On the user equipment side, the quality of the receiver defines the precision with which the phase shift between the received code and the internally-generated replica can be measured. Causes external to the equipment are the attenuation of the

satellite signal by the ionosphere and troposphere, and the noise and multipath effects caused by the environment surrounding the equipment. On the GPS system side, the validity of the data contained in each satellite's ephemeris slowly fades with time. Consequently, as time passes, it becomes increasingly outdated and prone to cause larger errors. Moreover, the satellite clock may experience shifts, which also induce errors.

Some empirical studies address the performance and accuracy of consumer-grade GPS receivers in forest settings. In [25], six commercial receivers in the range of $150 to $320 (values of 2005) were tested in static conditions. Measurements were taken in open sky, under young forest canopy, and under closed forest canopy. For the later scenario, the reported values of average positioning error vary between 2.7 and 11.4 meters, depending on the device. The work in [26] provides an updated version of these results, but only for two devices and under heavy canopy conditions. Reported values for the average positioning error were 4.0 and 6.9 meters. Studies about GPS position accuracy in smart phones are scarce. One example is the work of [27], which analyses the location accuracy of the iPhone obtained from different information sources (GPS, cell ID, WiFi) in different settings (rural and urban, and indoor).

In the measurement campaigns reviewed in the previous section, the user/base-station campaigns [14, 15] used the map-based method is used to determine the distances precisely. Recent work has addressed the impact of GPS errors in parameter estimation from crowdsourced data in cellular links and coverage estimation [28]. The VANET propagation studies we reviewed [13, 17, 18, 19, 20, 21, 22, 23]. report the use of GPS for distance calculation. The work described in [21], that addresses the particular scenario of intersections (and thus deals with short ranges and inter-vehicle distances), report that the vehicle positioning was improved with the car sensors and map matching. The authors of [19] refer that GPS position estimates result in inaccurate distances when the actual distance between terminals is small, which confirms our conclusion. Only two of the mentioned works [13, 18] report having used Differential GPS and the WAAS (Wide Area Augmentation System). This system uses additional information when available to improve the accuracy of the GPS estimates to sub-meter precision. The remainder of the works does not address the reality of GPS errors and their impact in model estimation.

2.1.3 Discussion

In the context of device-to-device propagation modelling, we observe that the current literature reports extensive use of the Global Positioning System (GPS) for obtaining position estimates. GPS is one of the most widely used positioning technologies, but it is subject to numerous sources of errors that affect its position estimates. We verify that most recent works addressing characterization of wireless device-to-device channels in M/VANETs fail to account the impact of GPS positioning errors in the estimation of the path loss model. This motivates the work done in Chapter 3.

2.2 I2V Link and Service Characterization

We start by reviewing the literature regarding the link level characterization of channels in the context of terminals with relative speed between them. We proceed to discuss tools for estimating data volume transfers from road-side static clients to vehicular nodes and for optimal placing of those road-side nodes.

2.2.1 Link-Level Characterization of V2X Channels

We will focus on IEEE 802.11a/b/g-based V2I and I2V communication, as it is the context of our work in Chapter 4. Most V2I/I2V studies using this technology address the specific application of providing Internet access to mobile nodes. Typical driving scenarios explored in these studies are urban, suburban and highway, with particular incidence for the last one [29, 30]. Some works also study the association time to Access Points (APs) and IP assignment time [31, 32, 33]. To study IEEE 802.11a/b/g-based V2I communication, the authors of [29] equipped a car with a IEEE 802.11 a/b/g WIFI client and placed an AP in the middle of a two-kilometer highway stretch. UDP and TCP measurements were performed with different payload sizes, transmission rates and vehicle velocities. At a velocity of 120km/h, UDP throughput reached up to 35 Mbits per second (Mbps) for payload sizes of 1250 bytes, during a 400 meter long stretch centered at the server. Results for other velocities are similar, showing that speed has little impact in throughput. The authors of [30] also address V2I in an highway scenario. Throughput and PLR between a mobile node and a fixed node using the IEEE 802.11b standard are evaluated at speeds of 80, 120 and 180 km/h, over all distances at which a connection exists. The IEEE 802.11b standard is used and UDP and TCP streams are tested, alternating the fixed node and the mobile node as senders. The authors report a 200 meter window around the fixed node in which link throughput can reach the nominal value and PLR can be almost to zero, if the mobile node is sending (close to 5 Mbits/s). Transmitted data volume was 9 MBytes. The work described in [32] also uses IEEE 802.11b for V2I communication and reports similar values. The authors set up a fixed AP in a radiation-free zone (desert) and evaluate AP range, association times, packet losses (using UDP) and throughput of UDP, TCP and web traffic connections, for a range of velocities (5, 15, 25, 35, 55 and 75 miles per hour). A data volume of 6.5 MBytes is reported at 75 miles per hour. It is concluded that speed has little impact on packet loss and throughput, and that the main factor limiting transfered data volumes is the connection lifetime, which is lower for higher speeds. The authors of [31] also tested V2V and V2I communication using IEEE 802.11b in a radiation-free zone (desert). A V2I experience is described in which a vehicle is moving circularly with respect to a fixed AP at different radius, at speeds up to 130 km/h. It is observed that packet loss rates, jitter and number of retransmissions at the MAC layer remain close to zero.

2.2.2 Placement Strategies Driven by Large-Scale I2V Service Estimation

The discussion in this section is two-fold. We first discuss works or models that aim at large scale characterization or estimation of service quality between mobile and infrastructural nodes. Afterwards, we address the literature placement strategies for road-side nodes and for coverage of target points.

Regarding the first topic, the work described in [33] presents a large-scale study (9 cars, 290 hours) on the feasibility of using domestic and commercial AP to provide Internet access. Personal cars were fitted with IEEE 802.11 clients that would search for APs and would, when possible and progressively, attempt association, IP assignment, and Internet connection. Results include connection duration and setup latency, impact of speed on AP association and connection duration, and packet losses in the link between AP and client. The work presented in [34] presents a theoretical evaluation of the capacity and coverage of various technologies (cellular and vehicular) to support infrastructure-to-vehicle communication at large scale. The work of [28] discusses coverage estimation from cellular towers within the scope of Minimization Drive Tests (proposed by 3GPP), that seek to crowdsource user RSSI and position samples to support propagation estimation. The authors of [35] describe CARM, an algorithm to generate RSSI maps from crowdsensed datasets. The authors identify as a major problem the lack of calibration (or error model of) of RSSI sample values from devices of different makes. The algorithm attempts to estimate simultaneously the parameters of unknown RSSI measurement error model and signal propagation model.

Regarding the problem of optimal placement for road-side clients to bridge pre-deployed sensor units and a vehicular network, discussed in the Chapter 4, it is relevant to review the placement literature on two research areas: sensing coverage/sensor placement in sensor networks, and road-side unit placement in vehicular networks. In the field of wireless sensor networks, coverage problems are broadly classified into area coverage, point coverage, and barrier coverage problems [36]. Our problem can be seen as an instance of point coverage problems, in which a placement solution for a set of sensors that covers a set of target points must be found. The most common goal is network cost minimization [37], i.e., to minimize the number of sensors (to be located also at vertices) necessary to cover the set of target points. The authors of [37] model the problem as an Integer Linear Programming (ILP) problem and solve it using a LP solver, which may be a computation-demanding task for large fields [38]. An alternative way of solving the network cost minimization ILP problem is by equating it to the combinatorial set-cover problem [39], known to be NP-hard but for which greedy heuristics may provide near-optimal solutions. The works in [40, 41] propose greedy algorithms that take into account sensor imprecision and the presence of occasional obstacles. In [38], the authors propose a greedy algorithm to handle the existence of multiple types of sensors with different cost and range, and in [42] algorithms that approximate the optimal solution within some δ are presented.

In the context of vehicular networks, the goal of optimal RSU placement is to find a placement solution that optimizes the performance of some V2I or I2V communication metric while minimizing the number of RSUs.Common target metrics are probability of V2I contacts [43], number

of V2I contacts and their duration [44, 45], delay in reporting an event [46, 47], or the average trip time at city-scale provided by an information dissemination system [48]. In [44], the authors address RSU placement (at intersections only) for information dissemination addressing maximization of the number of contacts between vehicle and RSU and service time by the RSU. The problem is formulated as a Maximum Coverage Problem (MCP), and real-world traces are used for performance comparison. The work presented in [45] extends this approach by discretizing the city into finer-grained cells and defining migration ratios among them, obviating the need to know the trajectory of every single vehicle as in [44]. Using realistic vehicle traces, the authors of [43] discretize the city in zones and compute the transition probabilities of vehicles between zones. In [48], the authors propose a VANET-based traffic information system for minimizing trip times. RSU placement is performed by a genetic algorithm from a pool of tentative locations, in which the fitness function assesses the reduction in the mean vehicular trip times. The authors of [49] present topological metrics to assess the centrality of a node in a graph and select the best nodes for RSU placement. A quantitative comparison using simulations is presented.

2.2.3 Discussion

Regarding the work presented in Chapter 4, we looked into the existing state-of-the-art on V2X channel characterization, application-driven service quality by vehicular networks, and strategies for placement of road-side static clients. On the first point, there is a large body of work regarding the characterization of V2I/I2V communication, detailing in particular the impact of speed and distance on throughput of such channels. Regarding the characterization of the service quality provided by a vehicular backhaul to road-side static nodes is object of a smaller set of literature. We contribute to these fields by undertaking I2V measurements between a road-side WiFi client and a large-scale vehicular network that: (i) corroborate the literature conclusions regarding the impact of speed on throughput; and (ii) present the attainable data volumes in a real-world scenario of vehicular data collection. Regarding the last topic, we tackle a system design problem concerning the placement of the road-side clients in a way that: (i) optimizes service by the vehicular network; and (ii) guarantees service to an end-system (a monitoring platform). This topic has been studied in the fields of vehicular networking (as the road-side unit placement problem) and wireless sensor network (as the optimal coverage problem). In spite of the wide body of work in both fields, we did not find a work that addresses placement of data aggregation/relay units constrained by distance to static clients (the sensor units). We provide a formulation in which a geographical constraint limits potential locations within a service range to the static clients. We also introduce a procedure to obtain a city-scale data volume characterization using real-world datasets from vehicular traces and a measurement campaign.

2.3 Data Collection Protocols

In the context of our work on data collection over dynamic topologies described in Chapter 5, we provide a review of the state-of-the-art data collection protocols in the fields of mobile ad hoc

and sensor networks, as well as protocols that use network coding. A brief background and some operational aspects of network coding are provided, and an overview on opportunistic routing protocols concludes the section.

2.3.1 Related Work on Protocols

Unicast routing protocols for mobile networks are broadly classified according to the timeliness and opportunity of route discovery [50]. Protocols are called pro-active if nodes actively share link-state information in order to keep an updated global or partial representation of the network. The second category are reactive protocols, in which case route discovery is triggered upon request. Other categories listed in [51] include hybrid, location-aware and multipath. Some of the most relevant source-initiated on-demand routing protocols are the Ad hoc On-Demand Distance Vector (AODV) [52], the Dynamic Source Routing (DSR) [53] and the Temporally Ordered Routing Algorithm (TORA) [54]. Some examples of table-driven protocols are the Destination-Sequenced Distance Vector (DSDV) [55], the Optimized Link State Routing (OLSR) [56] and the Wireless Routing Protocol (WRP) [57]. Concerning protocols oriented specifically for data collection in MANETs, we found little work. A survey on urban vehicular sensing platforms can be found in [58], in which some high-level architectures for pervasive sensing over VANETs are described, such as MobEyes [59]. In this case, nodes produce meta-data and opportunistically distribute it up to a maximum number of hops. The protocol COL [60] presents a more routing-oriented approach to data gathering in VANETs. COL allows any node to request data from its neighbours within a pre-defined range. Nodes build a local and temporal representation of the paths to any other node and validate it periodically by sharing that data with other neighbours. Back off-based Per-hop Forwarding (BPF) [61] is also a data gathering protocol for VANETs that uses wireless broadcast transmissions and geographical location. Packets carry the location of the sending node and of the destination, and receiving nodes, when attempting to seize the medium to broadcast the packet, compute a back-off time that is proportional to their distance to the destination.

Wireless Sensor Networks (WSNs) [62] are networks of wireless-enabled sensor nodes that monitor environmental data (e.g. temperature, humidity) over a target area (e.g. a forest or a building). The application motivating most WSNs protocols is information gathering. Due to the static nature of WSNs, route discovery is made very sparsely in time. A typical categorization of WSN data collection protocols includes cluster-based (or hierarchic), location-based (geographic), and flat routing. The operation of cluster-based protocols is based on the aggregation of nodes in clusters and election of cluster heads. Some examples are Low-Energy Adaptive Clustering Hierarchy (LEACH) [63], Power-Efficient Gathering in Sensor Information Systems (PEGASIS) [64] and Threshold-sensitive Energy-Efficient Network (TEEN) [65]. In geographic protocols, nodes are assumed to be location-aware. Relevant geographical protocols in WSNs are Minimum Energy Communication Network (MECN) [66], Geographic Adaptive Fidelity (GAF) [67] and Geographical and Energy-Aware Routing (GEAR) [68]. Flat protocols assign all nodes the same relevance.

Some examples are Sensor Protocols for Information via Negotiation (SPIN) [69], the Directed Diffusion [70], and the Collection Tree Protocol (CTP) [71].

From the presented classes of protocols, a WSN flat protocol can be a good candidate for application in M/VANET scenarios. Most M/VANETs tend to be organized in flat hierarchies, thus cluster-based protocols do not bring particular advantages. Geographical routing protocols are a powerful approach for MANETs, but face the problem that minimum-distance wireless communication may not be possible or be sub-optimal regarding delivery rates and energy consumption [72]. The principle of CTP is to set up a minimum-cost routing tree from the base station to every source node. The routing gradient is the number of expected transmissions, ETX, at link level. CTP assumes the existence of a link quality estimator to learn the single-hop ETX of a node to its neighbours. The ETX of a root node is 0, whereas the ETX of a node is the ETX of its parents plus the ETX of its link to its parent. To set up the minimum-cost tree, base stations advertise themselves periodically as tree roots. CTP advertisements, called beacon frames, carry the ETX field of the forwarder node. From all the advertisements a node receives from its neighbours, it chooses the neighbour with the smallest ETX as a parent, as it identifies the minimum-cost route to the base station.

There are some data collection protocols using network coding. An hybrid approach is proposed by SenseCode [73], which performs opportunistic coding on top of an existing routing structure, such as the minimum-cost routing tree created by CTP. Nodes can overhear packets and code them with their own, therefore propagating linear combinations of their own packets and overheard packets. This results in additional redundancy as, if links suddenly fail, some of the information from upstream nodes might still be recovered. Regarding performance, the SenseCode authors report an improvement over CTP performance, and state that systematic coding achieves a reliability similar to full-coding while consuming less resources. In [74], a network coding-based protocol for collecting energy consumption data from wireless-enabled energy meters to base stations is proposed. Packet forwarding is performed over a pre-computed shortest path routes. Measurements are taken every 15 minutes and nodes retransmit the same data periodically until a new measurement takes place; the scale of the simulation scenario (123 nodes) motivates this option. A performance evaluation is performed by comparing against reference protocols and varying the density of households and the reliability threshold. These two works constitute the main references to the work described in Chapter 5.

2.3.2 Opportunistic Forwarding Protocols and Strategies

A taxonomy of opportunistic protocols for vehicular routing is provided in [75]. Opportunistic protocols do not confine their operation to identifying and maintaining a single path between source and destination. In the Extremely Opportunistic Routing (ExOR) [76], once the source broadcasts its packet, nodes run a protocol to determine the subset of nodes that received it, and the node from this subset that is closer to the destination broadcasts the packet. The Opportunistic Routing in Ad Hoc Networks (OPRAH) [77] protocol uses a route request and route reply to identify the minimum hop count route from source to destination, as in AODV. Nodes can forward

a packet if their hop count to destination is inferior than to packets' current hop count. The Resilient Opportunistic Mesh Routing (ROMER) [78] protocol uses a credit-based system. Costs are assigned to links, and a credit is assigned to a packet at creation time and in excess of the minimum cost to reach the destination. A packet can travel through a paths as long as it as credit.

Given its natural match to the multicast application, network coding finds in opportunistic protocols and forwarding a natural partner. One of the first works to explore the conjunction of the two concepts in wireless mesh networks, for inter-session flows, was COPE [79]. COPE is not a routing protocol *per se*, but a network module that sits between the MAC and IP layer of each node and explores coding opportunities in packet flows. The MAC-independent Opportunistic Routing and Encoding (MORE) [80] protocol uses a similar concept, although oriented for intra-flow sessions, and merges it with the fundamental idea of ExOR. The source creates N coded packets that are linearly independent combinations of its N data blocks. Moreover, in each packet, a list of nodes that could participate in forwarding the packet is included. The Coding-aware Opportunistic Routing (CORE) [81] protocol extends MORE to optimize the forwarding and coding decisions in the presence of multiple flows. Each source broadcasts its coded packets, and the subset of receiving nodes for each broadcast is identified. From these, it is identified the node that can perform the coding operation and transmission that will be the most beneficial for all destinations.

2.3.3 Background on Network Coding

We now provide a brief review of the base concepts and seminal literature on network coding. Network coding was first suggested in the seminal paper by Ahlswede et al. [82]. The work addresses the particular scenario of a point-to-point communication network in which one or more sources multicast their data to multiple sinks thorough intermediate nodes. The authors prove, through an information-theorethic approach, that the minimum of the individual max-flow bounds from each source to the destination nodes can be achieved by employing coding at the intermediate nodes. In [83], it is shown that linear coding, in which intermediate nodes perform linear combinations of received packets before forwarding them, suffices to achieve the optimum in the multicast scenario. An algebraic approach to network coding and an algorithm to compute the solutions if the full network topology is known is proposed in [84]. The concept of random linear network coding (RLNC) is introduced and explored in [85]. The performance of RLNC has been deeply studied and is dependent on a variety of factors, especially in multicast scenarios [86].

One of the most fundamental examples of network coding is the multicast application in the butterfly network. In this network, source S intends to send packets p_1 and p_2 to destinations D_1 and D_2. S transmits packets p_1 and p_2 to nodes X and W respectively. Nodes X and W then broadcast the received packet, with node D_1 receiving p_1, node D_2 receiving p_2, and node Y receiving both packets. Whereas in a normal routing scheme turns would be necessary for Y to relay packetsp_1 and p_2 sequentially, using coding Y can perform a linear combination of both packets and send it immediately. In the simplest form, the linear combination may be a XOR. After Z has relayed the linear combination to the two destinations, both D_1 and D_2 can retrieve the missing packet by performing a XOR with the already received packet. In order to achieve

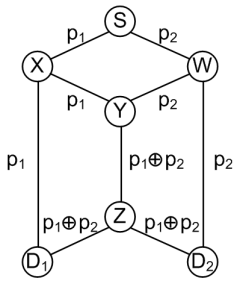


Figure 2.1: Butterfly network.

full packet delivery, the destination (or destinations) must receive N linearly independent (l.i.) combinations of the N original packets of the source (or sources). Each linearly independent combination available at the destination can also be refered to as *degree of freedom*, a term motivated by the algebraic/geometric nature of network coding.

In Random Linear Network Coding (RLNC), an extension of the previous approach, coding coefficients are associated to each packet when coding operations take place at a given node, chosen randomly from a Galois field (or finite field) $GF(2^q)$. The value of q defines the size of the Galois Field or, in other words, the range of available coefficients to use. Typical values of q in the literature are 1, 2, 4 or 8, leading respectively to 2, 4, 16 or 256 available coefficients. As a strategy to keep complexity and memory requirements low, native or coded packets are aggregated into groups of manageable size according to a specific criteria (typically interval of creation) [87]. Such groups are called *generations* and only packets belonging to the same generation may be mixed.

In a network coding protocol, the packet payload must be shared by the application data and the coefficient vector. The coefficient vector is a vector of identification elements for each data block contained in the payload. The triplets contain source node id, sequence number of data block and associated coefficient, (`src`, `seqnr`, `coeff`). There are two strategies to store the coefficient vector: on-demand or pre-assigned slots. In the first case, the triplets are stored on-demand in the packet payload, requiring all meta-data fields to be explicitly stored and added everytime a new data block is coded with the coded payload. In the pre-assigned case, a vector of coefficient placeholders, with the size of the number of the data blocks per generation, is reserved in the packet payload. The data blocks of each node are assigned a position in that vector; the coefficients of the data blocks present in that packet are stored at the respective locations. The selection of one of the two strategies must account for a number of impacting parameters: generation size, number of nodes, target coding density and Galois Field size [88].

2.3.4 Discussion

We reviewed the existing literature on routing protocols for M/VANETs, in terms of structured and opportunistic protocols. The bulk of data collection protocols, both from the wireless sensor networking and the mobile/vehicular ad hoc networking areas, relies on structured protocols, or

protocols that form well defined routes between source(s) and destination(s). Opportunistic protocols offer an alternative strategy by exploring a range of routes available at which moment. The pairing of opportunistic forwarding and network coding strategies has been proposed earlier, but a extensive design exploration over real-world mobility traces has, to best of our knowledge, not been reported. The work of Chapter 5 contributes substantially to increase the body of knowledge in this area.

2.4 Final Remarks

We presented the state of the art on the relevant fields for this thesis, namely on channel modelling between devices terminals, characterization of I2V links and service and placement of roadside static nodes, and on data collection protocols operating in mobile/vehicular ad hoc networks. These specific research topics stem from the network design aspects identified in the previous chapter, and underlie the contributions of the following chapters. The discussions at the end of each subsection motivate the contributions in the light of the existing literature.

Chapter 3

Propagation Models for D2D Channels and Impact of Erroneous Positioning

Ad hoc mesh networks can be set up to provide communication where infrastructure support is unavailable or not dependable. An important tool in the design of wireless protocols is the channel model, that provides an estimation-based relationship between device-to-device distance and received power. There are various types of propagation models with different degrees of accuracy and complexity and considering specific details about the environment, but generic models such as the log-distance path loss model are more useful as they are applicable to a wider range of scenarios. Futhermore, in the perspective of application designers, the communication range is the most relevant information for protocol development, for which the log-distance path loss model suffices. The estimation of the path loss model parameters involves undertaking measurement campaigns, in which the received signal strength readings are taken at known distances from the transmitter. The distance between the mobile devices is oftentimes obtained from the Global Positioning System, a feature that many portable devices nowadays are equipped with. This facilitates estimation of path loss model parameters in settings where exact distances are difficult to obtain or unavailable. However, we observed that the GPS positioning, particularly in lower-end devices, is prone to errors. Devices with errors in the range of centimeters are expected to cost thousands of Euros, whereas in consumer electronic products errors can reach tens of meters. The GPS positioning errors have a negative impact in distance estimation and in turn diminish the accuracy of the estimated propagation model.

In the context of a practical use-case – ad-hoc communication in a forested environment among smart phones using WiFi –, we conducted field measurements in a forest scenario to estimate model parameters of a device-to-device channel between mobile devices. This use-case is motivated by a feasibility study for a smart phone-based information system for firefighters [89]. The distance between terminals was obtained in two different ways: position estimates of the Global Positioning System, and centimeter-accurate distance measurements from a laser range meter. In addition to obtaining model parameters for propagation in forest environments, we used our dataset to address the problem of quantifying the impact of erroneous GPS position estimates in

the estimation of the path loss model. After analyzing the impact of GPS positioning errors on the estimation of the range between devices and consequently on the estimated propagation model, we derive guidelines for the design of future device-to-device path loss measurement campaigns, and propose a practical method to correct those errors based on Monte Carlo simulations. The conclusions we present are applicable to any outdoor scenario, provided that a good estimate of the standard variation of the GPS positioning error is available.

Our main contributions are as follows:

1. a model of the impact of GPS positioning errors on the ranging and path loss model estimation;
2. a method to improve path loss estimation using only GPS distances;
3. guidelines for designing measurement campaigns for path loss model estimation that reduce the impact of GPS errors.

The remainder of this chapter is organized as follows. In Section 3.1, we review channel models for the particular setting we address, a forest scenario. In Section 3.2, we describe the methodology for channel data collection and modeling, and present results using real-world measurements. In Section 3.3, we discuss the impact of GPS errors on the path loss model estimation. In Section 3.4, we provide guidelines for measurement data collection and a path loss model parameter retrieval methodology. In Section 3.5, a overview of the obtained results is presented.

This work was done in collaboration with Dr. Traian Emanuel Abrudan and has been published in the IEEE Transactions on Wireless Communications journal [6]. The text of this chapter was adapted from that article with minor modifications.

3.1 Related Work on Forest Channel Models

We provide a review on channel models for forest environments, the setting on which the experiments described in this chapter took place. A comprehensive survey of empirical path loss models for forested environments may be found in [90]. Typically such models are additive with respect to the free space path loss model [91]. The modified exponential decay (MED) model [92] is the basis for most empirical models concerning propagation in forests. MED uses the formula $A = \alpha f^{\beta} d^{\gamma}$, where the extra attenuation A is given in dB, and f identifies the signal frequency and d the tree depth. Parameters α, β and γ may be estimated from measured data. Given the wide variety of factors affecting propagation in forests (such as tree species, disposition of the trees, foliation), it is very difficult to find an universal set of values. The contribution of the various empirical models found in literature is to propose values that aim to be the most general possible, but these are strongly conditioned by the scenario in which data was obtained. Table 3.1 lists some of those models and corresponding parameter values. There is also a family of models based on the modified gradient model, such as the Maximum Attenuation (MA) [93] and Nonzero Gradient (NZG) [93] models. These models require additional parameters that are specific to the

Model	Note	α	β	γ
Weissberger [92]	$0 < d < 14$m	1.33	0.284	0.588
	$14 < d < 400$m	0.45	0.284	1
ITU [95]	$d < 400$m	0.2	0.3	0.6
COST 235 [96]	Not foliated	26.6	-0.2	0.5
	Foliated	15.6	-0.009	0.26
FITU [97]	Not foliated	0.37	0.18	0.59
	Foliated	0.39	0.39	0.25

Table 3.1: Parameters for various empirical models based on the MED model. f is set in GHz for the Weissberger model, and MHz for all others; d is in meters for all models.

measurement geometry and/or methodology. Both modified exponential decay and modified gradient models describe essentially propagation through canopies or at canopy-level. In [94], the authors analyze propagation at trunk level. Based on an extensive data set, the log-distance path loss model is shown to be the most accurate of existing models. An improved version of this model is proposed by incorporating scenario-specific parameters, namely tree density and trunk diameter. In summary, existing models for propagation in forests are tightly dependent on the particular conditions in which measurements take place. The application scope of our approach is not limited to forests, so we opted not to use such specialized models and use the more general log-distance path loss model.

Concerning measurement campaigns in forests, a large number of works, such as [98] and [99], focus on the propagation of GSM signals, in the 1900 MHz band. Propagation studies on the 2.4 GHz band are usually related to wireless sensor networks, which use lower transmission power and thus have limited range compared to typical WiFi or VANET communications [100]. None of these works clarifies which distance measuring method was used, nor account for distance measurement errors or for their impact on the accuracy of the estimated model. The only factor deemed relevant is the range of distances.

3.2 Measurement Collection and Parameter Estimation

We model the received signal power using a log-distance path loss model, whose formula in the logarithmic domain is given by

$$\rho_{[\mathrm{dBm}]}(d) = \rho_0 - 10\,\alpha \log\left(\frac{d}{d_0}\right) + X_\rho, \tag{3.1}$$

where ρ is the received signal strength (in dBm units) at an arbitrary distance d from the transmitter, and ρ_0 is the received signal strength at reference distance d_0 in the far field (typically 1 meter). The logarithmic signal strength measurements ρ are affected by normal fading, $X_\rho \sim \mathcal{N}(0, \sigma_\rho)$. In the following discussion, the term RSSI (Received Signal Strength Indicator) refers to the discrete readings of the received signal power ρ that are typically delivered by the drivers of wireless interface cards. The values of parameters α, ρ_0 and σ_ρ for each specific scenario are estimated

Figure 3.1: Experimental setting.

using distance-RSSI data pairs obtained in a measurement campaign, and then calculating the line that best fits the data.

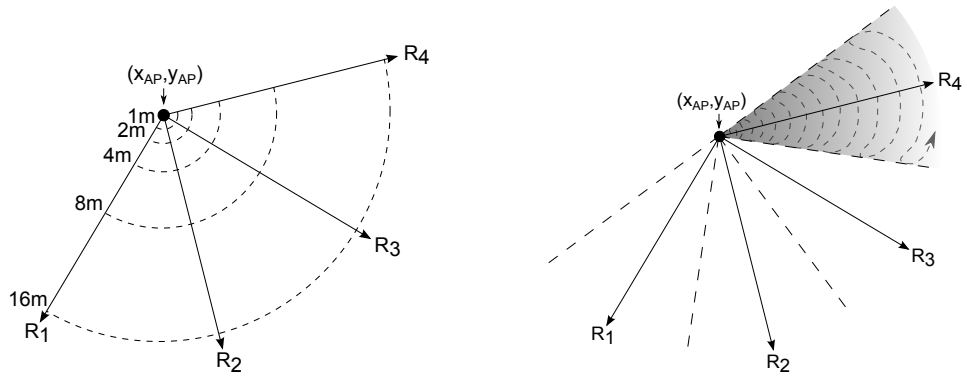
A laser-beam range meter could be used to determine the distances, but that would require line-of-sight between devices. A more practical solution is to record the coordinates obtained from the GPS receiver incorporated in the devices at different test distances, along with RSSI. However, since GPS positions contain errors, this solution requires a clearer understanding of how GPS-based distances affect the estimation of path loss model parameters, compared to the case when the actual distances are used.

In the remainder of this section, we explain the methodology we used to collect measurement sets of RSSI and distance data (Subsection 3.2.1) and present the results on path loss model estimation in the presence of GPS errors (Subsection 3.2.2).

3.2.1 Data Collection Methodology

We divided our measurement campaign into two phases. The first phase consisted of collecting pairs of RSSI and GPS measurements at known distances, to study the impact of GPS errors. In the second phase, we collected only RSSI and GPS samples. Measurements were taken in a forested area where the majority of trees were stone pines (*Pinus pinea*). The height of the devices was considerably lower than the bottom of the canopies (~1.5 m vs. ~6 m). The ground between transmitter and receiver was mostly covered with grass and small weeds. Most readings were taken in line-of-sight, or with a small number of tree trunks between receiver and transmitter. A small subset of the readings was taken with vegetation in-between, specifically bushes and smaller trees, most of them slightly taller than a human. Their density ranged from a single plant to a compact set of these. Figure 3.1 shows the various settings. The measurement equipment consisted of three standard, off-the-shelf, consumer electronics smart phones: one Samsung Galaxy Nexus S to act as access point (AP), which we call Device A, and two Samsung Galaxy Nexus to act as mobile receivers, referred to as devices B_1 and B_2. All models ran Android OS, and ran an application that stored GPS and WiFi RSSI values obtained from the Android APIs.

In the first phase of the measurements (henceforth called *Phase 1*), we placed device A at a fixed location, on top of a tripod of 1.5 meters, and set it in AP mode. This device periodically sent beacons announcing its presence to devices that aspired to join its network. It also recorded GPS measurements at a rate of 1 Hz. We then sequentially placed devices B_1 and B_2 together



(a) *Phase 1* of measurements – static case. The true distances were measured using a laser-based range meter. Solid lines show the radials along which readings were taken. Dashed lines mark the distances at which data was taken.

(b) *Phase 2* of the measurements – zig-zag walk. No true distances are available. Solid lines are the radials along which readings were taken. Dashed lines limit the data collection area for each radial. Gradient area marks data collection area for radial R4.

Figure 3.2: Spatial arrangement of measurements.

at several pre-defined distances from device A. For convenience, we call these distances "true distances". Their values were 0, 1, 2, 4, 8, 16, 32 and 64 meters, and we verified them *in loco* using a laser range meter. We chose these specific distances because they are equally spaced in logarithmic scale. At each true distance, we held devices B_1 and B_2 at a height of 1.5 meters, and each device recorded roughly 3 minutes of GPS and RSSI data, at an average rate of one GPS measurement every second and one RSSI measurement every two seconds. We repeated the same procedure for four radials roughly 45 degrees apart, as shown in Figure 3.2a. Device A recorded GPS measurements during the entire duration of each radial.

In the second phase of the measurements (henceforth called *Phase 2*), we again placed device A at a fixed location, set it in AP mode and activated it to record GPS measurements. We held devices B_1 and B_2 at a height of 1.5 meters next to the AP and initiated the recording application, collecting pairs of RSSI and GPS coordinates with the same frequencies as in *Phase 1*. We then carried them away from device A, following a zig-zag path that oscillated around each radial within a 45° angle approximately. We repeated the same procedure for the other three radials. Device A recorded constantly GPS coordinates during each radial measurement. Figure 3.2b shows the corresponding spatial arrangement.

3.2.2 Path Loss Model Parameter Estimation

From the data collected in *Phase 1* of the measurements, we obtain a set of GPS coordinates and RSSI values for each true distance. We compute the distances by pairing, via time-stamp, the coordinates recorded by device A and each device B ($B \in \{B_1,B_2\}$), and applying the Great Circle distance formula. This formula allows accurate computation of distances between points in a sphere whose positions are defined by decimal degrees. For convenience, we call distances

Phase	Model	ρ_0	α	σ_ρ
1st	True	-42.63	2.22	5.64
	GPS	-24.82	3.00	9.56
2nd	GPS	-42.86	2.13	7.34

Table 3.2: Model parameters derived from true and GPS distances using the Least Squares Estimator.

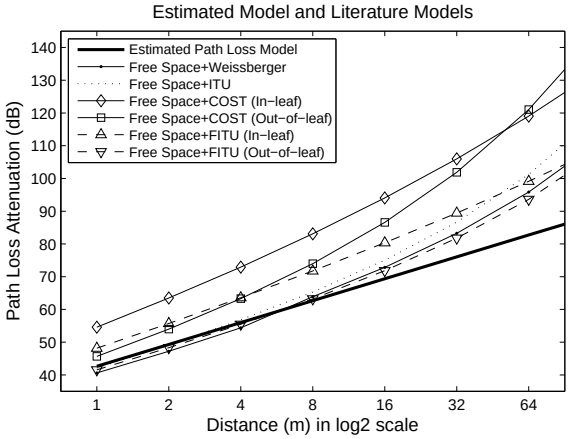


Figure 3.3: Measured RSSI data and "true" path loss model compared to literature models. See Table 3.1 for references.

calculated in this manner "GPS distances". During the analysis of the data, we found large clusters of consecutive, repeated GPS coordinates. We conclude that this range of commercial-grade devices may tend to fix on a set of coordinates if they do not detect significant movement for some time, in order to save energy. We substituted these clusters by a single measurement, to which we associated the median of the RSSI values of that cluster.

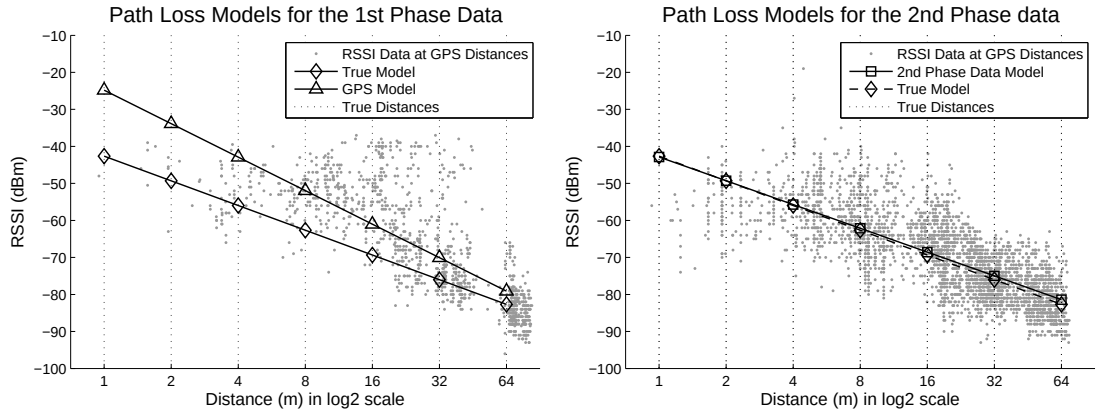
It is now possible to compute the parameters of the path loss model using two different data sets from *Phase 1*. In one case, we pair the RSSI measurements with the actual distances (obtained with a laser range meter). We call this the "true model". In the other case, we pair the RSSI values with the GPS distances, and we call this the "GPS model". Equations (3.2) and (3.3) describe both models as follows:

$$\rho(d) = \rho_0 - 10\,\alpha\,\log(d) + X_\rho, \tag{3.2}$$

$$\rho(d) = \tilde{\rho}_0 - 10\,\tilde{\alpha}\,\log(d_{\mathrm{GPS}}) + Y_\rho. \tag{3.3}$$

We perform regression over the two data sets using the Least Squares (LS) estimator which, for the true model, is also the maximum likelihood estimator. The parameters obtained from the measured data are shown in Table 3.2.

In Figure 3.3, our estimated model is shown against models proposed in literature. While for small distances the literature models are consistent with our data, they stray at larger distances. As



(a) *Phase 1* measurement data and models – static case

(b) *Phase 2* measurement data and models – zig-zag walk

Figure 3.4: Path loss models using the Least-Squares estimator. The true distances are marked by vertical dashed lines.

mentioned in Section II, most models reported in literature focus on propagation through canopies. For large distances, the extra-attenuation factor of the canopy models, which grows exponentially with distance, takes the overhand in the total attenuation. Given that our measurements were taken mostly in line-of-sight, with a very few occasional trunks or canopies between devices, our data does not show the effect of the additional attenuation, being best fitted by the log-distance path loss model. This explanation is supported also by the results in [94].

The GPS model exhibits overestimated parameter values α, ρ_0 and σ_ρ compared to the true model. Figure 3.4a shows both models derived from *Phase 1* measurement data. It also depicts the RSSI measurements at the distances provided by the GPS measurements, to give further insight on how their positions on the RSSI-distance plane condition the regression method. In Section 3.3, we provide a model for the erroneous distances, and analyse in more detail the impact of GPS errors on distance estimation.

As for the *Phase 2* measurement data (see Figure 3.4b), for which the exact distance was not recorded, we observe fairly different parameters when compared with the GPS model in the first phase (see Figure 3.4a). We expected them to be more similar, because both are computed using error prone GPS distances. However, unaccounted factors, such as user mobility, may explain this behavior. In the second phase of the measurements, more states of the channel fading are being captured due to the obstruction by trees, the user's body, and different device orientations.

3.3 Distance Estimation in Presence of GPS Errors

After discussing the previous motivating example, we now address the problem of estimating the distance between two GPS-equipped devices from error prone coordinates. Towards this end, we developed a model to describe the GPS positioning errors. This model will help explain the difference between the true path loss model and the GPS path loss model.

Following the characterization of GPS positioning errors and their sources in Section 2.1.2, we separate positioning errors into *systematic* and *non-systematic* errors, depending on the nature of the error source. Systematic errors affect all receivers within a certain area in similar manner, and hence are modelled as an identical position bias for all receivers. They are caused by atmosphere, quality of GDOP and ephemeris errors. Non-systematic errors are random in their nature, affecting each receiver and each measurement in a unique way. They are caused mainly by pseudo-range errors, multipath propagation, receiver noise, clock jitters and numerical errors. Based on this distinction, we now present a model for the error that affects distances computed from GPS coordinates. In this discussion, we assume local Euclidean coordinates, given that the scale of the distances we are using is small enough for the curvature of the Earth to be neglected. Our GPS position error model (see Figure 3.5) accounts for the distinction between the two types of errors mentioned earlier. Systematic positioning errors are modeled as a bias vector with respect to the actual position that is equal for all devices. Non-systematic positioning errors are modeled independently for each Euclidean coordinate as zero-mean circularly symmetric Gaussian random variables.

Given the exact Euclidean coordinates of the two terminals A and B, $(x_{\mathrm{A}}, y_{\mathrm{A}})$ and $(x_{\mathrm{B}}, y_{\mathrm{B}})$ respectively, the Euclidean coordinates corresponding to the measured GPS positions may be written as

$$(x_{\mathrm{A,GPS}}, y_{\mathrm{A,GPS}}) = (x_{\mathrm{A}}, y_{\mathrm{A}}) + (b_{x,\mathrm{A}}, b_{y,\mathrm{A}}) + (\varepsilon_{x,\mathrm{A}}, \varepsilon_{y,\mathrm{A}}),$$
$$(x_{\mathrm{B,GPS}}, y_{\mathrm{B,GPS}}) = (x_{\mathrm{B}}, y_{\mathrm{B}}) + (b_{x,\mathrm{B}}, b_{y,\mathrm{B}}) + (\varepsilon_{x,\mathrm{B}}, \varepsilon_{y,\mathrm{B}}),$$

where the errors $\varepsilon_{x,\mathrm{A}}$, $\varepsilon_{y,\mathrm{A}}$, $\varepsilon_{x,\mathrm{B}}$, $\varepsilon_{y,\mathrm{B}} \sim \mathcal{N}(0, \sigma_{\mathrm{GPS}})$ are assumed to be mutually independent, and σ_{GPS} is the standard deviation of the GPS positioning errors in each coordinate (x, y). The aggregated systematic errors along each axis are equal for both terminals, i.e., $b_{x,\mathrm{A}} = b_{x,\mathrm{B}}$ and $b_{y,\mathrm{A}} = b_{y,\mathrm{B}}$, effecting a translation of the terminal positions with no impact on the distance, as shown in Figure 3.5. For simplicity, we choose the local two-dimensional Euclidean system of coordinates with the origin centered at the exact location of device A, and with the abscissa-axis pointing in the direction of the device B, i.e., $(x_{\mathrm{A}}, y_{\mathrm{A}}) = (0,0)$ and $(x_{\mathrm{B}}, y_{\mathrm{B}}) = (d,0)$. Therefore, the expression of the GPS-based Euclidean distance between A and B reduces to

$$d_{\mathrm{GPS}} = \sqrt{(\varepsilon_{y_{\mathrm{A}}} + \varepsilon_{y_{\mathrm{B}}})^2 + (\varepsilon_{x_{\mathrm{A}}} + \varepsilon_{x_{\mathrm{B}}} + d)^2}. \tag{3.4}$$

Consequently, d_{GPS} follows a Rice distribution with the location parameter being the actual distance d, and the scale parameter $\sqrt{2}\sigma_{\mathrm{GPS}}$, i.e., $d_{\mathrm{GPS}} \sim \mathrm{Rice}(d, \sqrt{2}\sigma_{\mathrm{GPS}})$. The probability density function (p.d.f.) of the GPS distances given the actual distance is

$$\mathrm{p}(d_{\mathrm{GPS}}|d) = \frac{d}{2\sigma_{\mathrm{GPS}}^2} \exp\left(\frac{d_{\mathrm{GPS}}^2 + d^2}{4\sigma_{\mathrm{GPS}}^2}\right) I_0\left(\frac{d \cdot d_{\mathrm{GPS}}}{2\sigma_{\mathrm{GPS}}^2}\right), \tag{3.5}$$

where $I_0(\cdot)$ is the zero-order modified Bessel function of first kind.

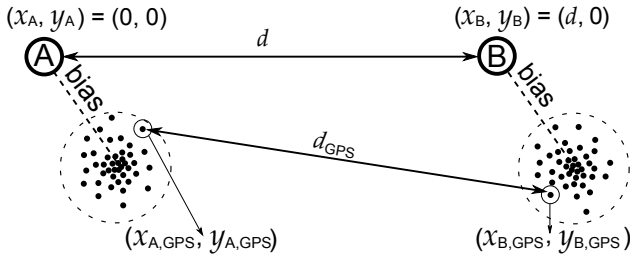


Figure 3.5: Characterization of the measured GPS distances.

Next, we provide a simple method to estimate the variance of the GPS error in each coordinate. The advantage of the proposed method is that the true coordinates of A and B are not required, the true distance being sufficient. Some receivers provide reliability information on the estimated position that can be used, for example, in a weighted LS estimation of the path loss model. Here, we assume that such information is unavailable, and we estimate an overall "average" reliability of the GPS position estimates. The second raw moment of the Rice distribution, μ_2', can be analytically related to the variance of the GPS coordinates as

$$\mu_2' \triangleq E\left[d_{\text{GPS}}^2\right] = 4\sigma_{\text{GPS}}^2 + d^2. \tag{3.6}$$

Empirically, the second raw moment of the GPS-based distances μ_2' can be obtained from the measured data by simply averaging the squared GPS distances. However, there are multiple true distances d_i. We compute the empirical second raw moment for each of the true distance, $\mu_2'(d_i)$, and use LS to estimate the overall variance of the GPS coordinates

$$\widehat{\sigma_{\text{GPS}}^2} = \frac{1}{4N}\sum_{i=1}^{N} \mu_2'(d_i) - d_i^2. \tag{3.7}$$

Finally, we compare the histogram of the measured GPS distances obtained in *Phase 1* with the Rice p.d.f. predicted by our model for each true distance. Figure 3.6 shows the normalized histograms of the GPS-based distances (dotted lines), and the Rice p.d.f. corresponding to that true distance $d \in \{1, 2, 4, 8, 16, 32, 64\}$ (solid lines). The true distance at which we took the respective GPS measurements is marked by a thick vertical solid line. Our model proves able to predict the major trends of the data. The overall standard deviation for GPS coordinates estimated using Equation (3.7) is $\widehat{\sigma_{\text{GPS}}} = 10.09$ meters. It may be noticed that for very small distances, i.e., $d \ll 3\sqrt{2}\sigma_{\text{GPS}} \approx 42$ meters[1], the GPS distances are overestimated by far, i.e, the mode of the Rice p.d.f. corresponds to a value much larger than the true distance. For larger distances, i.e., $d > 3\sqrt{2}\sigma_{\text{GPS}}$, the Rice p.d.f. is very close to a normal p.d.f. with the mode slightly larger than the true distance. In conclusion, we observe that GPS errors hamper significantly the distance estimation between two devices if the actual distance is smaller than $3\sqrt{2}\sigma_{\text{GPS}}$.

[1]The factor $\sqrt{2}$ appears due to the fact that the scale parameter of the standard Rice p.d.f. would be σ_{GPS}, whereas in our case, it is $\sqrt{2}\sigma_{\text{GPS}}$. The factor of three corresponds to the ratio d/σ_{GPS} for which a standard Rice p.d.f. can be approximated by a Gaussian p.d.f..

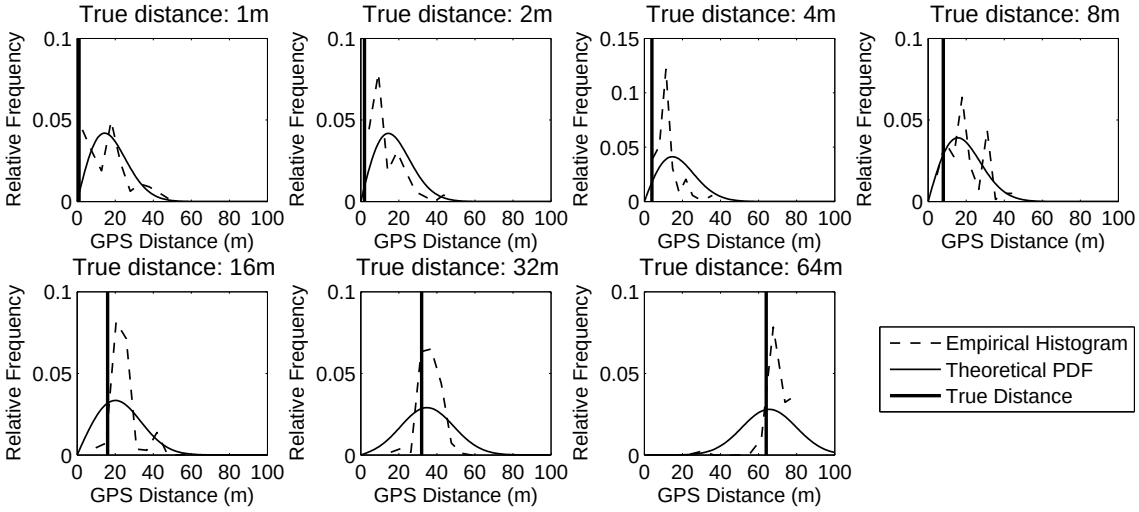


Figure 3.6: Normalized histograms of the GPS distances measured at each true distance.

3.4 Coping with Distance Errors in Path Loss Model Estimation

The error model for GPS distances introduced in the previous section can be used to improve path loss model parameter estimation. In Subsection 3.4.1, we detail the process by which GPS errors condition the estimation of the model parameters, and use this analysis to learn how can measurement campaigns be designed in order to mitigate the impact of such errors. Then, in Subsection 3.4.2, we provide a method based on Monte Carlo simulations to retrieve the true model parameters from measurements affected by GPS errors.

3.4.1 Guidelines for Selecting the Measurement Distances

In this section, we provide a selection policy concerning the distances at which measurements should be taken to mitigate the impact of GPS errors. We start by explaining the mechanism by which GPS errors impact parameter estimation using a linear regression method. In the previous section, we have seen that GPS distances tend to over-estimate small true distances. Consequently, for such distances, the RSSI values are paired with GPS distances larger than those at which we actually took them. This causes the distance-RSSI data pairs to be shifted to right side of the distance-RSSI plane. Due to this phenomenon, the estimated line departs from the true model, resulting in the GPS model. For large distances, the GPS distances are closer to the true distances. The corresponding distance-RSSI pairs are not shifted significantly, and therefore the perturbation they introduce in the path loss model estimation is little. Figure 3.7 illustrates the perturbed data, the mean and standard deviations for the RSSI-GPS distances pairs taken at each true distance, and the true and GPS models.

Based on this knowledge, a simple approach to improve the model estimation is to take more measurements at large distances, since those are less affected by GPS errors. Table 3.3 shows the

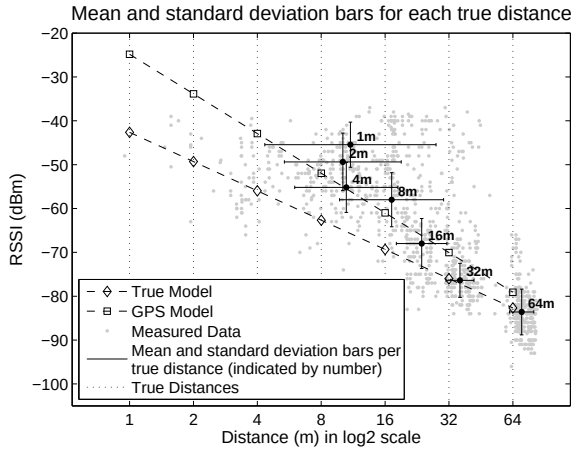


Figure 3.7: Error metrics for the data measured at each true distance.

Model	d [m]	ρ_0	α	σ_ρ	RMSE [dBm]
True	1...64	-42.63	2.22	5.64	-
Range	1...64	-24.82	3.00	9.30	11.70
	2...64	-25.20	3.06	9.32	11.03
	4...64	-26.27	3.05	9.31	10.19
	8...64	-24.63	3.17	9.37	11.00
	16...64	-30.85	2.85	9.27	7.18
	32...64	-38.44	2.44	9.36	2.55

Table 3.3: Model parameters and Root Mean Square Error (RMSE) if using the measurements taken at subsets of the true distances.

regression using only measurements taken at different subsets of the true distances d, where $d \in \{1,2,4,8,16,32,64\}$. We keep the maximum true distance constant, while varying the minimum true distance to be considered. We observe that, as we restrict the data set to the measurements taken at larger true distances, the estimated model tends to the true model. From our data sets, the largest distances available are 32 and 64 meters. Although the condition $d > 3\sqrt{2}\sigma_{\text{GPS}} \approx 42$ meters (see Section 3.3) is satisfied just approximately for the lower distance, the corresponding estimated GPS model is the closest to the true model.

This conclusion allows us to propose guidelines for the selection of the distances at which measurements should be taken. If the communication range allows, the measurements should be taken at true distances larger than $3\sqrt{2}\sigma_{\text{GPS}}$. Otherwise, the model derived from the GPS distances needs to be corrected. A correction method is provided in the next section.

3.4.2 Retrieving the True Model from Imprecise Distances

In many practical scenarios, true distance measurements are not available and GPS distances need to be used instead. This is the typical case when the GPS positioning that comes embedded in wireless devices and RSSI measurements from the wireless driver are used to derive a channel model. Note that direct estimation of the true distances from the GPS distances is not possible due

to the insufficient number of i.i.d. samples that are taken at a fixed distance, especially in mobile scenarios. Therefore, we propose a simulation-based method[2] to improve the path loss model estimation. The essence of this method lies in the fact that the RSSI samples used for estimating both the true and the GPS model are the same. This equality is shown in the following equations.

$$\rho(d) = \rho_0 - 10\,\alpha\,\log(d) \tag{3.8}$$

$$\rho(d) = \tilde{\rho}_0 - 10\,\tilde{\alpha}\,\log(d_{\mathrm{GPS}}) \tag{3.9}$$

Both equations hold in Least Squares sense, and therefore fading variance needs not to be included. As we assume that only GPS distances were recorded during the measurements, we can only compute the parameters of the GPS model $\tilde{\rho}_0$ and $\tilde{\alpha}$. Our method consists of defining a set of known reference distances, which are then perturbed according to the GPS error model of Section 3.3. The resulting simulated GPS distances are mapped into RSSI samples using the GPS model obtained from the measured data. Using the equality between (3.8) and (3.9), we pair the simulated RSSI values with the reference distances. By applying a regression method, we are able to retrieve corrected path loss parameters α and ρ_0 that are closer to the true model than if erroneous distance measurements had been used.

A crucial aspect of this procedure is the selection of the set of known reference distances. As seen in the previous subsection, the quality of the model parameters output by the regression method is very much affected by the distribution of the measured data with respect to the corresponding true distances. Therefore, we use Monte Carlo simulation to generate a set of simulated true distances $\hat{d}_i$ that resemble the actual true distances d_i as closely as possible. This ensures that the estimated model is correct. Our method consists of selecting the simulated true distances in such a way that the corresponding erroneous distances $\hat{d}_{\mathrm{GPS}_i}$ resemble the compound p.d.f. $p(d_{\mathrm{GPS}_i}|d_i)$ associated with the measured GPS distances d_{GPS_i}. We use the direct and inverse cumulative distribution function (c.d.f.) methods [101, Sec. 3.3] for this purpose. The direct c.d.f. method maps samples taken from a random variable X with some proposal distribution $p_X(x)$ into a uniformly distributed random variable $U \sim \mathscr{U}(0;1)$ using the c.d.f. of X as transformation function (by the uniform transformation theorem [101, Sec. 3.3.1]). Then, the inverse c.d.f. method maps U into a random variable Y that follows a target distribution $p_Y(y)$ using the inverse c.d.f. of Y as transformation function (by the inverse transformation theorem [101, Sec. 3.3.1]). In other words, $p_X(x)$ is used as a proposal distribution in order to sample from $p_Y(y)$. In our case, we use this procedure to sample from a proposal compound distribution of simulated erroneous distances $\hat{\delta}_{\mathrm{GPS}_i}$ in a way that the distribution of the transformed samples matches the compound p.d.f. $p(d_{\mathrm{GPS}_i}|d_i)$ corresponding to the measured GPS distances d_{GPS_i}. In the end, our procedure outputs a set of simulated true distances $\hat{d}_i$ paired with a set of simulated GPS distances $\hat{d}_{\mathrm{GPS}_i}$ whose distribution matches the one of the measured GPS distances. However, given that the mapping of the true distances d_i into GPS distances d_{GPS_i} is not bijective, there is no guarantee that the distribution

[2] Our attempt to derive a closed-form, or iterative estimator for the path loss model parameters (e.g. maximum likelihood) led to intractable calculations. Many of the Rice p.d.f parameters have complicated expression (e.g. moments are expressed in terms of Laguerre polynomials), and the expression of the density itself contains a Bessel function.

1.	Generate initial reference distances: $\hat{\delta}_i \sim \mathscr{U}(0, \max_i d_{\text{GPS}_i}), i = 1, \ldots, N$
2.	Simulate the GPS errors: $\hat{\delta}_{\text{GPS}_i} \sim \text{Rice}(\hat{\delta}_i, \sqrt{2}\sigma_{\text{GPS}})$
3.	Transform the resulting distances: $\hat{\delta}_{\text{GPS}_i} \longrightarrow v_i = P_D(\hat{\delta}_{\text{GPS}_i}) \sim \mathscr{U}(0,1)$
4.	Transform $v_i \longrightarrow \hat{d}_{\text{GPS}_i} = P_V^{-1}(v_i) \sim p(d_{\text{GPS}_i} \mid d_i)$
5.	Transform the uniform true distances to the simulated true distances: $\hat{d}_i = P_V^{-1}(P_D(\hat{\delta}_i))$
6.	Return: simulated true distances $\hat{d}_i$ and the simulated GPS distances $\hat{d}_{\text{GPS}_i}$

Table 3.4: The sampling procedure that generates samples from the p.d.f. of the measured GPS distances, and the corresponding simulated true distances pairs.

of the true distances will be reproduced exactly.

We now describe the procedure in more detail. We start by generating an initial set of reference distances uniformly distributed between zero and the maximum GPS distance recorded, i.e., $\hat{\delta}_i \sim \mathscr{U}(0, \max_i d_{\text{GPS}_i})$, $i = 1, \ldots, N$. We then perturb each reference distance $\hat{\delta}_i$ according to a Rice p.d.f. with the location parameter $\hat{\delta}_i$ and the scale parameter $\sqrt{2}\sigma_{\text{GPS}}$, i.e., $\hat{\delta}_{\text{GPS}_i} \sim \text{Rice}(\hat{\delta}_i, \sqrt{2}\sigma_{\text{GPS}})$. The standard deviation σ_{GPS} is the one computed from the empirical data in Section 3.3, using Equation (3.7). We then map the samples $\hat{\delta}_{\text{GPS}_i}$ of the proposal distribution to a standard uniform random variable $v_i \sim \mathscr{U}(0,1)$, using the c.d.f. of the samples themselves (by the uniform transformation theorem). We denote the corresponding direct c.d.f. transformation $\hat{\delta}_{\text{GPS}_i} \longrightarrow v_i$, by $P_D(\cdot)$. Then, we map the samples v_i to the target distribution of the measured GPS distances by using the inverse c.d.f. method, with the corresponding transformation $v_i \longrightarrow \hat{d}_{\text{GPS}_i}$, denoted by $P_V^{-1}(\cdot)$, where $P_V(\cdot)$ is the empirical c.d.f. of the measured GPS distances. We obtain the simulated true distances by applying the same direct and inverse transformations to the original uniformly distributed reference distances $\hat{\delta}_i$. The sampling procedure is summarized in Table 3.4.

This approach is able to approximate the true model parameters with good accuracy. Figure 3.8 shows the histograms of the path loss parameters α and ρ_0 estimated by 10000 independent runs of the Monte Carlo simulations. Based on the shape of the histogram, we assumed a normal distribution for the estimation error, and the confidence intervals were computed accordingly. The standard deviation for the estimated path loss exponent $\hat{\alpha}$ is around 0.1, whereas for the estimated reference RSSI $\widehat{\rho}_0$ is around 2dB. The true model parameters lie between one and two standard deviations with respect to the mean of the distribution of the estimated parameters. Table 3.5 presents the true model parameters and the mean values for the parameters estimated by the Monte Carlo simulations. Although the procedure does not guarantee retrieval of the exact true model parameters, it provides considerable improvement over the GPS model parameters.

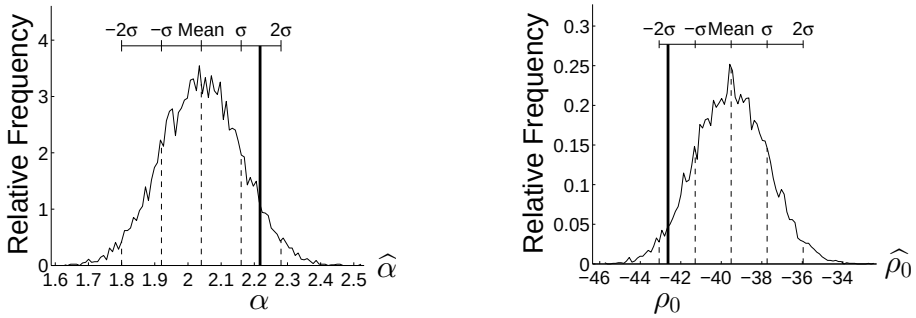


Figure 3.8: Histograms of occurrences for the $\hat{\alpha}$ and $\widehat{\rho}_0$ values in 10000 runs of the Monte Carlo simulation. Thick vertical line corresponds to true value.

Model	ρ_0	α
True	-42.63	2.22
Mean of MC runs	-39.51	2.04
GPS	-24.82	3.00

Table 3.5: The true model parameter values, the mean of the estimated parameters obtained from the Monte Carlo simulation, and the parameters derived from plain GPS distances.

3.5 Final Remarks

We addressed the topic of propagation modelling in device-to-device channels. We carried out field measurements to estimate the parameter values of the log-distance path loss model in a forest environment between two mobile devices. Distance between terminals was collected using the in-built low-end GPS receiver of the mobile devices, which can reach up to tens of meters. We proceeded to analyse the impact of positioning and distance errors on the estimation of a path loss model, using consumer-electronics (smart phones) communicating in the 2.4 GHz ISM band. We conclude that distances obtained from GPS measurements lead to the overestimation of the communication range, and show that this is caused by the distance errors that result from GPS inaccuracy. We modeled the impact of position errors on the distance estimation and provide guidelines for the selection of the distances at which measurements should be taken for path loss model parameter estimation. As a rule of thumb, if the communication range is sufficiently large, the measurements should be taken at distances larger than $3\sqrt{2}\sigma_{\text{GPS}}$. Otherwise, the path loss model derived from error prone GPS distances has to be corrected. A simulation-based method to correct the model estimated from erroneous distances was provided.

Chapter 4

I2V Service Characterization and Static Node Placement Driven by I2V Service

In a variety of distributed systems and applications involving equipments with sensors deployed over different locations, the data produced by the various equipments must be gathered at a back-office. If available, existing wireless backhauls, either fixed (i.e. infrastructural) or vehicular, can be explored for this purpose. We envision that dedicated road-side communication hubs are necessary to aggregate the data produced by nearby sensor units and forward it to the backhaul gateways (e.g. WiFi access points). In cases in which the sensor equipments are not equipped with wireless transceivers, communication hubs need to be deployed in a dedicated installation campaign. A placement process must occur prior to deployment, in which deployment sites for the hubs are identified under constraints related to the I2V service and other relevant factors while seeking for minimizing the number of hubs. This infrastructure planning task allows network designers to guarantee the collection service to the sensor equipments while using resources judiciously. A particular challenge of the placement process is the estimation of I2V data tranfers throughout all potential deployment locations to evaluate if the service requirements of the target application are meet. Given the unfeasibility of carrying out measurements at all potential locations, we can develop extrapolation models of I2V service based on the characterization of I2V links. This characterization can be obtained from a few dedicated measurement campaigns.

On a first stage, we carried out an experimental characterization of the wireless connectivity between the vehicular network and a road-side wireless client, in the city of Porto. Under the scope of the smart city platform PortoLivingLab, the city has been equipped with the vehicular network *BusNet* – a deployment of on-board units (OBUs) in the public bus fleet –, and a platform of small-footprint weather sensing stations named *UrbanSense*. We deployed a prototype (called Data Collection Unit – DCU) that encompassed a weather station as data producer and a communication hub atop a traffic light pole of a major street. Over the course of approximately one month, we characterized the duration and bandwidth of the wireless links between the communication hub and buses' OBUs using UDP streams. Daily connection time reached almost one hour on weekdays, resulting in a transferable daily volume in the order of 4+ Gigabytes. We also measured

the buses' speed and position, and found speed to have little correlation with throughput.

On a second stage, we addressed the placement process and estimation of I2V service. The selection of deployment for the communication hubs must take into account constraints relating to end-system equipments, logistic aspects (e.g. utility availability), and communication service by available backhauls (fixed or vehicular). We propose a decision support framework that produces a placement solution for communication hubs taking into account those restrictions and seeking minimization of the number of hubs. We formulate a minimization problem and propose a solution strategy that encompasses two steps: identify potential deployment locations, and find a minimal set of deployment locations. The second step maps the problem of finding the minimal set of hubs into an instance of the Set Cover problem, and a heuristic is proposed to solve it. To support the framework regarding I2V service, we developed a procedure to estimate and predict data transfers in I2V links at city-scale, using historical position traces of the vehicular nodes and a throughput-distance model obtained from our initial measurement campaign.

Finally, we provide an example application of the framework to a real-world deployment in the city of Porto, Portugal. The target end-system to be served are the full set of weather station units of UrbanSense, with 73 planned deployments and 22 actual deployments, to be installed at the municipality traffic lights and served by the BusNet on-board APs and the municipality-owned network of outdoor APs. Through parameter-space exploration using the dataset of 73 tentative locations, we find that by, sharing hubs over multiple equipments, the number of required hubs is 20% less than the number of serviceable equipments, for a range between equipment and hub of 300 meters. We also compare the quality of the placement output by the framework against the actual end-system deployment. For hubs serviced by the fixed backhaul, we observed that close to 60% of the deployed hubs are located up to 100 meters of a solution location, and 87.5% had good or sufficient WiFi service. For locations served by the vehicular backhaul, we conclude that our model of I2V data transfers estimates data volumes measured in the field to within a order and a half of magnitude and accurately ranks locations according to relative performance.

- A measurement campaign of I2V WiFi communications between nodes of a vehicular network and a prototype road-side unit deployment.
- Formulation of a minimization placement problem for road-side communication hubs that serve pre-placed equipments, depend on existing wireless backhauls and have logistic requirements, and a two-step solution strategy for this problem.
- A method to obtain a city-scale estimation of transferable data volumes in I2V connections, and discussion on the impact of incurring assumptions caused by limited availability or granularity of input datasets.
- Evaluation of quality of the framework placement against an actual deployment and of service estimates against field measurements, through application of the framework to a medium-sized European city scenario (Porto, Portugal).

The remainder of this chapter is organized as follows. In Section 4.1, the PortoLivingLab smart city platforms – UrbanSense and BusNet – are presented. Section 4.2 presents a characterization of I2V WiFi communications in an urban testbed, composed of a vehicular network and a prototype road-side unit deployment. In Section 4.3, we describe a decision support framework for communication hub placement, in terms of an optimization problem formulation and solution strategy for the DCU placement problem. In Section 4.4 we describe a procedure to obtain a city-wide characterization of I2V data volumes. Results of our framework for example input datasets and respective quality evaluation are presented in Section 4.5. Finally, in Section 4.6 we draw the main conclusions of our analysis.

The first part of this work – the initial measurement campaign at a prototype DCU described in Section 4.2 – has been published in the ACM MobiCom Workshop on Challenged Networks 2015 [7]. The remainder – the decision support framework for communication hub placement – has been submitted to the ACM Transactions on Sensor Networks [8]. The text of this chapter was adapted from those two articles with minor modifications.

4.1 Background on PortoLivingLab Platforms

In this section we provide background information about the smart city platform PortoLivingLab, a large-scale IoT-based multi-source sensing platform deployed in Porto, Portugal, and the two sub-infrastructures that motivated and supported the work presented in the chapter. The two infrastructures are the city-scale sensing platform UrbanSense [102], and the vehicular network BusNet [103], comprising of 600 vehicular nodes (400 of which buses) equipped with WiFi and DSRC. Relevant inputs from the partners of PortoLivingLab are also described.

UrbanSense aims to perform comprehensive monitoring of environmental parameters at selected locations of the city and subsequent storage and processing of sensor data at a backend server. The components of the platform are the following:

- *sensor units*: a collection of weather and air quality sensors;
- *communication hubs*: wireless-enabled devices that collect data from multiple sensor units and forward their data;
- *backend server*: central repository of the data collected by all sensor units, located in the cloud;
- *communication backbone*: infrastructure through which the data of all sensor units reaches the backend server.

The end-to-end communication architecture of UrbanSense is shown in Figure 4.1.

In the scope of the *Urbansense*platform, we often refer to co-located sensor units/hubs as Data collection units (DCUs). The sensor units incorporates ten sensors addressing weather metrics (wind vane and speed, rain gauge, thermometer, hygrometer), air quality (particles meter, O_3, NO_2) and life quality (luminance meter, sonometer, solar UV radiation), that produce a sample

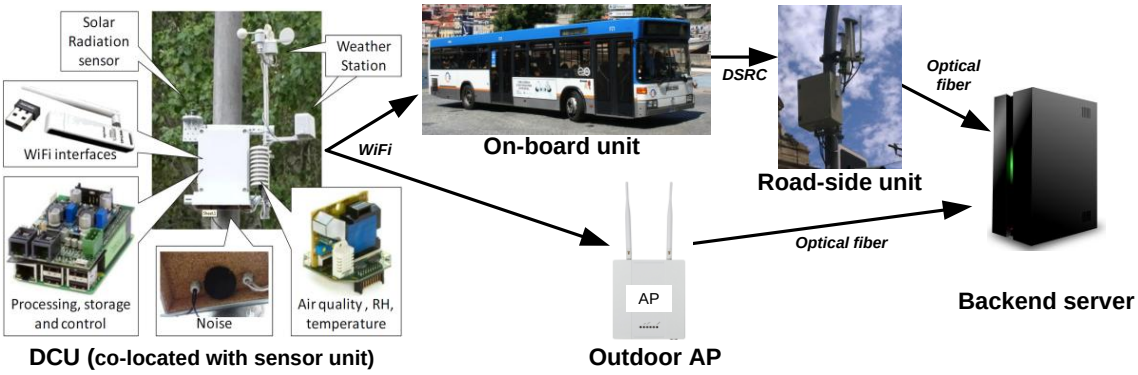


Figure 4.1: Architecture of sensing platform. Arrows indicate data collection flow.

every 15 minutes. The communication hubs in the Urbansense platform is equipped with a WLAN interface that can be configured to connect to infrastructural or vehicular backhauls. In the case of vehicular backhauls, the hub constantly scans the wireless medium for hotspot advertisements from buses. When a connection is established, data from the local database is transferred to the vehicular AP. The hubs also bookkeeps in a local database which data is in transit, and does not erase it until acknowledged by the backend server.

In operational terms, the first stage of deployment of the UrbanSense platform comprised 22 DCUs deployed throughout the city, progressively and starting from mid-2015. Since then, three DCUs had to be removed due to excessive wear and corrosion, specifically those deployed near the sea. Regarding produced data volume, sensor units are currently configured to sample all sensors every 15 minutes. A packet containing all sensor data of an interval of 15 minutes takes the size of 818 bytes. Over the course of one hour, this translates into 26.18 kbits, into 628.22 kbits over an day, and 4.39 Mbits over the course of week.

The vehicular network *BusNet*is a large-scale privately-operated deployment of on-board units (OBUs) in Porto's public bus fleet (400+ buses) and road-side units (RSUs) that interface OBUs and the Internet. OBUs are embedded computing devices equipped with a GPS receiver and modules for 3G, WiFi (IEEE 802.11 b/g/n) and DSRC (IEEE 802.11p) communication. The DSRC technology enables OBU-to-OBU and OBU-to-RSU communication within the vehicular network. The WiFi interface advertises a hotspot to which passengers and external clients can connect. Internet-access is provided via the 3G interface or the DSRC interface, directly to RSUs if possible or using delay-tolerant services. The GPS position of all OBUs is recorded every $\tau = 15$ seconds and later transmitted to a backend server of the service operator. RSUs are infrastructural routers deployed at strategic locations of the city, equipped with DSRC antennas and a wired connection to a fiber-optics ring. Sensor data from UrbanSense reaching any RSU is forwarded to the backend server. Additional detail about the vehicular technology deployed in Porto buses can be found in [104].

The municipality, also a partner of PortoLivingLab, provided access to utility facilities for equipment installation and power supply, specifically traffic lights, and granted use of the cost-free infrastructural access points *Porto Digital* that are connected to the Internet via a metropolitan

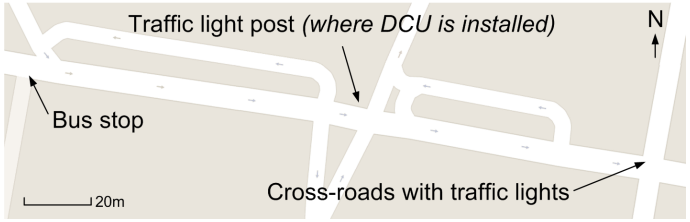


Figure 4.2: Spatial configuration of stopping opportunities at prototype DCU site.

Figure 4.3: West-bound, prototype DCU, and East-bound view respectively

fiber ring. Location of traffic light and free access points served by the city fiber ring were made available by the municipality. There are 290 traffic lights and 63 outdoor infrastructural APs available. Our dataset does not discriminate individual traffic light poles, but only indicates the geographical center of intersection the traffic lights are associated to.

4.2 Experimental I2V Characterization in an Urban Testbed

We performed an experimental characterization of the wireless connectivity between a prototype DCU and the nodes of the vehicular network BusNet. We describe next the experiment setting and methodology, present the obtained measurements in Section 4.2.2, and discuss some insights about site features obtained from additional processing in Section 4.2.3.

4.2.1 Experiment Description

Our experiments involved a DCU deployed at a street of our city where substantial bus traffic exists. The co-located sensor unit/communication hub was placed on a traffic light pole, at a height of approximately 4.5 meters, from where it draws electrical power. This pole is located roughly in the middle of a 600-meter long straight stretch of road, one-way and three lanes wide, with no curves or loss of line-of-sight for at least 300 meters in both directions. Variation in elevation amounts to 9 meters over that 600 meters stretch (less than 1° slope). Figures 4.3 shows the DCU installation and the street from the DCU point of view.

There are three bus routes going through the street where the DCU is located. This site features multiple stopping opportunities for buses: the traffic light where the DCU is placed, a bus stop at approximately 100 meters towards West and a cross-roads with traffic lights, at 125 meters in the other direction. At the cross-roads, three distinct bus routes pass on the perpendicular street, on

both ways. The only stopping opportunity with line-of-sight to the DCU for those buses is at the cross-roads' traffic lights. Night bus routes (0am-6am) only exist on the perpendicular street.

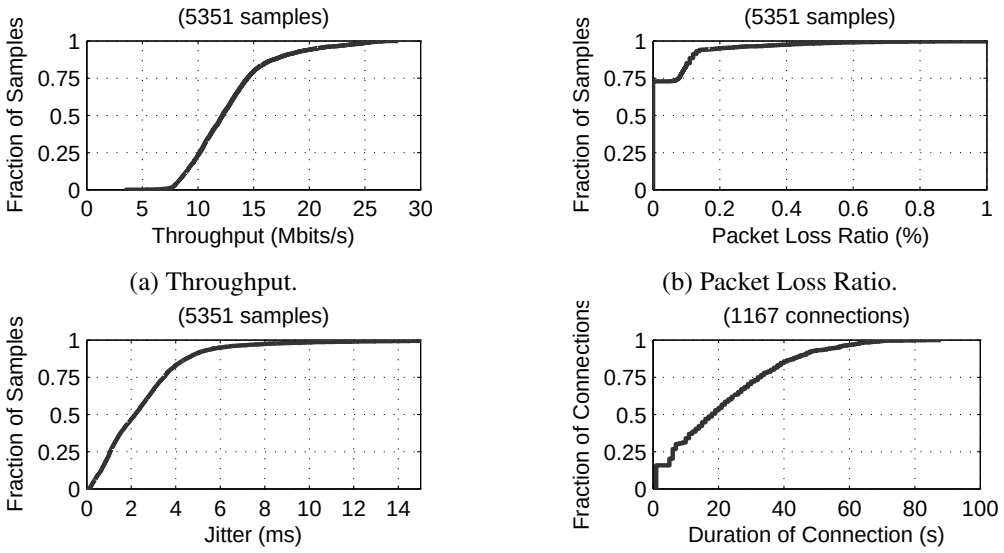
We now describe our measurement setup and its operation during the experiment. The measurements are managed by an application-level script at the DCU. In our setup, the IP of the DCU is assigned dynamically by the OBUs' access points. The script monitors the DHCP client (*dhclient*) service as it continuously waits for an IP assignment. The instant at which a new IP is detected is tagged as the beginning of the connection. To speed up the process of IP acquisition, the DHCP client parameters `initial-interval` and `backoff-cutoff` were set to to the minimum possible, 1 and 2 seconds respectively. The script then initiates a GPS query to the OBU and link quality measurements in a sequential fashion. The end of the connection is identified via timeout, when a query for new GPS data or a session of link quality measurements became unresponsive for 5 seconds. We do not have the information, in real or deferred time, to associate an OBU's MAC/IP address to specific routes.

We perform unidirectional (DCU to OBU) UDP link quality measurements. We use the tool *Iperf* [105] to generate load traffic (i.e., attempts to use the full bandwidth of the link) for a period of one second in each measurement. At the end of the Iperf measurement session, measured throughput, packet loss ratio and jitter for that second is reported. The Iperf client resides on the DCU and the Iperf server at the OBU. The GPS information of the bus is obtained via a query at application-level to the OBU, and contains time, longitude, latitude and speed of the bus. We pair a new set of GPS and Iperf measurements approximately every 2-3 seconds during the period the connection is alive. The MAC address of the associated AP is also stored.

This experiment took place during 25 days, starting in August 19 and ending in September 12 of 2014. During this period, our equipment was the sole user of the OBUs' APs.

4.2.2 Measurement Data Analysis

We describe the pre-processing and analysis performed to our measurements. The obtained raw samples were pre-processed to remove invalid or corrupt measurements. We observed that Iperf provides some anomalously large throughput values. Our measurements record throughput only at the application level and, even if the expectable drop in throughput due to IP/UDP overhead is disregarded, in no circumstance values higher than the physical layer's maximum nominal bit rate (55 Mbits/s for IEEE 802.11g) can be expected, which was the case. We filtered the measured samples depending on whether their value exceeded the physical layer's nominal bit rate. On the set of samples below 55 Mbits/s, we observed by histogram analysis that frequency of samples above 30 Mbits/s is null. This accounts for the expectable overhead and corroborates the validity of the selected samples. Regarding analysis of connections, we use the stored MAC addresses to identify the beginning and end instants, and quantify the duration. In case two consecutive sets of samples associated to the same MAC address are apart by more than 60 seconds, we consider those to be two independent connections. Total transferred data volume for a connection is estimated by multiplying the total connection time and the average throughput for that connection.



(a) Throughput.

(b) Packet Loss Ratio.

(c) Jitter.

(d) Connection duration.

Figure 4.4: Empirical CDFs over all measured samples and identified connections.

Additionally, estimated transferred data volumes per day were not consistent throughout the experiment period. The operation of our setup was interrupted on day 21 for the delay tolerant communication experiment, and due to malfunction on 22 and 23 of August, and from 5 to 9 of September. Other interruption periods, namely day 25 of August and 3 of September, may be associated to external problems such as disruptions in the WiFi service operation, in the public authority vehicles operation and/or in road traffic. We restricted our dataset to a period of stable operation, between August 26 and September 4 (indicated in Figure 4.5 by the grey length bar). For this period we obtained 5351 samples and identified 1743 connections to buses.

An analysis of the measured data after pre-processing is now presented. The empirical cumulative distribution of throughput is presented in Figure 4.4a. The mean throughput considering all samples is 12.82 Mbits/s and the respective standard deviation is 3.78 Mbits/s. The cumulative distribution of measured jitter is presented in Figure 4.4c. The mean is 2.627 ms and the standard deviation is 2.628ms. Figure 4.4b depicts the ECDF of the packet loss rate. Packet loss was zero in 72.82% of the samples. The ECDF of connection duration is shown in Figure 4.4d. The mean connection time is 21.38 seconds and the standard deviation is 17.39 seconds.

The transferred data volume and total connection time per day is shown in Figure 4.5. We highlight the apparent correlation between the daily data volume and the daily total contact time visible in the figure. The Pearson's correlation test provided a correlation coefficient of 0.965 between contact time and transferred data volume for all individual connections, and of 0.99 for the daily totals. The total daily data volume can reach close to 5 Gigabytes (see days 1 and 2 of September). Daily total data volumes on the weekend of August 30-31 are smaller than in week days. The average measured number of connections per day was 86.5 on weekend and 124.3 on week days, and the average connection time was 18.7 and 21.9 seconds respectively. A smaller number of connections on weekends is consistent with less scheduled passing-bys, whereas faster

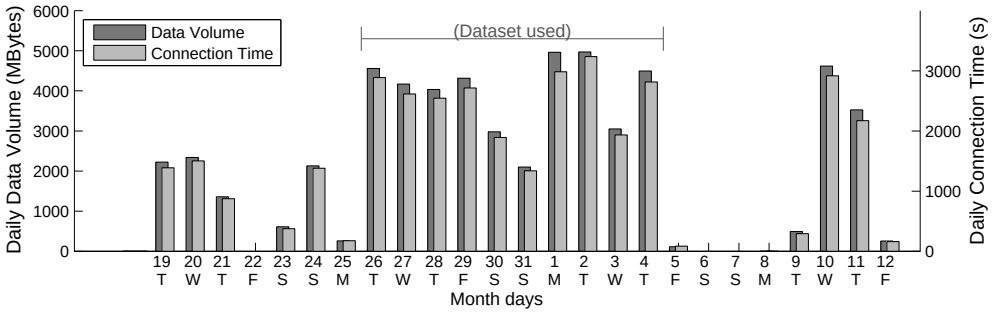


Figure 4.5: Average data volume transferred per day, over all days. Length bar indicates period of stable operation.

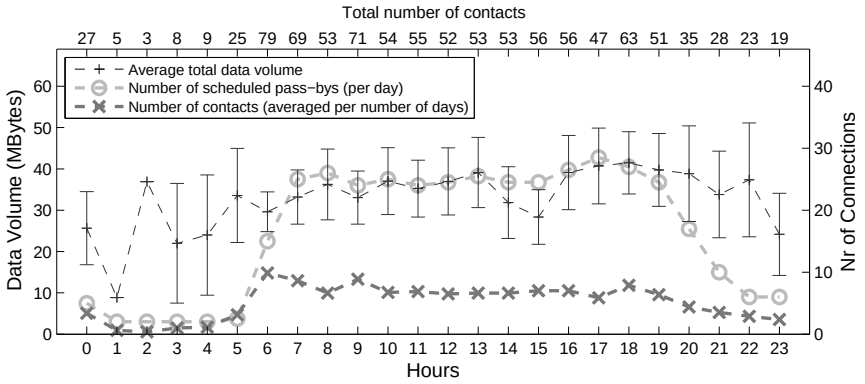
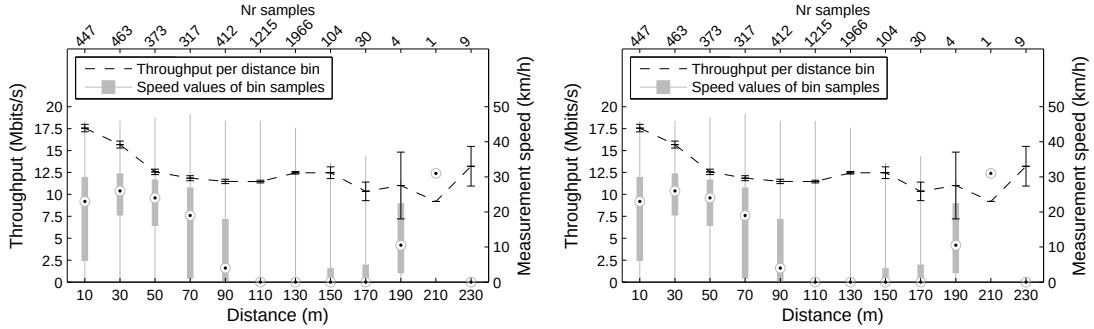


Figure 4.6: Average data volume, number of connections and number of scheduled bus passing-bys per hour for weekdays of stable period. Confidence intervals of 95%[1].

connections may be caused by less traffic and/or less passengers.

The hourly average data volume per hour during weekdays of the stable period, the number of contacts averaged per number of days and the number of scheduled bus passing-bys, are shown in Figure 4.6. An apparent trend is that data volumes transferred during the night period (11pm-7am) are lower than those of peak hours (7am-10am and 4pm-8pm). We hypothesize that this is related to the corresponding number of measured connections/scheduled passing-bys. Despite correlation values with hourly data volume not being very strong (Pearson coefficient of 0.562 for the number of connections and 0.663 for the number of scheduled passing-bys), the curves exhibit similar shapes. Regarding the relation between the number of measured connections and of scheduled passing-bys, recall that, as described in Section 4.2.1, there are two sets of routes passing within range of the DCU with different stopping opportunities. The number of scheduled passing-bys shown in Figure 4.6 include both sets. Pearson coefficient between both metrics is 0.829.

The impact of distance and speed on throughput is presented in Figures 4.7a and 4.7b. The measured throughput decays as the distance between terminals increases, which indicates lower signal-to-noise ratios caused by a decay in received power (as modeled by the line-of-sight path loss model). Samples at distances of 110 and 130 meters have usually velocity zero. These distances are coincidental with the nearby bus stop and cross-roads. Regarding speed, throughput showed little variation over the whole range of measured velocities, which is consistent with the



(a) C.I.s (of 95%) of throughput binned by distance, and boxplots of the speed associated.

(b) C.I.s (of 95%) of throughput binned by speed, and boxplots of the distances associated.

Figure 4.7: C.I.s of throughput binned by distance and speed.

existing results [29, 30, 32]. Note that higher speeds are typically recorded at closer distances.

The main takeaways are the following: (i) daily connection time reached almost one hour on weekdays, resulting in a transferable daily volume in the order of 4+ Gigabytes; and (ii) the buses' speed was found to have little correlation with throughput, which is in line with the conclusions in the literature.

4.2.3 Discussion on Site Selection

With the available dataset, we looked further into understanding the characteristics of opportunistic connectivity that impact the most the total transferable data volume at a potential DCU deployment site. Colloquially, we ask if larger hourly data volume transfers are achieved at sites where few but long connections occur (near a bus stop), or in sites where connections are numerous but short (a major road without traffic lights or bus stops).

From the results presented earlier, we observed that transferred data volumes were highly correlated, according to the respective large Pearson coefficient, with connection duration, per instance and per day. This seems to indicate that locations with the largest daily total connection times should be favoured. Also, throughput measurements present negligible influence of speed and exhibited a small variance over the whole range of velocities recorded. We conclude that connections with buses moving at urban speeds support similar throughput as with stopped buses, indicating that locations with few stopping opportunities may sustain high daily data volumes.

Additional insight was obtained from analysis over hour-long segments. We selected the hourly number of contacts (`nrconn`), the hourly average throughput (`thr_avg`) and the hourly average of connection duration (`dur_avg`) as potential features to provide indication about the hourly total transferred data volume (`tdv`). We computed `nrconn`, `thr_avg` and `dur_avg` from the data of the stable operation period, totalling 223 observations, and calculated their correlation with `tdv`; results are shown in Table 4.1. The number of connections of a site, `nrconn`, exhibited the highest correlation to `tdv`, indicating that locations with a large number of connections

[1] C.I.s for hours 1 and 2 omitted due to lack of statistical significance.

Method	thr_avg	dur_avg	nrconn
Pearson	0.288	0.577	0.828
Kendall	0.205	0.463	0.683
Spearman	0.306	0.635	0.846

Table 4.1: Correlation of tdv and selected features.

Crit.	thr_avg	dur_avg	nrconn
AIC	0.305	1.527e-49	1.437e-88

Table 4.2: p-values for selected predictors. Inclusion/exclusion criteria was Akaike Information Criterion.

to buses are preferential. Multiple regression analysis further supported this conclusion. We input tdv as the variable to be modelled and, as predictors, nrconn, thr_avg and dur_avg. We used stepwise regression and constrained it to finding a linear model without predictor interactions within 20 rounds. The p-values at the last round are presented in Table 4.2. The best predictor is the hourly number of connections, nrconn. Again, a large number of connections seems to be preferential over a large average connection duration. Incidentally, this particular result ended up not being used in the work described in the remainder of this chapter, but left an open-ended hypothesis to be validated in future measurement campaigns.

4.3 Decision Support Framework for Communication Hub Placement

We now present our decision support framework for hub placement. An optimization formulation that incorporates real-world limitations as constraints is presented next. A solution strategy that outputs a placement solution for communication hubs is described in Section 4.3.2.

4.3.1 Problem Statement

We formulate the optimization problem that sits at the core of our decision support framework, the *Min-Hub Problem* . A mathematical formulation of Min-Hub Problem is shown in Problem 1.

We refer to the system/platform that the hubs must serve as *end-system*, and composing nodes that produce data as *sensor units* $\underline{\mathbf{S}}$. The set of communication hubs is referred to as $\underline{\mathbf{P}}$. A potential hub location associated to a particular sensor unit s_i is referred to as p_i. A single hub may serve multiple sensor units. The location of a single unit for an arbitrary class (locations or units) is indicated by x_i with a superscript indicating the respective class (e.g. for sensor unit, x_i^s). Exception applies to potential placement locations, referred simply as x_i.

The cost function we use is the minimization of the number of necessary communication hubs while serving the maximum possible number of sensor units. The Min-Hub Problem encompasses a number of constraints and inputs that are categorized into four classes, that we introduce next.

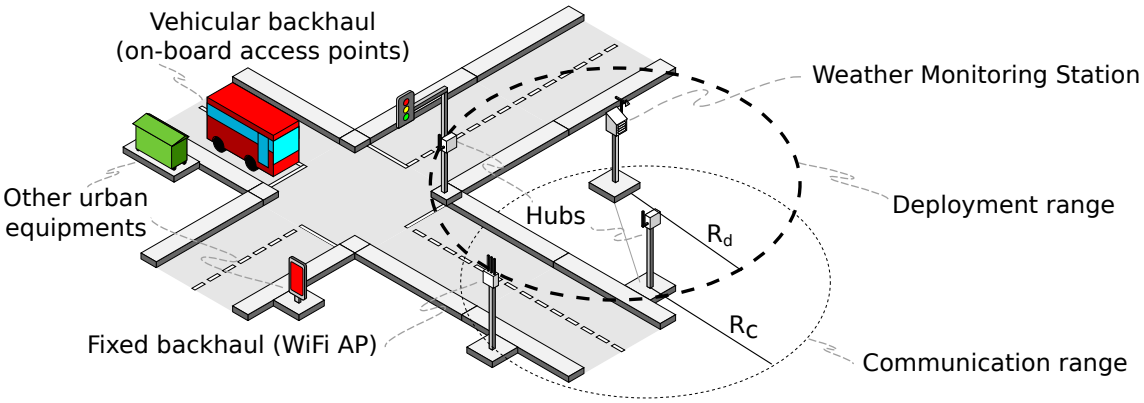


Figure 4.8: Application scenario for communication hubs and depiction of some technical aspects constraining hub placement (e.g. deployment range to sensor units, communication range to gateways).

The mathematical formulation of the listed constraints is found in Problem 1, and Figure 4.8 provides graphical support of how some of the constraints relate physically to each other. Throughout this discussion, superscripts f and v identify the fixed and the vehicular backhauls respectively.

- **End-system:** we define that sensor units produce a fixed amount $v_{\min}$ of data over an arbitrary period of time (constraint **c1**), and the number of communication hubs must be equal or inferior to the number of sensor units (constraint **c2**). The location of the sensor units ($\underline{\mathbf{S}}$) is used in constraint **c3**.

- **Logistic:** we incorporate the possibility that hubs require a power supply (and/or other utilities) for operation and that, given that hubs may have to be installed in outdoor spaces, that their deployment is constrained to authorized locations. Thus, hubs must be deployed at logistic locations, where: (a) necessary utilities are available; and (b) permission is granted. In addition, logistic locations should be up to a maximum distance to sensor units that a communications link can be established. Formally, we define an input dataset of logistic locations $\underline{\mathbf{U}}$, and eligible locations must be within a maximum deployment range r_{d} to each sensor unit $(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})$ (constraint **c3**). The data transfers attainable by a hub deployed at one of these logistic locations must be sufficient to serve all associated sensor units (constraint **c4**).

- **Communication:** we assume the existence of both fixed and mobile backhauls, i.e. composed of static and mobile APs respectively. Due to their different nature (even among backhauls of the same type), it can be necessary to define backhaul-specific constraints and inputs. The fixed backhaul encompasses a set of access point locations $\underline{\mathbf{A}}$, and in the current formulation, an isotropic unit disk of radius r_c is used to model service availability (more accurate models can be used if available). Thus, we set the constraint that a tentative deployment location x_i must be within wireless communication range r_c of an access point (constraint **c5**). The vehicular backhaul requires data volumes transfered in I2V connections at the potential location x_i to support all served sensor units (constraint **c6**). A model of I2V

Problem 1: Min-Hub Problem

$$\text{minimize} \quad \sum_{i}^{|\underline{\mathbf{S}}|} c_i^f \cdot q_i^f + c_i^v \cdot q_i^v$$

$$\text{subject to:} \quad v_i^s \geq v_{min},\ \forall\, i \qquad \textbf{(c1)}$$

$$|\underline{\mathbf{P}}| \leq |\underline{\mathbf{S}}| \qquad \textbf{(c2)}$$

$$x_i \in \underline{\mathbf{U}}, d(x_i^s, x_i) < r_d\ \forall\, i \qquad \textbf{(c3)}$$

$$v_i \geq v_{min} \cdot |cover(x_i, x_i^s)|,\ \forall\, i \qquad \textbf{(c4)}$$

$$q_i^f = \begin{cases} 1, \text{if } \{\, x^a \in \underline{\mathbf{A}} : d(x_i, x^a) < r_c \,\} \neq \emptyset \\ 0, \text{otherwise} \end{cases} \qquad \textbf{(c5)}$$

$$q_i^v = \begin{cases} 1, \text{if } M_v(x_i) \geq v_{\min} \cdot |cover(x_i, x_i^s)| \\ 0, \text{otherwise} \end{cases} \qquad \textbf{(c6)}$$

$$c_i^f = f^f(\text{user-defined criteria}), c_i^v = f^v(\text{user-defined criteria}) \qquad \textbf{(c7)}$$

Notes: $d(x_i, x_j)$ indicates Euclidean distance between elements x_i and x_j; $cover(x_i) = \{s^i : d(x_i, \underline{\mathbf{x}}_i^s) < r_d\}$, i.e., function $cover(x_i, x_i^s)$ outputs the set of sensing units s_i within deployment range of a potential deployment location.

data transfers $\underline{\mathbf{M}}_{\mathrm{V}}$ is used to estimate service by the vehicular backhaul; its generation is described in Section 4.4.

- **User-defined:** the user-defined costs shown in constraint **c7** include any relevant constraints that are not of technical nature concerning the backhauls (although similar costs may be defined for logistic aspects). These may be service fees of a commercial backhaul and existence of particular institutional partnerships. The user-defined costs c are represented as outputs of functions $f(\text{user-defined criteria})$ and may take arbitrary values.

In the light of formal definitions just presented, we refine the definition of our cost function that is formalized in Problem 1. We seek to minimize the sum, over the set of sensor units to be served, of a viability factor q and a cost factor c per communication backhaul for each potential location x_i. The viability factor q summarizes whether a potential deployment location meets the technical constraints (end-system, deployment and communication), and thus can only take binary values. The user-defined cost factor c incorporates non-technical constraints.

4.3.2 Solving Strategy

The solution strategy of our framework has two stages:

1. Produce a set of locations where hub deployment is possible from the input datasets. If possible, compute also a ranking of such tentative locations.

2. Find a subset from the set of potential deployment locations that optimizes the cost function.

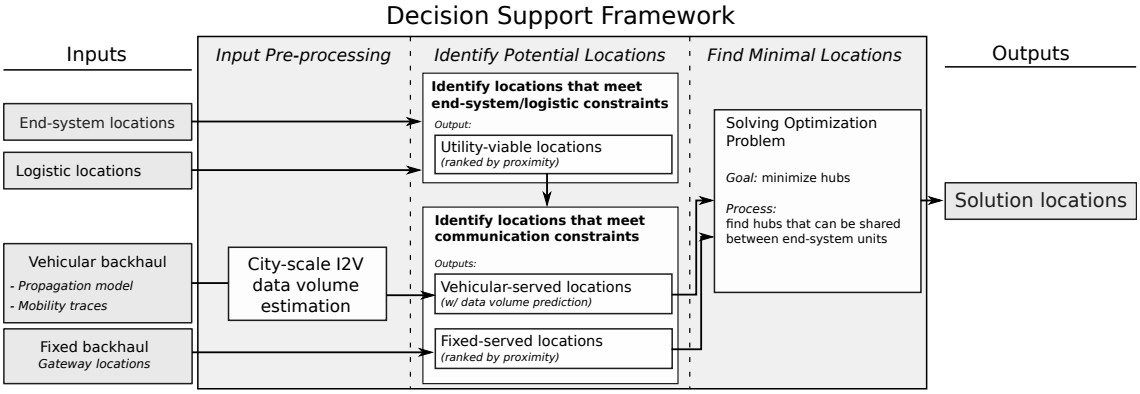


Figure 4.9: Decision support system for site selection.

Figure 4.9 describes the workflow of the decision support framework, system as it will be discussed throughout this section.

4.3.2.1 Identifying Potential Locations

In this stage, we seek to identify the logistic locations that are viable for deployment – that are within deployment range r_d and can be served by at least one of the backhauls.

We review the necessary inputs for the first stage of the solving procedure: (i) sensor locations $\underline{\mathbf{S}}$; (ii) logistic locations $\underline{\mathbf{U}}$; (iii) locations of fixed access points $\underline{\mathbf{A}}$; (iv) the map of I2V data transfers $\underline{\mathbf{M}}_{\mathrm{V}}$. We first identify the logistic locations $\underline{\mathbf{U}}$ that are within r_d meters of any sensor unit $\underline{\mathbf{S}}$. The process is done by computing the Euclidean distance and thresholding, and the output set is referred to as the logistic-viable locations $\underline{\mathbf{U}}^{\mathrm{s}}$. With this step, constraint **(c3)** (hub must be co-located with logistic location) is observed. We further narrow down viable locations independently for each backhaul. Regarding the fixed backhaul, given the use of an isotropic unit-disk to model service availability (as discussed in the previous section), we apply the same approach to identify logistic locations $\underline{\mathbf{U}}$ within communication range r_c of the access points $\underline{\mathbf{A}}$. By intersecting the resulting logistic locations with $\underline{\mathbf{U}}^{\mathrm{s}}$, the logistic-viable locations serviceable by fixed backhaul $\underline{\mathbf{U}}^{\mathrm{f}}$ can be obtained. Constraint **(c5)** is meet if there is at least one AP in range of a $\underline{\mathbf{U}}^{\mathrm{s}}$ for a given s_i. As for the vehicular backhaul, we must identify the logistic-viable locations $\underline{\mathbf{U}}^{\mathrm{s}}$ that support sufficient I2V data transfers to serve covered sensor units, thus meeting constraint **(c6)**. The map of I2V data transfers $\underline{\mathbf{M}}_{\mathrm{V}}$, whose generation is detailed in Section 4.4, is used for this purpose, producing set $\underline{\mathbf{U}}^{\mathrm{v}}$ (logistic-viable locations serviceable by the vehicular backhaul).

Formally, the procedure for can be summarized as follows.

$$\textbf{Step 1:} \quad \underline{\mathbf{U}}^{\mathrm{s}} = \{u_i : \{s_j \in \underline{\mathbf{S}} : d(x_j^s, u_i) < r_d, \forall j\} \neq \emptyset\} \tag{4.1}$$

$$\textbf{Step 2:} \quad \underline{\mathbf{U}}^{\mathrm{f}} = \underline{\mathbf{U}}^{\mathrm{s}} \wedge \{u_i : \{a_j \in \underline{\mathbf{A}} : d(x_j^a, u_i) < r_c, \forall j\} \neq \emptyset\} \tag{4.2}$$

$$\textbf{Step 3:} \quad \underline{\mathbf{U}}^{\mathrm{v}} = \underline{\mathbf{U}}^{\mathrm{s}} \wedge \{u_i : M_{\mathrm{v}}(u_i) > v_{\min} \cdot |\mathrm{cover}(u_i, \underline{\mathbf{u}}_i^s)|\} \tag{4.3}$$

where $cover(u_i, \underline{\mathbf{u}}_i^s)$ is the set of sensor units serviceable by logistic location u_i.

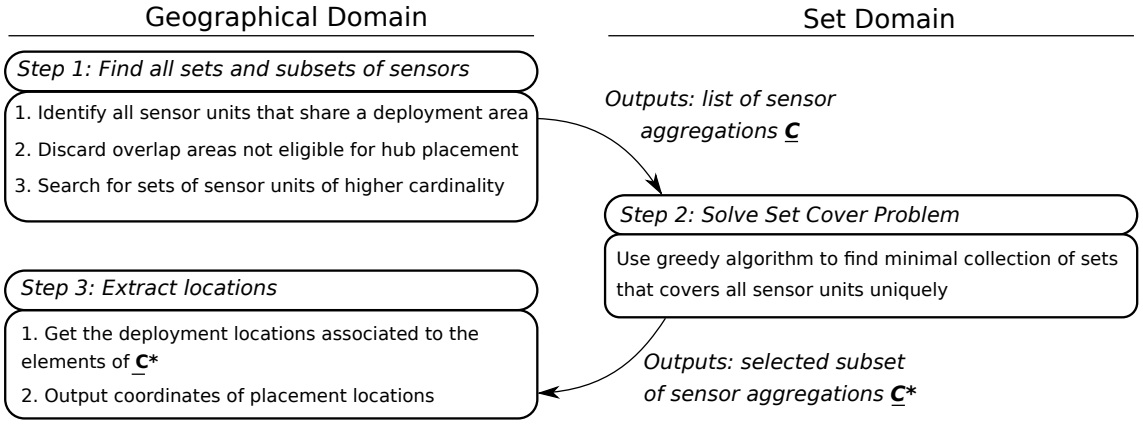


Figure 4.10: Solution workflow for Min-Hub Problem.

In the context of our mathematical formulation, we wish to compute the viability factor q for each element of set of logistic-viable locations. For this purpose, for each $i = 1...|\underline{\mathbf{U}}^{\mathrm{s}}|$, a binary value 'one' is assigned to q^v and/or q^f if the logistic location can be served by the vehicular and/or fixed backhaul respectively, or 'zero' otherwise.

Ultimately, the final output of this stage is the set of logistic-viable locations that can be served by one or either backhaul.For some input parameter values, it may occur that not all sensor units are serviceable by a hub – i.e., constraint **(c1)** is not observed for all sensor units. In such circumstances, we discard the sensor units that are not assigned a hub location. In the following step, constraint **(c2)** is thus applied only to sensor units that can be served.

4.3.2.2 Minimum Cost Problem

After logistic locations have been classified as viable under one or more communication backhauls, we address the task of minimizing the number of communication hubs. Our solution procedure for this stage is composed of three distinct steps, and the associated workflow is shown in Figure 4.10. The procedure is applied independently to the two backhauls.

We note that the deployment range r_d defines a circle around each sensor unit where a hub must be deployed; we refer that that area as the *deployment region* of a sensor unit. We tackle the problem by leveraging the insight that hubs can only be shared if the deployment regions of two or more sensor units have a common area. We start by applying a clustering algorithm to identify all deployment regions that overlap and aggregate the respective sensor units into sets. Then, we evaluate if the shared deployment regions contain fixed or vehicular backhaul-served logistic locations and remove those that do not respect this condition. After this step, we are faced with an instance of the Set Cover problem, an optimization problem known to be NP-hard [106]. A reduction proof of this identity can be found in Appendix A. We compute a solution using an greedy heuristic, and then search the selected sets of shared deployment regions for the logistic location with the largest data volume transfer to find a hub placement location. In Figure 4.10, the steps in the "Geographical Domain" or the "Set Domain" columns are those that work over

geographical information (distances or coordinates) or over sets of sensor unit associations (sets) respectively.

In the first step, we identify the sets of sensor units that have common areas in their deployments regions. Two sensor units share deployment areas if the Euclidean distance between both is less than $2r_{\mathrm{d}}$ and if served by the same communication backhaul. We identify all sets of sensor units that respect this rule. The respective shared deployment areas are filtered to evaluate their viability for hub placement: if the area area does not have any logistic location, it is not eligible and thus excluded. We refer to the subset of sensor units that have valid shared deployment areas as $\underline{\mathbf{S}}^{\mathrm{o}} \subseteq \underline{\mathbf{S}}$. For the complementary subset, i.e., the set of sensor units that do not have a shared area with another sensor unit that are eligible for hub placement, the closest logistic-viable location is selected. The output of this step is a collection $\underline{\mathbf{C}}$ of all sets C of sensor units that have shared overlap areas:

$$\underline{\mathbf{C}} = \left\{ C : C = \{s_{\mathrm{i}}^{\mathrm{o}}, ..., s_{\mathrm{k}}^{\mathrm{o}} : s^{\mathrm{o}} \in \underline{\mathbf{S}}^{o},\ i \neq ... \neq k,\ d_{\mathrm{i}} \cap ... \cap d_{\mathrm{k}} \neq \emptyset \},\ |\{i, ..., k\}| \geq 2 \right\} \tag{4.4}$$

where d_{i} is the deployment region of sensor unit s_i^{o}, $d = \{(x,y) : d((x^{s_{\mathrm{i}}^{\mathrm{o}}}, y^{s_{\mathrm{i}}^{\mathrm{o}}}), (x,y)) \leq r_{\mathrm{d}}\}$. Each set in $\underline{\mathbf{C}}$ is assigned a cost c_i equal to the cost of the associated backhaul: c^f or c^v. Note that each set of C is binded to a specific hub that is to be shared by all sensor units in that set. We refer to the set of shareable hubs as $\underline{\mathbf{P}}^{\mathrm{o}}$.

The second step of our solution hinges on the realization that our current problem is an instance of the Set Cover problem [39]. The objective of this optimization problem is, for a set of elements $e \in \mathscr{E}$ and a collection of sets $f \in \mathscr{F}$, seek a collection $\mathscr{G} \subseteq \mathscr{F}$ of disjoint sets that contains all elements e. Our current problem maps into the Set Cover problem with little differences. The universe of elements $\mathscr{E}$ being the set of sensor units that have shared deployment areas $\underline{\mathbf{S}}^{\mathrm{o}}$ and the collection $\mathscr{F}$ being the collection of associations of sensor units that can share a hub $\underline{\mathbf{C}}$. To minimize the number of necessary hubs, we are faced with the problem of finding the minimal collection of sets $\underline{\mathbf{C}}^* \subseteq \underline{\mathbf{C}}$ that covers all sensor units s^{o} *without repetitions.* The requirement of not repeating sensor units in the solution sets is addressed by requiring that the solution sets $\underline{\mathbf{C}}^*$ must be disjoint. Although this requirement is not part of the original Set Cover formulation, it does not alter the nature of the problem fundamentally. Greedy heuristics (among others) have been proposed to solve the Set Cover problem [107]. We use a weighted greedy algorithm, presented in Algorithm 1, to search for the elements of $\underline{\mathbf{C}}$ that have the largest cardinality.

The third and last step of our solution is now performed. The output of the previous heuristic is a set of selected elements $\underline{\mathbf{C}}^* \subseteq \underline{\mathbf{C}}$ that provide a solution to the Set Cover problem. The logistic locations associated to those sets of sensor units s^o are selected for deployment.

4.4 City-scale Characterization of I2V Data Volume Transfer

We propose a procedure to produce a city-scale model (or *map*) of estimated data transfers in I2V connections $\underline{\mathbf{M}}_{\mathrm{V}}$. The fundamental design principle of this procedure is that transferable

Algorithm 1: Greedy Algorithm

Data: $\underline{\mathbf{C}}$
Result: $\underline{\mathbf{C}}$*
$\underline{\mathbf{R}}$ = sort_by_cardinality($\underline{\mathbf{C}}$)
$\underline{\mathbf{E}} = \emptyset$
$\underline{\mathbf{C}}^* = \emptyset$
while $\underline{\mathbf{E}} \neq \underline{S}^o$ **do**
  $C_{max\_card} = \emptyset$
  for *i = 1 to* $|\underline{\mathbf{R}}|$ **do**
    if $|R_i| > |C_{max\_card}| \cdot c_i \wedge (\forall\, r \in R_i :\ r \cap \underline{\mathbf{E}} = \emptyset)$ **then**
      $C_{max\_card} = R_i$
    end
  end
  $\underline{\mathbf{C}}^* = \{\underline{\mathbf{C}}^*,\ C_{max\_card}\}$
  $\underline{\mathbf{R}} = \underline{\mathbf{R}} \setminus C_{max\_card}$
  $\underline{\mathbf{E}} = \underline{\mathbf{E}} \cup C_{max\_card}$
end

data volume, at a given distance and over a period τ, can be approximated by the product of two elements:

1. throughput at given distance;
2. contact time with vehicular nodes at that distance.

An important aspect affecting the design and accuracy of the I2V data transfer model is the tight dependency on the available datasets. For our procedure, we assume the existence of following datasets and models:

- measurement pairs of throughput and distance for IEEE 802.11b/g/n links;
- mobility traces or model of the vehicular nodes;
- vehicular terminal selection and connection setup latency models.

We see as reasonable to assume the existence of the two datasets in a vehicular network, as they entail only that GPS and WiFi are available in the OBUs. We describe next the input datasets and models and required pre-processing, the procedure to build the model of I2V data transfers $\underline{\mathbf{M}}_{\mathrm{V}}$, and discuss limitations of the model in terms of source, impact and dataset.

4.4.1 Inputs and Procedure for Model Generation

We model connectivity range and data transfers in 802.11b/g/n links at a given distance with an unit disk model. The model can use global or location-specific values obtainable from throughput-distance measurement pairs. We define a function $\zeta(d((x_z, y_z),(x,y))$ where d is the Euclidean distance between the coordinates of a micro-cell of interest z, (x_z, y_z), and the coordinates (x,y) of an arbitrary micro-cell, that outputs the value of throughput ζ measured for the range of distances

to which d belongs. Eq. 4.5 presents the mathematical formula of this unit disk model. The communication range r_c indicates the maximum distance at which communication can occur. With this model, we incur in the assumption of isotropic radiation.

$$\zeta(d) = \begin{cases} \zeta_1, & r_{\mathrm{u,0}} < d \leq r_{\mathrm{u,1}} \\ \zeta_2, & r_{\mathrm{u,1}} < d \leq r_{\mathrm{u,2}} \\ \ldots & \\ \zeta_{\mathrm{n}}, & r_{\mathrm{u,n\text{-}1}} < d < r_{\mathrm{u,n}} \end{cases} \tag{4.5}$$

The second input, the mobility traces, should correspond to GPS entries of OBUs of the vehicular network that operates as wireless backhaul. For convenience in computation, we discretize the raw traces both spatially and temporally. In spatial terms, we broke down the city map into a grid and respective cells are called "micro-cells". The exact coordinates of all relevant infrastructural elements (e.g. traffic lights, infrastructural APs, etc.) are mapped into coordinates of this grid. Temporally, the dataset may be discretized in periods τ, again for ease of use. The resulting "map of presence time" $\underline{\mathbf{M}}_{\mathrm{T}}$ keeps a binary value regarding the existence of at least one OBU at a micro-cell with coordinates x at time interval t, for all micro-cells.

We introduce two model abstractions to address limitations of the aforementioned datasets. Note that we estimate I2V transfers for the subset of micro-cells where communication hubs can potentially be installed. If, for a given micro-cell, multiple OBUs are within communication range r_c, a *vehicular terminal selection model* is necessary to decide which OBU is the hub connected to. Currently, we define that the hub always connects to the closest OBU. Furthermore, the time interval that each OBU passes within r_c of the tentative deployment location (referred to as *presence time*) does not account for the latency in connection setup. We define a *connection setup latency model* as a time period ε that reflects setup latency and is subtracted from the overall presence time, resulting in the actual connection time. We are currently assuming said ε to be null. The selected model implementations are motivated in the following subsection.

Overall, the calculation of the data volume v transmitted between a micro-cell z and the vehicular network, over the course of a pre-defined time period $(\tau \cdot \Gamma)$ is as follows. For each time interval t_i of set $\{1,...,\Gamma\}$, we extract from $\underline{\mathbf{M}}_{\mathrm{T}}$ the coordinate of the closest micro-cell that contains record of an OBU at that instant t_i, $x_{closest}(t_i)$, and calculate its distance to z, $d(x_z, x_{closest}(t_i))$. The data volume $v(x)$ is then approximated by the sum of the product of the throughput corresponding to $d(x_z, x_{closest}(t_i))$ (drawn from the unit disk model), over all time intervals $\{1,...,\Gamma\}$, and the time resolution τ.

$$v_{\mathrm{z}}(d) = \tau \sum_{i}^{\Gamma} \zeta(\ d(x_z, x_{\mathrm{closest}})\ ), \tag{4.6}$$

Figure 4.11 describes the process of transferable data volume for a hub located in a specific micro-cell.

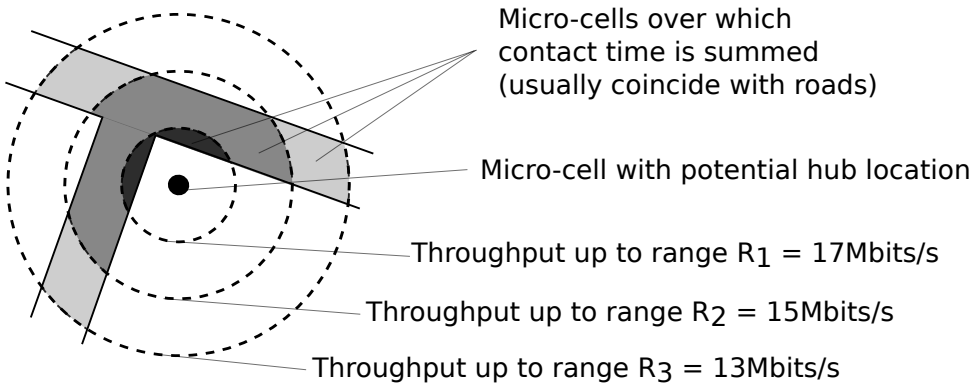


Figure 4.11: Computing data volume for a micro-cell.

4.4.2 Discussion on Model Accuracy

We discuss the assumptions of the model, their impact, and how they stem from the accuracy and granularity of the available datasets. They are:

1. *isotropic radio propagation* (i.e., communication is assumed equal in all directions within fixed range r_c);

2. *null connection setup latency* (i.e, presence time of an OBU equals actual contact time with the hub and no time is spent in connection setup);

3. *preferential connection to the closest vehicular terminal* (i.e., hub is always connected to the closest OBU).

The three assumptions may concur to an over-estimation of the actual transferable data volume of a micro-cell.

The impact of an isotropic propagation assumption with a fixed communication range r_c is two-fold: (i) the value selected for r_c may be inaccurate and even differ based on orientation; (ii) it may account for data transfers with OBUs that are not in line-of-sight due to buildings or other obstacles. The incorporation of topological information (e.g. Open Street Maps [108]) or the use of a connectivity map (as discussed in [109]), if available, may alleviate the impact of this assumption.

The second assumption ignores the the overhead time ε spent in the association and IP assignment process. This latency is dependent of implementation-dependent and circumstantial factors. In the first case, it depends on whether dynamic addressing is used and, if so, on the values assigned to the DHCP user-configurable timing parameters. An I2V system may use static addressing to obviate the overhead involved in IP assignment, although at the cost of a tighter integration with vehicular backhaul. If dynamic addressing is used nevertheless, many DCHP timing parameters may be varied to obtain smaller connection setup latencies, as we did in prior work [110] for example. A circumstantial factor stems from DHCP advertisement messages being missed before an IP assignment takes place and connection starts. This introduces a random process component

into the latency duration. Another circumstantial factor is lease re-utilization. Given that it is dependent of a previous connection to the same on-board AP, it introduces additional uncertainty on the duration of the setup. Given that an universal value or distribution cannot be assumed for this latency, we opted for a null ε for the purpose of this explanation.

The third assumption results from the model's uncertainty about which OBU is the communication hub currently connected to, if multiple are available. If an incorrect OBU is used, the associated OBU-hub distance will also be incorrect, ultimately leading to an incorrect throughput value. As mentioned earlier, we currently assume that the connection is always with the closest OBU. More sophisticated models would require estimating terminal selection and connection maintenance tendencies, which would involve dedicated measurements or simulation.

4.5 Framework Application to a Medium-Sized City and Evaluation

We now present an example application of our framework to a medium size European city: Porto, Portugal. We introduce the scenario in Porto, and use it as a basis to explore the parameter-space of our framework and compare a selected placement solution against an actual deployment.

4.5.1 Input Datasets for Framework

We describe the inputs necessary to apply our framework to the scenario of Porto described in Section 4.1. The location of the sensor units $\underline{\mathbf{S}}$ encompasses 73 planned *UrbanSense* locations to be monitored. The set of logistic deployment locations $\underline{\mathbf{U}}$ corresponds to the location of traffic lights. This dataset was made available by the municipality, and counts 290 traffic lights. Our dataset does not discriminate individual traffic light poles, but only indicates the geographical center of intersection the traffic lights are associated to. The set of fixed gateways $\underline{\mathbf{A}}$ concerns the location of the 63 *Porto Digital* access points, that the municipality also provided.

The remaining input is the model of estimated I2V data transfers $\underline{\mathbf{M}}_{\mathrm{V}}$ specific for Porto, that must be generated from the model of throughput vs. distance and mobility traces of the vehicular nodes. The model of throughput with respect to distance was sourced from data of the measurement campaign described in Section 4.2. Regarding the mobility dataset, it corresponds to one week of GPS traces of the *BusNet* OBUs, specifically days 12 to 19 of January 2015, provided by the service operator. The micro-cell size was set to 10 meters tall (in latitude) and 7.5 meters wide (in longitude), and the resulting grid is 528 micro-cell tall per 1511 micro-cells wide. The raw dataset was discretized in periods τ of 15 seconds. A "map of presence time" $\underline{\mathbf{M}}_{\mathrm{T}}$ was generated from this dataset, by summing the number of GPS entries recorded at each micro-cell. After applying the procedure of Section 4.4, the produced estimations of I2V data transfers $\underline{\mathbf{M}}_{\mathrm{V}}$ were limited to locations of utilities $\underline{\mathbf{U}}$. Although it could be performed for the entire city, we opted to do so partly because the locations $\underline{\mathbf{U}}$ are the only relevant locations for the problem (i.e., where hubs can be installed), and partly to limit the computation times involved in the model generation.

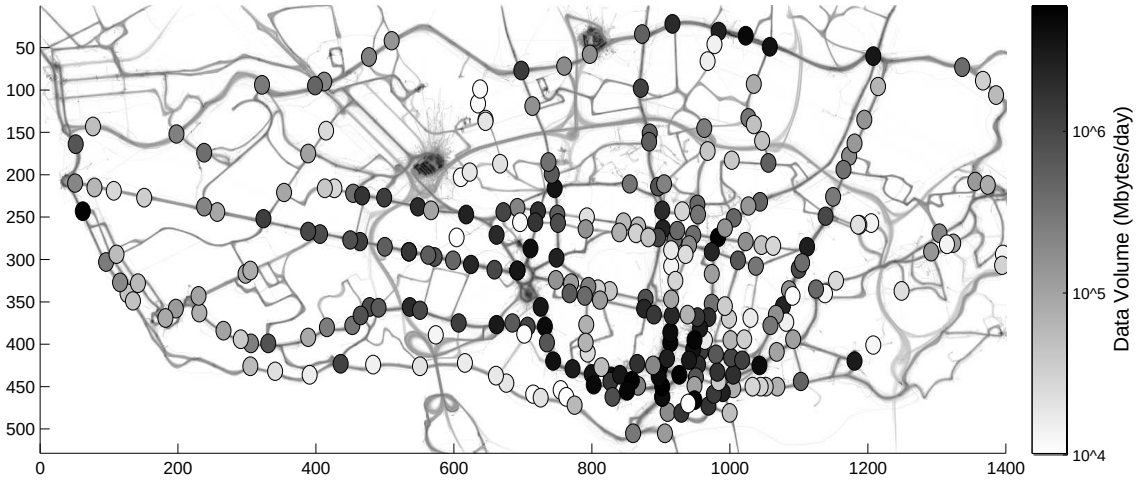


Figure 4.12: Map of transferable data volume for each micro-cell for r_c equal to 100 meters. Diameter of circle is r_c meters, face color indicates data volume in grey scale. Map of accumulated presence time M_t shown in background for geographical reference.

Figure 4.12 presents the map of estimated I2V data transfers $\underline{\mathbf{M}}_{\mathrm{V}}$ at the logistic locations, indicated by the circles whose color in gray scale is proportional to our data transfer estimation, and in background the histogram of "presences" $\underline{\mathbf{M}}_{\mathrm{T}}$ of vehicular nodes over the entire period.

4.5.2 Solution Production and Parameter-Space Exploration

We evaluate the gains brought by our framework for a placement of communication hubs to serve the set of 73 tentative sensor locations, and how the balance of hubs served by each backhaul changes as the values of the framework parameters are modified. We contextualize the gains brought by solving the Min-Hub Problem by presenting them against the placement of hubs under a one-hub-per-equipment policy (i.e., not sharing hubs). We call these results the *baseline* solution.

We define as independent variables the maximum range for deployment r_{d} and communication r_{c} and user-defined costs c^f and c^v enforcing a particular policy; the minimum data volume to be served $v_{\min}$ is kept fixed. Table 4.3 presents the values used for relevant parameters; underlined values correspond to default values used throughout this discussion. The set of values chosen for r_{d} aims to model the distances between hub and sensor unit at which an off-the-shelf wireless communication technology (e.g., ZigBee, Bluetooth or WiFi) could reliably connect the two terminals, or a wired infrastructure could be installed in an urban environment connecting the two terminals. The two values chosen for r_{c} provide respectively a conservative and a optimistic estimations of the maximum range at which WiFi can be expected to operate in urban I2V scenarios. The minimum data volume transfer $v_{\min}$ that must be guaranteed for each sensor unit is kept fixed at 1 Mbits/week. The value chosen for $v_{\min}$ is aligned with the estimate of the data volume produced by a sensor unit for a period of a week – as aforementioned, a sensor unit produces approximately 628.22 kbits over the course of one day, amounting to 4.39 Mbits over the course

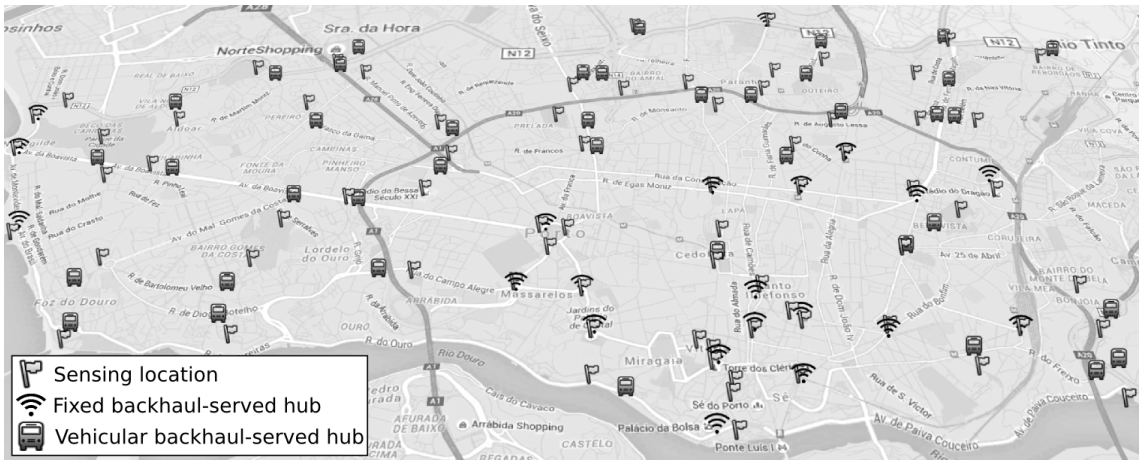


Figure 4.13: Sensor units locations, and solution locations for Min-Hub Problem with r_{d}=300m, v_{min}=1 Mbit and preferential used of fixed backhaul.

of one week. The two policies for setting the value of user-defined costs are those discussed at the end of Section 4.3.1: preferential use of fixed backhaul and preferential use of vehicular backhaul.

An example placement solution for the Min-Hub Problem is shown in Figure 4.13 for r_{d} = 300m, r_{c} = 200m, v_{min} = 1Mbits and preferential use of the fixed backhaul ($c_i^f = q_i^f; c_i^v = \bar{c}_i{}^f$). For computability reasons, the values of binary variables c^f and c^v were approximated by very distant values (0 as 10^{-5} and 1 as 10^5).

4.5.2.1 Baseline Solution

The baseline problem is identical to the Min-Hub Problem, but the requirement of minimizing the number of hubs is relaxed: each sensor unit is to be served by a dedicated hub. The same cost-function and constraints of the Min-Hub Problem are used, but we refrain from carrying out the minimization procedure. Instead, the closest logistic location is associated to each sensor unit. It may occur that some traffic lights serve more than one sensor unit. We relax constraint **(c6)** by not verifying if a given traffic light pole support the I2V volume transfer of all assigned sensor units (for the vehicular service). This relaxation may lead to inexact solutions but always produces the minimum number of hubs that can be placed under the one-hub-per-equipment policy for a given r_d (and thus does not result in advantage to the solutions found for the Min-Hub Problem).

A characterization of baseline solutions over a range of r_{d} values is shown in Figure 4.14. For r_c=200 meters, Figure 4.14a shows the maximum number of hubs that can be placed independently per backhaul (even if serving the same sensor unit), and the aggregated number of served sensor units (number of sensors units that can be served by one or other backhaul). We observe that, for both backhauls, an increase in the deployment range r_d increases the number of deployable hubs and thus of served sensor units. The coincidence between the number of served sensor units and the vehicular backhaul-served sensor units shows that the vehicular backhaul could support all locations on its own. Three aspects concur to this result: (i) the larger area of coverage of

Parameter	Value
Number of sensor locations **S**	73
Minimum data volume $v_{\min}$	1 Mbits
Deployment range r_{d}	{50, 150, 200, 250, 300}m
Communication range r_{c}	{100, 200}m
User-defined costs c^f, c^p	{fixed-preferential,vehicular-preferential}

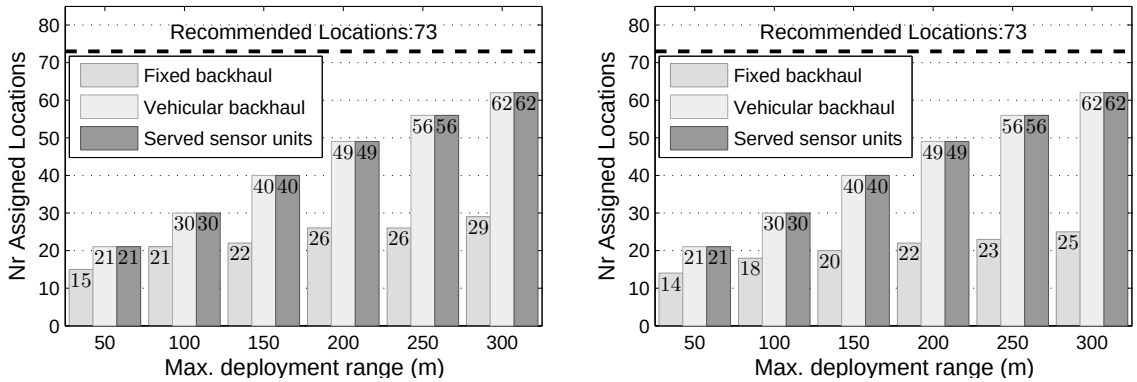
Table 4.3: Parameter values.

the vehicular backhaul with respect to the fixed backhaul; (ii) a disparity between the data volume requirements of *UrbanSense* sensor units and the estimated I2V data transfers – at most 15 Mbits/week required vs. 10+ Gigabyte/day (see Figure 4.12); and (iii) the estimated I2V data transfers may in turn be overestimated (as discussed in Section 4.4.2). Figure 4.14b presents the same metrics for r_c=100 meters. The number of fixed backhaul-served hubs decreases with respect to the case of $r_c = 200$ meters, whereas the number of vehicular backhaul-served hubs does not vary.

4.5.2.2 Min-Hub Problem Solution

We now evaluate the solutions of our framework for the Min-Hub Problem, again over a range of r_{d} values (Figure 4.15). Figure 4.15a presents the ratio between the number of placements of the shared-hub solution over the number of placements of the baseline (one-to-one) solution. It can be seen that this ratio becomes inferior as r_c increases, showing that increasingly larger deployment radii improve the benefit that hub sharing brings with respect to assigning each sensor unit a dedicated hub. Figure 4.15b breaks down the total number of deployed hubs in the shared-hub solution per backhaul. It is noteworthy that, for small r_d values, the fixed backhaul has more associated hubs than the vehicular backhaul and, as r_d increases, the trend inverts. Due to the policy of preferential use of fixed backhaul, the maximum number of fixed backhaul-served potential deployment locations becomes assigned throughout the full range of r_d values. However, given the relatively small number of APs in the city, the number of vehicular backhaul-served locations surpasses the first as r_d increases and more sensor units can be served.

We now look at the impact of selecting a different policy to compute the user-defined costs (Figure 4.16). Figure 4.15c presents the output of the placement strategy if a policy of preferential use of the vehicular backhaul is followed. Across the range of r_d, the framework assigns all hubs to be served the vehicular backhaul, and for some deployment radii (e.g. 300 meters) attains a number of placements inferior to the policy of preferential service by the fixed backhaul. The cause is that independent solutions for the vehicular and fixed backhauls present an high overlap of serviceable sensor units, as shown in Figure 4.15b. Finally, we discuss how different communication ranges r_c impact the deployment solution. Figure 4.15d depicts the breakdown of the placement solution for r_c=100 meters. We observe similar trends to those of Figures 4.15b and 4.14b: the fixed backhaul



(a) Number of hub locations for unshared placement solution, with r_c=200m.

(b) Number of hub locations for unshared placement solution, with r_c=100m.

Figure 4.14: Solution metrics for a range of r_{d} values, fixed $v_{\min}$=1 Mbit and r_c=200 meters.

has less hubs assigned over all the r_d range and, as the value of r_d increases, the number of fixed backhaul-served hubs stabilizes as the vehicular backhaul-served increases.

4.5.3 Solution Evaluation against Real-World Deployment

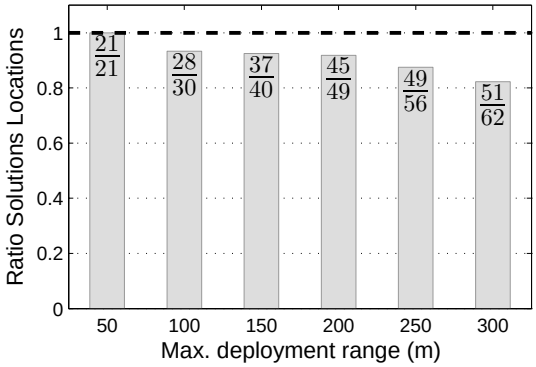
We compare now selected placements produced by our framework against the *UrbanSense* deployment in Porto. The *UrbanSense* platform comprised, in a first stage, 22 co-located sensor units/hubs (DCUs) deployed at a subset of the recommended 73 tentative locations, installed from mid to end of 2015. Since then, three DCUs had to be removed due to excessive wear and corrosion, specifically those deployed near the sea. All field deployment locations are shown in Figure 4.17.

Due to the very different set of constraints that each backhaul encompasses, we perform our evaluation separately for the fixed (Section 4.5.3.1) and vehicular backhauls (Section 4.5.3.2). Separate placement solutions have been computed for each case.

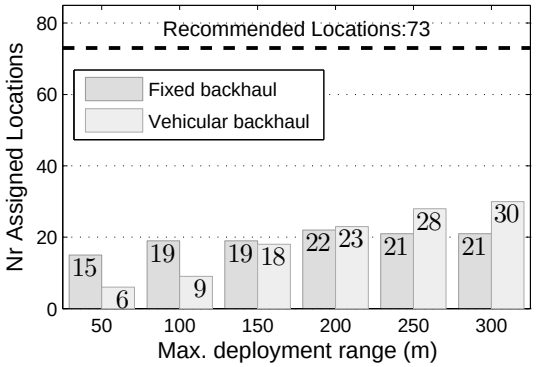
4.5.3.1 Service by Fixed Backhaul

We computed a placement solution produced for comparison purposes that takes ample communication and deployment radius, particularly r_c=200 meters and r_d = 300 meters, and preference for the fixed backhaul. The overall solution resulted in 53 placed hubs, of which 23 are served by fixed backhaul and 30 by the vehicular backhaul. This particular placement solution is also shown in Figure 4.17. The remainder of this discussion focus only on locations served by the fixed backhaul. Regarding the *UrbanSense* deployment, all 22 installations were in range of an fixed backhaul gateway. Connectivity between server and DCUs was evaluated with a remote connection.

The metrics used in our comparison between the placement solution and the field deployment are explained next. The main figure of merit is the ratio of placements used with respect to the all *UrbanSense* sites. A placement is considered *used* if a DCU was deployed up to 100 meters of

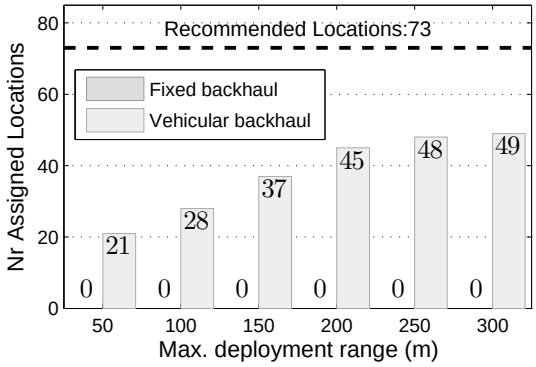


(a) Ratio of shared-hub placement solution to one-to-one placement solution.

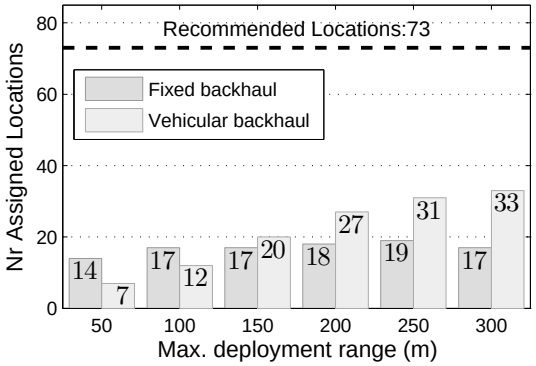


(b) Number of hub locations for shared-hub mixed-backhaul service placement.

Figure 4.15: Solution metrics for a range of r_{d} values, with default parameter values of r_c=200 meters, preferential use of fixed backhaul, and fixed $v_{\min}$=1 Mbit.



(c) Number of hub locations for shared-hub mixed-backhaul service placement, with preferential use of vehicular backhaul.



(d) Number of hub locations for shared mixed-served placement, with r_c=100 meters.

Figure 4.16: Solution metrics for a range of r_{d} values, with default parameter values of r_c=200 meters, preferential use of fixed backhaul, and fixed $v_{\min}$=1 Mbit.

the recommended location. Other metrics are related to discarded, re-located, unplanned or non-operational placements (fully or partially). Discarded placements are framework-recommended locations where installation was absolutely not possible. A placement is considered re-located if changed to location farther than 100 meters. Unplanned locations are those where a placement was not recommended by the framework but a DCU was installed. Finally, unsuccessfully-operated locations are those suffering of insufficient service from utility or communication infrastructures, or that have been deactivated due to unforeseen factors. The later metric can only be assessed after installation, whereas the first four metrics report to a pre-installation stage.

We now evaluate the placement produced by our framework application. Table 4.4 summarizes the results. With respect to pre-installation metrics, our placement compares to actual deployment in the following manner:

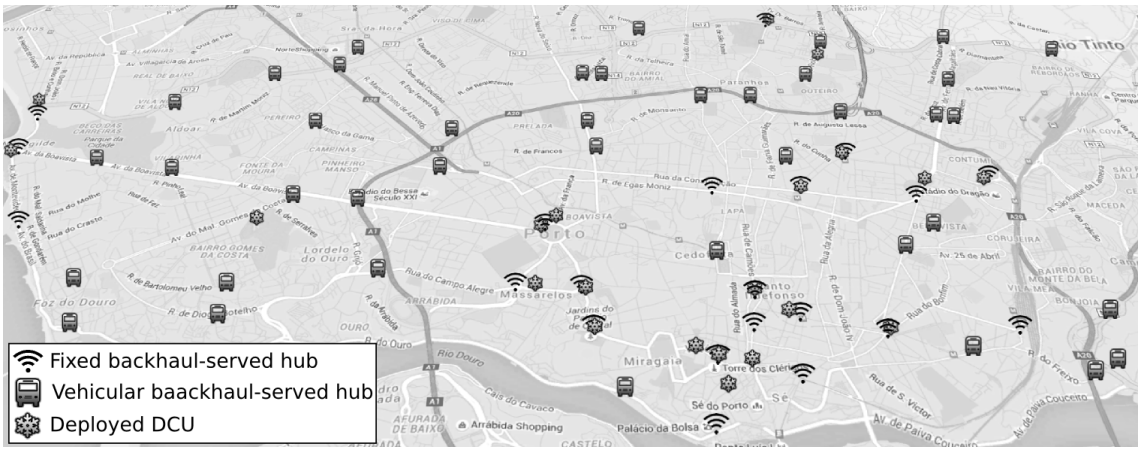


Figure 4.17: Locations of placement solution and field deployment in Porto.

- **Used placements:** of the 22 deployments, 13 DCUs (around 60%) are located at logistic locations (traffic lights) recommended by the framework.
- **Discarded placements:** there were four locations where deployment was recommended but discarded at the planning stage. In two locations the APs lacked wired connection to the municipality's fiber ring and were thus nonoperational. The third required an expensive intervention to install power cables, and the fourth was considered redundant with other deployments due to the similarity of the monitored area (seaside).
- **Re-located placements:** six deployments had to be re-located to a location farther than 100 meters than the framework-indicated traffic light, as the municipal services did not deem those traffic lights suitable for DCU installation.
- **Unplanned placements:** Three deployed locations were not recommended by the framework, of which two locations refer to buildings of partner institutions where power supply was available. Of these, one location had no traffic lights nearby; the other location had nearby traffic lights, but the solution location for that sensing location referred to a different set of traffic lights, as a result of the DCU-sharing goal. The third non-recommended location refers to a test unit.

We observe that reasons for discarded and/or re-located placements were due to insufficient or out-dated dataset detail from municipal authorities. The unplanned placements did not meet the constraints imposed to the framework and/or are motivated by constraints not included in our framework.

As for post-deployment evaluation, we observed after deployment that 19 of the 22 deployed DCUs had sufficient or good WiFi service. The remaining three DCUs, two of which belonging to the subset of 13 DCUs deployed at recommended locations, are considered nonoperational given that there is insufficient WiFi service to sustain data collection. We took a closer look to the deployment scenario of the two solution-recommended placements to understand what failed in the framework operation or inputs. Given the wireless nature of the communications links,

Global Numbers	Fixed backhaul-served placements	23
	Deployed DCUs	22
Action w.r.t. placement solution	Used	13
	Discarded	4
	Re-located	6
	Deployed although not placed	3
Result w.r.t. placement solution	Insufficient service	2

Table 4.4: Action/outcomes of field deployment with respect to placement for r_c=200m, r_d=300m and preferential use of fixed backhaul.

the assumptions made in our framework are likely to have a greater impact than in the utilities case. The two locations, which we name "Stadium" and "Trindade", had different reasons causing insufficient connectivity to nearby APs:

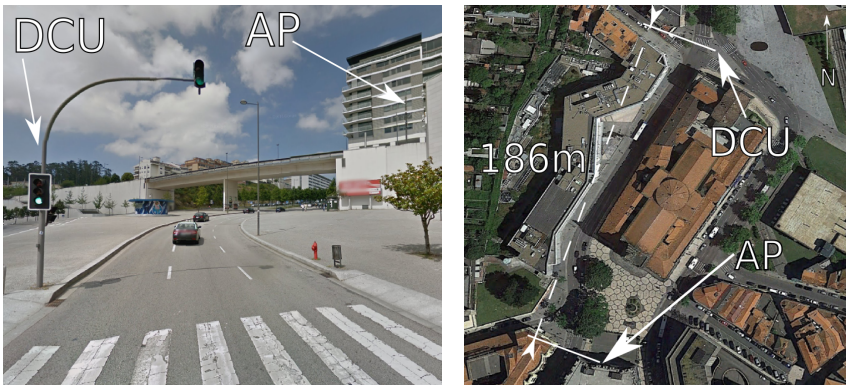
- Stadium: there is a nearby AP, but it has a considerable altitude offset to the DCU that greatly affects link quality. The solution, which is on-going, is for the municipality services to tilt the AP antenna. The reason for the framework not to predict insufficient service at this location (and perhaps not assign this placement) was thus due to outdated or insufficiently detailed datasets.
- Trindade: it is located close to a municipality AP, but experienced connectivity quality is poor due to obstructing buildings. This was caused by our assumption of isotropic radiation within a communication range r_c.

Figure 4.18 provides graphical description on these particular sites. The solution found for these sites was to install a cellular hotspot to provide stable communication. Two DCUs have been removed due to unforeseen factors, namely excessive wear and corrosion in sea-bordering areas, and a third due to malfunction.

The framework may be adapted to reflect some of these lessons, such as including altitude differentials or identify areas where the potential for wear/corrosion is higher. As discussed in Section 4.4.2, as datasets become more comprehensive, the solution quality also improves.

4.5.3.2 Service by Vehicular Backhaul

We now evaluate the performance of our placement framework for service by the vehicular backhaul. It is not possible to carry out field tests at all potential deployment locations in the city, nor test all potential associations of sensor units and hubs. Furthermore, the heuristic used to approximate a minimal solution for the Min-Hub Problem and its algorithmic performance has been analyzed previously in the literature [111], and thus we will not evaluate the quality of produced



(a) Stadium setup. (b) Trindade setup.

Figure 4.18: Framework-recommended locations suffering from insufficient service.

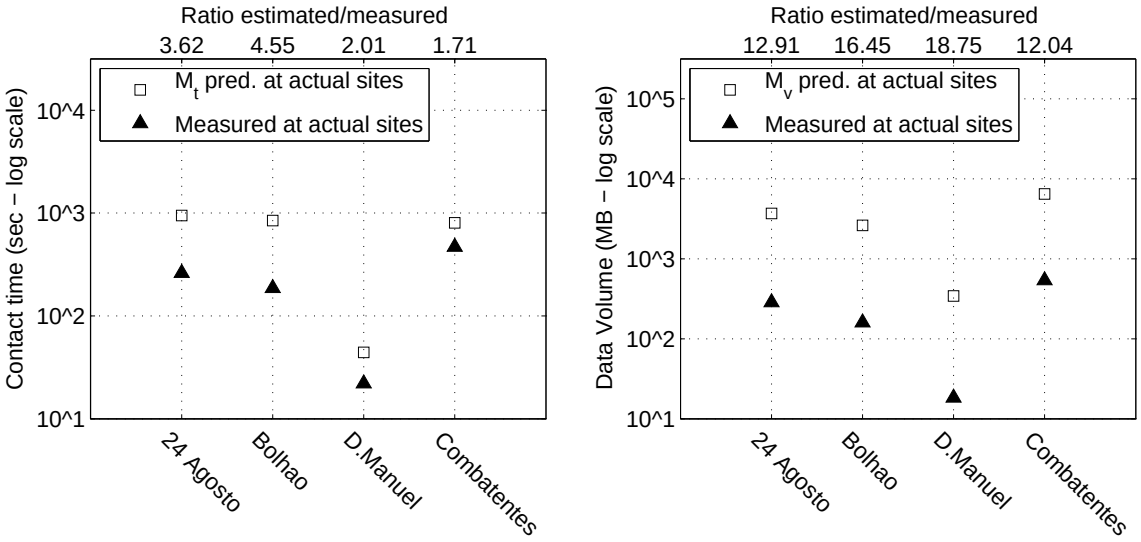
solutions. Thus, we limit our validation to the tool used to perform the placement, the map of I2V data transfers $\underline{\mathbf{M}}_{\mathrm{V}}$.

The accuracy of $\underline{\mathbf{M}}_{\mathrm{V}}$ is evaluated against actual infrastructure-to-vehicle measurements carried out at the DCUs during September of 2016. To avoid any impact of the I2V measurements in regular bus WiFi service operation, the service operator requested the I2V measurements to be performed with vehicles of the municipal fleet, composed of garbage-collecting and street-cleaning trucks, and which we designate by *CMP vehicles* and *network*. Given this restriction, we computed a new $\underline{\mathbf{M}}_{\mathrm{V}}$ specifically for this network by using the GPS traces of the CMP vehicles. Four DCUs were assigned to this experiment, which we call for convenience as "24 Agosto", "Bolhão", "D. Manuel" and "Combatentes". Measurement software was installed in ten OBUs of the CMP network. The experiment ran from September 5th to 25th 2016.

We produced a placement solution that took a communication radius of r_c=100 meters, deployment radius of r_d = 300 meters, and preference for the vehicular backhaul. The overall solution resulted in 48 placed hubs, all of which served by the vehicular backhaul. Of the 48 placements produced, two were co-located with *UrbanSense* DCU locations ("24 Agosto" and "Combatentes").

The values obtained in the I2V measurements and the predictions of the model $\underline{\mathbf{M}}_{\mathrm{V}}$ are plotted in logarithmic scale, as average of daily totals over the aforementioned days, in Figures 4.19a and 4.19b for the contact time and transfered data volume respectively The ratios between estimated and measured are also shown on the top of the plots. We observe that, in all four locations, the estimated transferable data volumes are over-estimated between 12 to 18-fold with respect to the measurements. It is worthy noting that, despite the error in the absolute value estimation, the model is able to capture the relative order of the locations in both predicted and measured values.

The discrepancies in the absolute values result from the three assumptions inherent to the current version of our procedure to generate the map of I2V data transfers, discussed in Section 4.4.2. Those are: (i) isotropic radiation; (ii) the DCU is always connected to the closest OBU within range; and (iii) connection lasts for the entire period an OBU is within range (no connection setup



(a) Measured and estimated contact time at selected DCUs. Average of daily totals shown.

(b) Measured and estimated data volumes at selected DCUs. Average of daily totals shown.

Figure 4.19: Comparison between measurements and framework predictions.

latency). The impact of the first assumption cannot be verified without a city-wide wireless propagation map, which we do not have available. The second assumption could not be evaluated due to the nonexistence of occurrences. We observed no cases of multiple OBUs being simultaneously within range of a DCU in our dataset, due to the small number of nodes of the CMP network and the nature of their task (garbage-collection trucks). This also causes this assumption to have a residual impact in the our estimated values of $\underline{\mathbf{M}}_{\mathrm{V}}$.

The third assumption implies that our procedure to generate $\underline{\mathbf{M}}_{\mathrm{V}}$ does not account for latency-inducing processes such as AP discovery and IP assignment. To evaluate its impact, we compared with measured and estimated contact times per connection. For each connection, we computed from the GPS traces the moment the participant OBU comes into communication range of the DCU and the moment the OBU leaves the range, to estimate the predicted contact time for that connection. Figure 4.20 plots the CDF of all connections with respect to the ratio of measured over predicted contact time. We observe that the model over-estimates the contact time in 90% of the connections, which accounts for explaining the mismatch between measured and estimated data volumes. The remaining fraction of measured connection times outlasts the respective estimated contact duration, accountable to connections for which the maximum connection range (r_c = 100m) was conservative.

In conclusion, our procedure to model the transferable data volumes provides over-estimated but reasonably approximate estimates for each potential location in the city. We point out that our model observes the relative ordering of the locations, which helps identifying which locations perform better. Some alternatives to improve the model of I2V transfers estimates are discussed in Section 6.3.

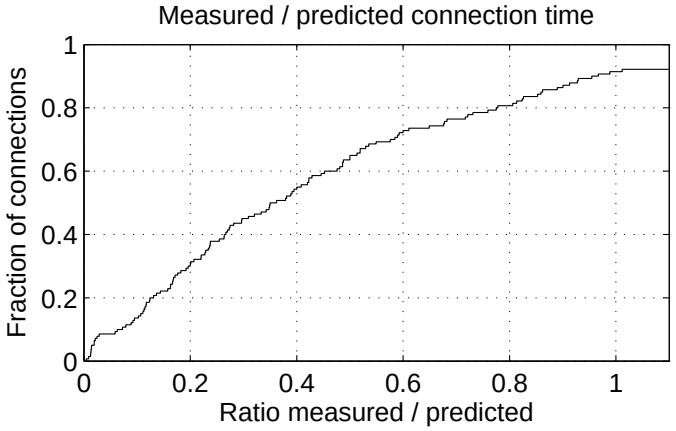


Figure 4.20: CDF of measured over predicted contact times per connection.

4.6 Final Remarks

We have addressed in this chapter the topic of characterizing wireless links at the data-link layer between terminals with a relative speed larger than zero, and the challenge of placing road-side communication hubs that support a sensor equipments deployed at disparate locations. In an initial stage, an experimental study in an urban testbed with a road-side WiFi-equipped unit and the buses of a public transportation fleet equipped with on-board WiFi access points allowed us to characterize the throughput, data volumes, number and duration of connections.

On a second stage, we developed an optimization-based decision support framework to produce a placement for road-side communication hubs that forward data from pre-placed sensor equipments via existing wireless backhauls. The placement is constrained by closeness to client equipments, logistic limitations (e.g. availability of power supply) and existence of wireless links to backhaul gateways, and it is assumed that a single hub can be serve multiple equipments. The problem is formulated as an optimization problem with the cost-function of minimizing the number of hubs that must be deployed. A solution strategy for the minimization aspect of the problem was devised and relies on a greedy heuristic to solve the Set Cover problem, a known-NP hard problem into which our minimization problem maps. In order to model the service by vehicular backhauls in our framework, we developed a model to estimate the transferable data volumes in I2V links from a dataset that we assume common in fleet operators (mobility traces of the vehicles) and a model of throughput with respect to distance obtained from a field measurement campaign.

Taking a platform of weather monitoring stations in Porto, Portugal, as an example case, our framework produces the result that the number of hubs can be 20% inferior than the number of sensors, if large deployment ranges are allowed. Comparing the placement solution of our framework with a field deployment of 22 equipments, we observe that, for service by an infrastructural backhaul, almost 60% deployed hubs are located up to 100 meters of a solution location, and 87.5% had good or sufficient WiFi service. For service by a vehicular backhaul, our city-wide estimation of I2V data volumes estimates the actual data volumes to within a order and a half of magnitude and accurately ranks locations according to relative performance.

Chapter 5

Data Collection in Dynamic Topologies and Design of Network Coding Protocol

In scenarios where data produced by mobile nodes must be collected at a static base station, many-to-one routing strategies can harness beaconing for faster and more efficient route setup. This solution consists in having the base station issuing beacon packets regularly that the nodes can use to set up a routing structure – for example a minimum spanning tree (MST), as done by the reference Collection Tree Protocol. This strategy is subject to a progressive degradation of the routing information at the nodes, which in turn may lead to packet mis-routing and losses. We propose the use of wireless broadcast and opportunistic forwarding to overcome the limitation of using rigid routes that fade over time in scenarios with dynamic topologies. The option for wireless broadcast involves abdicating of the link-layer retransmissions, but network coding can provide an alternative to implement reliability. By allowing nodes to code packets from various sources and in conjunction with the opportunistic forwarding, we promote data dissemination and replication across the network so that enough degrees of freedom arrive to the base station.

In this work, we study the design space of a data collection protocol for scenarios with mobility that builds on periodic beaconing, opportunistic forwarding, and network coding. In addition to aspects related to routing and reliability, a network coding strategy encompasses a number of possible configurations relating to coding breadth, coefficient pool size, generation size and associated payload, and coding policies at intermediate nodes. We build a framework protocol in a modular fashion to evaluate and compare the performance of alternative configurations of the protocol, via simulation over connectivity traces obtained from a real-world vehicular testbed. Finally, we benchmark our proposal protocol against a state-of-the-art structured protocol, the Collection Tree Protocol, and evaluate its resilience against topology changes.

Our contributions are as follows:

- An identification of the design and parameter aspects that a network coding protocol implies, and an analysis of the existing literature on those design aspects;

- Performance comparison of alternative implementations of the protocol, by means of simulation using connectivity traces of a real-world vehicular testbed;
- Performance comparison of the various protocol configurations against a benchmark structured protocol, CTP.

The remainder of this chapter is organized as follows. Section 5.1 motivates the problem with an analysis over traces and simulation results of a MST protocol from data sourced from a real-world setting. Section 5.2 overviews the open options in the design of network coding protocol for data collection, and describes the implementation of structural and modular elements of a framework protocol developed to test different design configurations. Section 5.3 describes the traces pre-processing and simulation setup, and presents the performance evaluation and comparison of the various configurations of the protocol. Finally, Section 5.4 draws the conclusions from our analysis, with practical insights regarding design of a network coding protocol.

The current work was done in collaboration with Prof. Daniel Lucani, and a publication is under preparation.

5.1 Routing Information Lifetime

The application we target is that of multi-hop data collection over an ad hoc network with mobile and static nodes to a base station. In spite of the wide body of routing for MANETs, this particular application is base station-centric, and thus it lends itself to the use of beaconing and routes that follow a minimum spanning tree (MST) configuration. A minimum spanning tree is a minimum set of paths that emanate from the base station and reach all nodes. Setting up routes over a MST is a strategy used by a number of data collection protocols, chief amongst which the Collection Tree Protocol (CTP). In situations where a fixed beaconing rate is used, the routing information stored at the nodes becomes progressively outdated. As larger intervals between beacons are used, the impact of the routing information degradation affects the performance of spanning tree-based protocols.

To provide some insight into this phenomenon, we analyzed the traces and simulated the operation of this protocol over the connectivity traces of a vehicular network. We evaluated first the rate of the degradation of the optimal routes from the connectivity traces, and afterwards we studied the impact in the performance of fixed-beaconing data collection protocols such as CTP by means of simulation. The vehicular network is installed in container port in Porto, Portugal [112], and counts 20 container-carrying trucks with on-board units (OBUs) and 7 road-side units (RSUs). The communication hardware between vehicles and road-side units is IEEE 802.11p/DSRC standard hardware and stack. The OBUs are programmed to broadcast 10 beacons per second (consuming a duty cycle of 10%) and registed to beacons from other nodes. In post-processing, the PDR of all the links available to a node, at any given second, is approximated by the ratio of how many beacons from that neighbour, out of 10, have been received. Figure 5.1 presents the premises where the vehicular network operates.

Figure 5.1: Harbour premises.

	Parameter	Value
Scenario	Data prod. bitrate v_i	10 KB/s
	Data Δt_{beacon}	Equal to t_{b}
	Max transmit rate	2 Mbit/s

Table 5.1: Parameters used in CTP operation simulation.

5.1.1 Rate of Topology Change

We evaluate the rate of topology change by computing the distribution of the duration of minimum spanning trees in seconds for each RSU independently for a dataset of one week of traces. For an arbitrary second t of the trace, we verify the duration of the optimal tree at t over subsequent seconds, even if the tree becomes non-optimal. If the tree of an arbitrary second $t+1$ is the same as that t, we do not include the duration of this tree in our sample pool. A tree under analysis becomes invalid if: (i) one of the links of the tree at t has disappeared; (ii) the number of connectable nodes by the optimal tree at $t+\Delta t$ is larger than that of the optimal tree at t; and (iii) the number of connectable nodes at t and $t+\Delta t$ is the same, but there are new nodes. If the number of nodes at $t+\Delta t$ is smaller than t's, but existing nodes are still connectable by the tree of t, the tree is still considered valid.

Figure 5.2 shows the results of this analysis. We can see that 25% of the trees last 3 seconds or less, and that more than 50% of the trees become invalid after 6 seconds (52% last 6 seconds or less).

5.1.2 Impact in CTP Performance

After the previous traces analysis, we simulated the operation of CTP with two beaconing intervals t_{b}– 3 seconds and 6 seconds –, for a single day of traces and the parameters shown in Table 5.1, All nodes are constant bit rate (CBR) sources, and beacons are broadcast by the RSU carrying a request for data that must be delivered before the next beacon transmission. The results are shown in Figure 5.3. It can be seen that the performance of CTP for larger values of t_{b} degrades, as the overall PDR for beacons/requests transmitted every 3 seconds is superior to that of 6 seconds. This

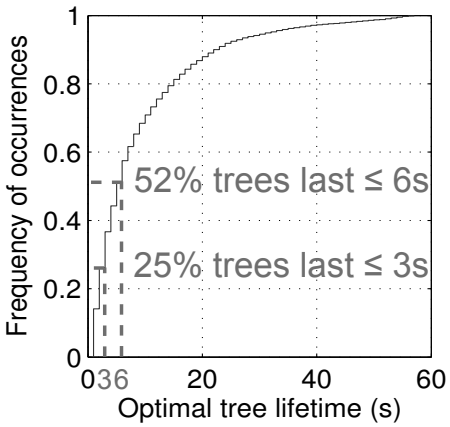


Figure 5.2: Lifetime of MSRTs in vehicular traces.

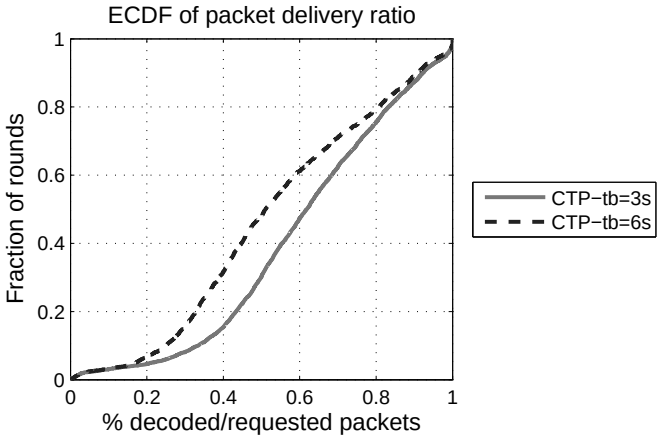


Figure 5.3: Decrease in performance of CTP as t_b increases.

shows how the performance of minimum spanning tree-based protocols with fixed beaconing may be impaired by outdated routing information.

5.1.3 Conclusion

These results provide the context and motivation to pursue an improvement over the existing solutions. We propose that the problem can be alleviated, still in the context of a base station-centric protocol, by harnessing wireless broadcast and opportunistic forwarding for data packet relaying. This allows to explore more routes that those of the minimum spanning tree and take advantage of the best available forwarders at any given instant, but turns the use of link-layer retransmissions impractical. An alternative reliability mechanism can be implemented by a network coding solution, that relies on end-to-end feedback as proposed in [74] and [73] for data collection applications. We discuss next the structure and design possibilities of such protocol.

5.2 Network Coding Protocol Design Space and Specification

The design of a data collection protocol based on opportunistic forwarding and network coding is subject to a variety of options and implementations. We created a framework protocol to structure and systematize the exploration of such design space, following a modular approach that allows different module implementations to be swapped in and out to produce different protocol configurations. The high level structure and design philosophy of the framework protocol is inspired by the previous work of network coding protocols, as discussed in Section 2.3.

We break down the protocol design and operation into three planes – **Routing**, **Network Coding** and **Reliability**, for convenience in the following discussion. These planes are supported by dedicated mechanisms that map into modules of the framework protocol.

- **Routing Plane:** includes the mechanisms that route packets and manage traffic towards the base station. These are the *Forwarding Engine* and *Congestion Mitigation.*

- **Network Coding Plane:** handles the design aspects related to coding, grouped into the mechanisms of *Generation Management* and *Mixing/Coding*, with a number of coding-related design options.
- **Reliability Plane:** addresses the performance improvement of the protocol by means of two mechanisms, *Redundancy* and *Feedback*.

In the remainder of this section, we provide an overview the design-space of a network coding protocol in Subsection 5.2.1, and describe the implementation of the modules and operation of the protocol in Subsection 5.2.2.

5.2.1 Design Space of a Network Coding Protocol

We review the design-space of the mechanisms associated to each of the aforementioned planes, and identify a subset of options worthy of evaluation.

5.2.1.1 Routing Plane

Forwarding Policies

In a data collection or point-to-point routing protocols, some of the most widely cost metrics are the number of hops to the base station and the the number of expected transmissions [113]. Based on this, we propose to test the following forwarding policies upon reception of a response packet:

- Hop count-based (**hop count**): forward packets from nodes with a larger hop count.
- End-to-end PDR-based (**end-to-end PDR**): forward packets from nodes with smaller end-to-end PDR.
- Mixed policy (**mixed policy**): forward packets that observe any of the above conditions.

Congestion Mitigation

The issue of congestion in mobile networks has been thoroughly described in [114]. We implemented a contention mechanism at the forwarding module to mitigate congestion. We tested packet deferral for a period drawn from a uniform distribution. With this mechanism, packets are held back by the protocol a uniformly distributed random time before relay transmission.

5.2.1.2 Network Coding Plane

Generation Size and Payload Length

The generation size g and the packet payload l must be such that the total data transmitted per node matches the base station request per generation and node, v. However, the packet payload in a network coding protocol must shared by the application data and the coefficient vector. Regarding the coefficient vector storage, we use the on-demand strategy (see Section 2.3.3). To avoid the

coefficient vector to grow unrestricted and dominate the packet payload, we set a limit r_{coeff} on the payload fraction that the coefficient vector can use. All parameters are related by the formula $v = g \cdot l_{\max} \cdot (1 - r_{\text{coeff}})$, where $l_{\max}$ bytes is the maximum packet payload.

We evaluate two sets of generation size and useful payload length. The values are discussed in Section 5.3.1 as they depend on the application requested data.

Galois Field Size

The Galois Field (GF) size affects a number of operational aspects of the protocol, of which we address two. The first aspect is the probability of creating linearly dependent combinations in coding operations performed during transit to the sink, eventually leading to data blocks elimination. Protocols using smaller GF sizes have a larger probability of generating linearly dependent combinations during network transit [115] with respect to large GF sizes. The impact is on delay and efficiency – a small GF size will require more time and packets for sufficient d.o.f.s to arrive at the base station. The second aspect is that the GF also affects the available payload space per packet and thus, indirectly, the size of the generation. In this case, smaller GF sizes have a smaller footprint than larger sizes (although we do not test this implication).

We propose to evaluate the protocol performance in two Galois Field sizes: a small – GF(2) – and a large – GF(2^8).

Breadth of Coding

The breadth of coding in network coding mechanisms falls into two classes: systematic and non-systematic (the later also known as full-coding, a term we will favour) [116]. Systematic mode involves transmitting the initial requested n data blocks unencoded and subsequent transmissions as lin. ind. combinations of the original n data blocks, whereas full-coding mode involves that all m packets are sent encoded. The systematic mode serves mostly a reliability purpose, whereas full-coding allows capacity to be reached in suitable scenarios [117]. Tunable Sparse Network Coding (TNSC) [118] has been proposed as a dynamic solution that increases the coding density as time progresses. In [74], this strategy is facilitated by allowing intermediate nodes to decode.

We will focus on the two first alternatives, systematic and full-coding.

Coding Policy and Buffer Size

The coding policy concerns whether intermediate nodes are able to decode packets and produce novel lin. ind. combinations, or merely capable of recoding received packets. The first option provides more control over the lin. ind. combinations generated by intermediate nodes, at the cost of the delay necessary for those nodes to start decoding original data blocks [119]. The second option requires less resources and processing time at the nodes, but implies a looser control over the lin. ind. combinations created in the network. Several buffer management policies for network coding scenarios can be found in the literature. In [120] and [121] three designs for finite memory buffers are proposed: (i) a simple FIFO buffer; (ii) the accumulator design, in which received

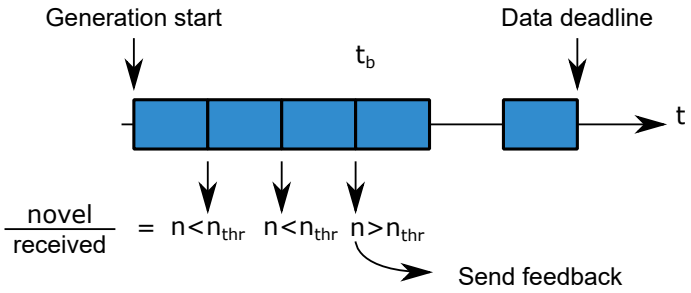


Figure 5.4: Operation of the Novelty Ratio mechanism.

packets are immediately randomly coded with all stored packets; and (iii) the recombinator design, in which new combinations are created using received and stored packets selected according to a uniform distribution. In [121] it is claimed that a single packet per node is sufficient for network coding to continue to perform optimally. The authors of [73] use the accumulator design and report that, from their experiments, the buffer size can be of just a few packets.

We test only the recoding policy and the accumulator design, and the buffer size is equal to the generation size per node.

5.2.1.3 Reliability Plane

Redundancy

For protocol designs using network coding, reliability mechanisms can be proactive or reactive [122]. In addition to the initial M packets, the source can send pro-actively $N > M$ packets to account for losses during transit. This solution requires some mechanism to compute expected losses (such as LQE/ETX information).

We propose to test two policies to compute the number of redundant lin. ind. combinations to be sent and inject them in the network:

- Hop-to-hop redundancy (**h2h red**): upon producing or receiving a response packet bound to be relayed, nodes produce a number of additional lin. ind. combinations inversely proportional to the minimum PDR of the links to its neighbours.
- End-to-end redundancy (**e2e red**): only the source creates redundancy lin. ind. combinations/packets in a quantity inv. proportional to the maximum end-to-end PDR.

Feedback

In a explicit feedback mechanism, the feedback request packet informs which data packets has the base station received, and nodes refrain from coding received packets in new transmissions [123]. In dense coding scenarios, the feedback information may be reduced to how many d.o.f.s are missing at the base station [124].

We developed a mechanism named *Novelty Ratio* to select the instants to send the feedback packets. A feedback request is sent each time the ratio of innovative packets over total received

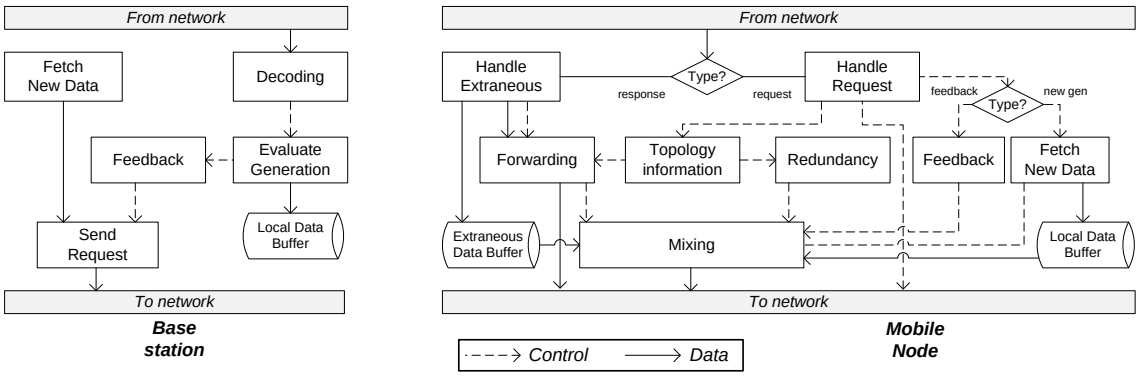


Figure 5.5: Internal architecture of sources and base station.

packets r_{nr} during a pre-defined time interval t_{nr} is lower than a some threshold. Figure 5.4 depicts the operation of this mechanism. We evaluate the performance of the protocol with and without the *Novelty Ratio* mechanism.

5.2.2 Protocol Specification

We describe the concrete implementation of the structural elements of the protocol and the modules that substantiate the mechanisms introduced in the previous section.

Two implementation architectures are discussed, one for the base station and other for the nodes. Each mechanisms can be implemented by one module (named after the mechanism) or multiple (which will be detailed), and can be distributed over base station and nodes or at just one of these. Two data structures support the operation of these mechanisms. Locally-produced data blocks are kept unencoded in the node, in a specific buffer called `local_data`. When a node receives an extraneous packet, it is stored in a buffer for extraneous packets, `extraneous_data`. It is assumed that a lower-layer LQE mechanism exists, allowing nodes to learn their neighbours and link quality to each. Figure 5.5 shows the internal architecture of base station and source, and the flow of control signals and data.

5.2.2.1 Routing Plane

Periodical beacon transmissions are sent by the module `send request` at the base station, at intervals t_{b}. Beacon/request packets carry hop count and accumulated PDR fields. From these, nodes can learn the neighbours with the smallest number of hops to the sink and the largest PDR to the sink, at module `topology information`. Nodes discard their previous neighbour list when a new beacon is received.

The `forwarding` module enforces the diverse forwarding policies discussed in Section 5.2.1.1. Response packets carry the last hop's latest hop count and end-to-end PDR to sink to support the diverse forwarding policies. The validity of the packet for forwarding is assessed with two tests: (i) whether lin. ind. combinations from previous generations are present in the payload, and (ii) if the packet has been seen previously. In the first case, packets with lin. ind. combinations from a previous generation must be discarded to avoid contamination of the current generation. In the

second case, each packet is assigned a unique identifier `uid` at production time. Nodes bookkeep received packets and do not relay packets already received as a mechanism to limit packet transmission.

5.2.2.2 Network Coding Plane

At the base station, the module `fetch new data` initiates a new generation every t_b. The broadcast beacons also act as requests for new N data blocks from each mobile node. At the sources, upon reception of new request packet by module `handle request`, the contents of `local_data` and `extraneous_data` buffer are discarded to eliminate response packets from the previous generation. The `fetch new data` module requests a new batch of N data blocks from the application and stores those in `local_data`. The `mixing` mechanism produces the N lin. ind. combinations to be sent. At the base station, received packets are stored in a buffer and Gaussian Elimination is performed to extract the encoded data blocks.

The `mixing` module performs the packet coding operations, and controls coding density and combination sources. This mechanism operates on request of the `handle request`, `redundancy` and `feedback` mechanisms. The mixing mechanism selects the sources from which combinations should be drawn, namely the `local_data` and `extraneous_data` buffers. Currently, all coded packets receive combinations from both buffers.

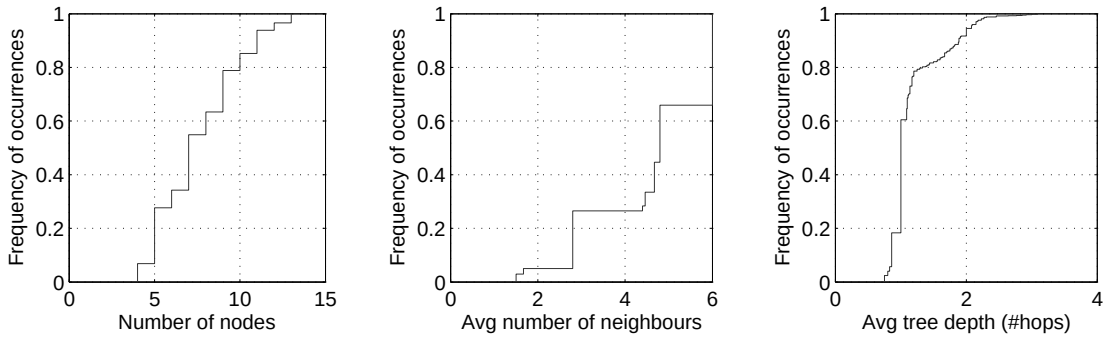
5.2.2.3 Reliability Plane

The `redundancy` module generates additional lin. ind. combinations to inject in the network. New lin. ind. combinations are created using packets sourced from `local_data` and `extraneous_data`. As discussed earlier, two policies are enforced: hop-to-hop redundancy (**h2h red**) and end-to-end redundancy (**e2e red**). After the initial request and during the generation interval, the base station sends feedback packets informing how many more d.o.f.s are necessary.

The `feedback` module implements the *Novelty Ratio* discussed in Section 5.2.1.3. The threshold ratio for which a feedback request is triggered is denoted by r_{nr}, and the interval over which the ratio is evaluation is t_{nr} seconds. To prevent excessive load on the network, nodes that have the required d.o.f.s only reply with a certain probability p_{fbr}.

5.3 Simulation Evaluation using Real-World Traces

In this section, we evaluate the protocol performance in various configurations and against CTP using simulation over connectivity traces sourced from a real-world testbed. We first describe the pre-processing performed on the traces and our simulation setup. We then present the design space exploration that we carried out regarding the protocol performance with various design options. We also evaluate the performance of a network coding protocol by comparison against CTP and evaluate if the premise laid out in Section 5.1 holds. Finally, a brief performance characterization with respect to topology features is presented, in Section 5.3.4.



(a) ECDF number nodes (b) ECDF avg number neighbours (c) ECDF average tree depth

Figure 5.6: Topology characterization from connectivity traces

5.3.1 Trace and Setup Description

Traces Pre-Processing

The connectivity traces used in this work were obtained from the container-carrying trucks at the Leixões harbour, in Porto, Portugal. These traces contain the PDR of the links of each node to their neighbours with a frequency of one second. We were provided a dataset of one week (from June 2 to 9, 2014) and extracted the data corresponding to June 3. After analysis of the dataset, the data from only three of the five RSUs was used.

For convenience, we opted to perform simulations for each RSU independently. For each RSU, from the total set of traces, we identified the seconds at which there is an active link to the target RSU (i.e. a link with a PDR larger than 0). For each RSU, we carried out a minimum spanning tree (MST)-search using Prim's algorithm for each second. From this result, we identified the nodes that participate in the MST, and isolated from the raw traces only the links between nodes of that subset. This information makes up a second of the input trace files that are fed to the simulation. In total, we obtained 5050 3-second generations for the selected day of traces for all three RSUs.

With the processed connectivity testbed traces, we computed selected metrics per generation (3 seconds-long) to evaluate the diversity of the wireless topologies occurring in the vehicular network, in addition to the analysis made in Section 5.1.1 regarding the distribution of the duration of optimal trees. Figure 5.6a shows us that the pool of topologies with 4 to 13 participant nodes is approximately evenly distributed. The CDF of the average number of neighbours features a few values that are more frequent more than the remainder, as seen in Figure 5.6b. The average tree depth, shown in Figure 5.6c, goes up to one hop in 60% of the cases, with the remainder not going above 3 hops.

Simulation Setup and Parameters

We use the OmNET++ simulator version 4.3.1 and the MiXiM Framework version 2.2.1. We developed our protocol in the C++ programming language and implemented it between the network and application layers. Random number generation was performed using OmNet++ in-built mechanism (Mersenne Twister), using the run number as seed. The ARQ mechanism of the MAC layer

of the IEEE 802.11 standard is used during the operation of CTP, with the maximum number of retransmissions for ARQ set to 4. The physical layer is a perfect packet erasure channel in which the packets are received or lost according to a pre-defined probability, the PDR extracted from the connectivity traces described previously. We use the IEEE 802.11b standard as data link/physical layer technology at the nodes. The simulation setup fixes the MAC layer bit rate to 2 Mbit/s, and the maximum payload size supported prior to fragmentation is 2348 bytes. We developed our own coding engine for GF(2^8) using logarithmic look-up tables. Our coding engine guarantees that each of n native packets is coded in at least one coded packet. We developed the decoding mechanism at the sink, that performs on-the-fly Gaussian Elimination as packets arrive. Presented results are averaged over 4 runs with different random generator seeds.

There are two application/scenario-driven parameters: the requested data volume per node v and the interval between beacons t_b. We define two data volume requirements to induce heavy and light load on the network, namely v = 10 kB/s and v = 3 kB/s. The value of t_b, as in Section 5.1, is also evaluated at two values: t_b =3 seconds and t_b = 6 seconds. The protocol parameters are discussed next. We define that the maximum payload size l_{max} of an higher layer datagram to be equal to the maximum payload size of IEEE 802.11b (thus no fragmentation is allowed). Given that we use an on-demand coefficient vector (i.e. of variable length), we opted to use fixed sizes for the payload fraction assigned to data, and the remainder of the packet payload can be fully used by the coefficient vector (thus implementing the r_{coeff} discussed in Section 5.2.1.2). The selected payloads sizes dedicated to coded data are of 1500 and 750 bytes, resulting in 20 and 40 packets/node respectively for v=10KB/s, and 6 and 12 packets/node for v=3KB/s. The parameters related to the *Novelty Ratio* feedback mechanism, p_{fbr}, r_{nr} and t_{nr}, were selected empirically. A list of all parameter values is shown in Table 5.2.

We consider as the main figure of merit to evaluate performance the overall packet delivery ratio p, the number of unencoded data blocks that were obtained by the sink (either received or decoded) over the data blocks that were generated at the sources (i.e., triggered by the reception of a request). To evaluate the effect of each single option at time, we defined a baseline configuration upon which we vary only the parameter or design option of interest. The underlined values in Table 5.2 correspond to the baseline configuration.

5.3.2 Design Space Exploration

We present the PDR of the various design- and parameter-space configurations of our protocol. A brief discussion follows to address a few operational aspects underlying some of the performance results. We evaluated the following configurations:

1. **Data Volume**
2. **Generation/Payload Size**
3. **Redundancy and Feedback** (reliability mechanisms)
4. **Congestion Mitigation**
5. **Forwarding Policy**

	Parameter	Value
Scenario/ Application	Data prod. bitrate v	3KB/s, 10 KB/s
	Beacon interval Δt_b	3s, 6s
Protocol	Forwarding policy	**hop count**, **end-to-end PDR**, **mixed policy**
	Redundancy policy	· **h2h red** (**hop count**) · **e2e red**(**end-to-end PDR**, **mixed pol.**)
	Feedback	Novelty Ratio
	Payload l / gen. size g	1500B/20, 750B/40
	Galois Field (2^q)	1, 8
	Coding breadth	systematic, full-coding
	Coding policy	re-coding
	Buffer size b	g
	Prob. fb. reply p_{fbr}	0.5
	Novelty interval t_{nr}	0.1 s
	Novelty threshold r_{nr}	0.2

Table 5.2: Parameter values (parameters for baseline underlined).

6. **Redundancy Injection Strategies**
7. **Galois Field Size**
8. **Coding Breadth**

We leave the analysis of the impact of the inter-beacon interval t_b for Section 5.3.3. The full range of results can be found in Appendix B, including results for energy efficiency (ECDF of decoded packets over total number of packets transmitted).

5.3.2.1 Performance Results

• **Data Volume:** The performance of the network coding protocol showed considerable difference in performance as the data volume requested from the nodes is varied. In Figure 5.7, we observe the difference in performance as we increase the requested volume from 3KB/s to 10KB/s, causing additional load on the network.

• **Redundancy and Feedback:** The impact of the reliability mechanisms – redundancy and feedback – proposed in Section 5.2.1.3 are shown in Figure 5.8. We observe that both mechanisms resulted in considerable improvement of PDR. The NC performance improved from around 50% of generations receiving half of the expected packets, to 60% of generations receiving the entire generation. The impact is particularly visible in the full-coding mode (see Appendix Figure B.27), as without any reliability it exhibits very poor performance due to an insufficient number of d.o.f.s received at the sink (discussed in more detail in the next subsection).

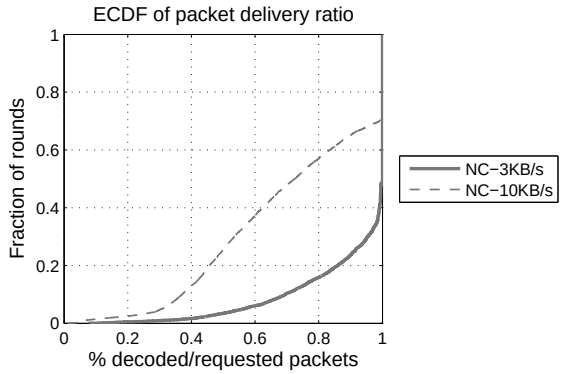


Figure 5.7: PDR of different data volumes (light and heavy load).

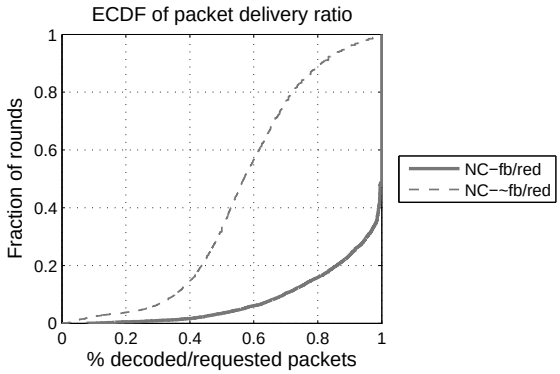


Figure 5.8: PDR with and without reliability mechanisms.

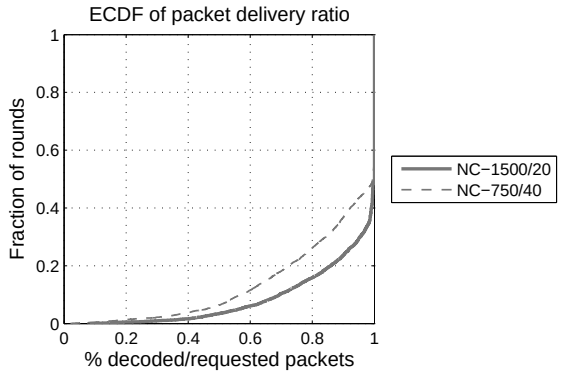


Figure 5.9: PDR with different generation/payload sizes.

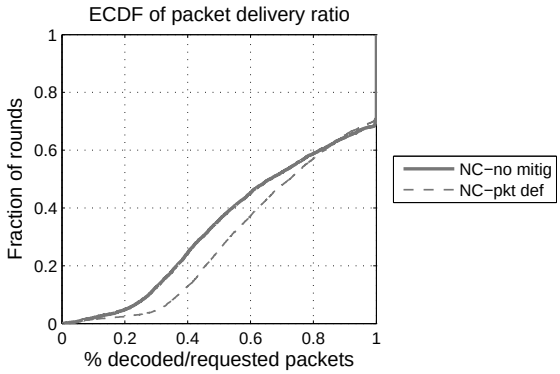


Figure 5.10: PDR with and without congestion mitigation under heavy load (10KB/s).

• **Generation/Payload Size:** Figure 5.9 presents the results for the two sets of parameter values for the generation and payload size. We observe that smaller payloads/larger generations yields, in the majority of the generations, a worse PDR performance than the complementary design solution. We hypothesize that two phenomena concur to this result: (a) there is an higher overhead involved with medium access and packet processing at nodes; and (b) coding over a larger coefficient pool promotes more inter-packet dependency when decoding at the base station.

• **Congestion Mitigation:** Inspection of the protocol performance under large data volumes at an early design stage showed a large number of collisions. For this reason, we plot the PDR of the protocol with and without a congestion mitigation mechanism for a large requested data volume per node (10KB/s) in Figure 5.10 (see Appendix Figure B.11 for lower requested volumes).

• **Forwarding Policy:** The performance profile of the three forwarding policies discussed in Section 5.2.1.1 – **hop count**, **end-to-end PDR** and **mixed policy**– is depicted in Figure 5.11. The **hop count** policy fares the worst of the three, and the **end-to-end PDR** policy provides slight overall improvement over **hop count**.

• **Redundancy Injection Strategies:** Figure 5.12 presents the results of the two redundancy policies, **e2e red** and **h2h red**. We observe that the **e2e red** (transmitting an additional number of

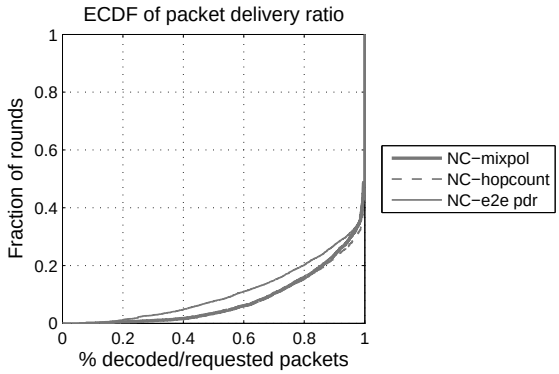


Figure 5.11: PDR for different forwarding policies.

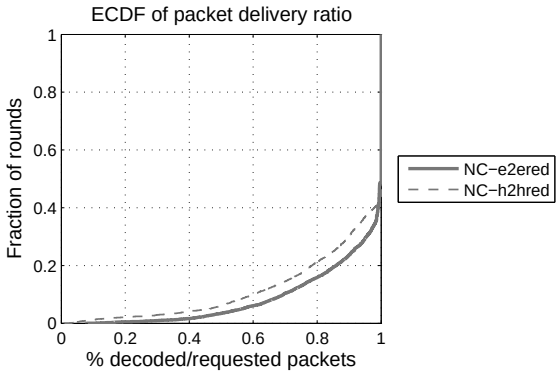


Figure 5.12: PDR for different redundancy injection policies.

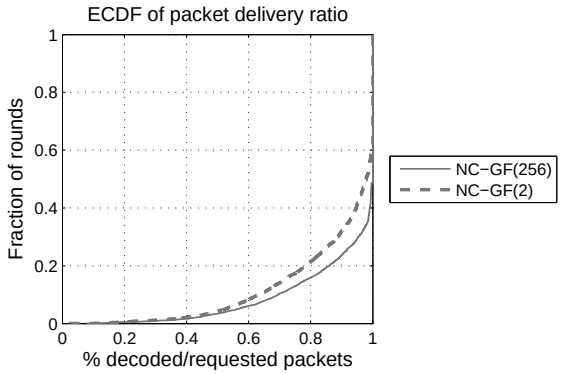


Figure 5.13: PDR for different Galois Field sizes.

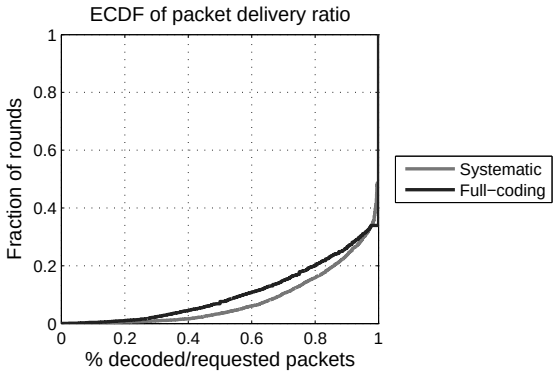


Figure 5.14: PDR for different coding breadths.

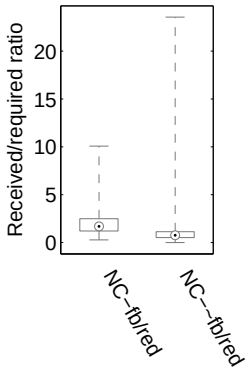
packets at the sources) outperforms a policy of creating new lin. ind. combinations at intermediate nodes (**h2h red**).

• **Galois Field size:** The two Galois Field sizes evaluated affected the protocol performance in the way shown in Figure 5.13. As discussed in Section 5.2.1.2, smaller field sizes can potentially lead to a larger probability of d.o.f.s disappearing during transit. The difference proved not to be significant.

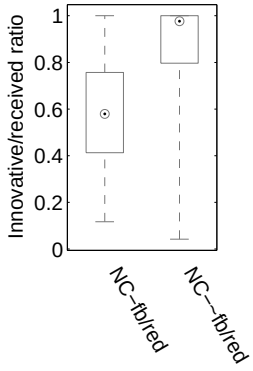
• **Breadth of coding:** The two modes of operation – systematic and full-coding – proved to have a marginal difference in performance at smaller requested volumes, as visible in Figure 5.14. The difference in performance is considerably larger as the requested volume increases, as seen in Appendix Figure B.25. As in the case of the reliability mechanisms, this is due to a higher inter-packet dependency, as we will discuss in the next subsection.

5.3.2.2 Discussion on Performance

We discuss two particular operational aspects of the previous analysis relating to network coding and traffic load.

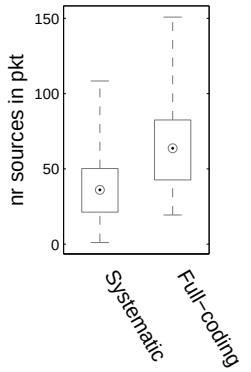


(a) Ratio of received to requested.

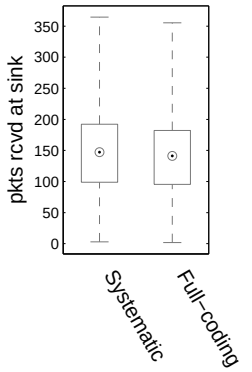


(b) Ratio of innovative to received.

Figure 5.15: Coding-associated metrics for reliability mechanisms.



(a) Number of sources per packet.

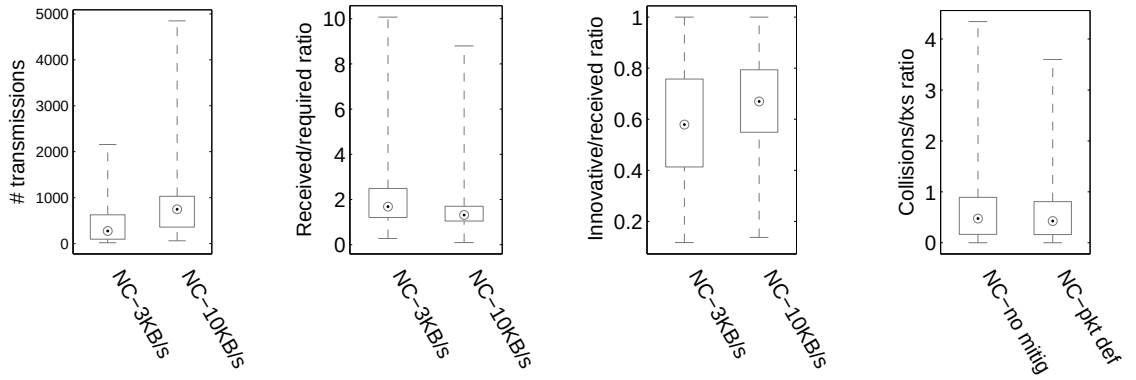


(b) Number of packets received at sink.

Figure 5.16: Coding-associated metrics for coding breadths under heavy load (10KB/s).

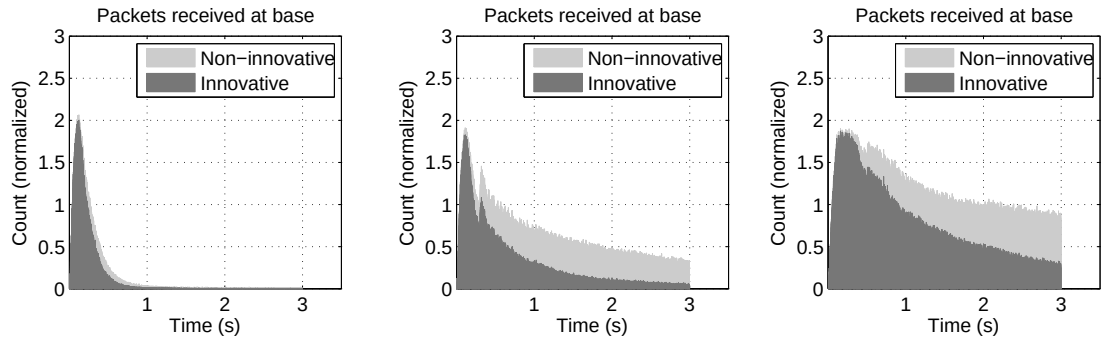
It is relevant to note that, in a network coding protocol, even if enough packets arrive to the base station, not all packets are useful for decoding the generation. We evaluate the impact of the coding mechanisms in the protocol performance by analyzing the ratio of received-to-requested and innovative-vs-received packets at the base station. The first ratio is affected by routing and reliability design options, whereas the second ratio is more dependent on network coding options. In Figure 5.15, we present the ratios for the protocol configurations with and without reliability mechanisms (feedback and redundancy). It can be seen that, despite the non-reliability configuration exhibiting high ratios of innovative packets (Figure 5.15a), there are not enough packets arriving to the base station to support a better PDR (Figure 5.15b). A similar behaviour is observed when we vary the generation and payload sizes towards larger generations and smaller payloads. We also observed that, as the requested data volume is increased, the selected breadth of coding – systematic and full-coding modes – begins to play a non-negligible effect on these ratios. This is due to the higher occurrence of inter-packet dependency that full-coding promotes. In Figure 5.16, we observe an higher number of sources per packet for the full-coding mode (Figure 5.16a) that it is not accompanied by an increase in the number of packets received at the base station (Figure 5.16b).

The load exerted in the network is affected by multiple options, chief amongst which the data volume requested from the nodes. In Figure 5.17a we observe the increase in the total number of transmissions averaged over all generations as the requested data volume per node goes from 3KB/s to 10KB/s. As in the previous analysis regarding the impact of network coding options, we still identify a network load condition at higher data volumes by noticing the inferior ratio of received-to-requested packets (Figure 5.17b) and the higher ratio of innovative-vs-received (Figure 5.17c). The impact of the congestion mitigation mechanism is shown in Figure 5.17d for the higher data volumes, without which the previous metrics would be even worse. The phenomenon of load being exerted on the network can be also understood from a temporal perspective, by depicting at the temporal profile of the reception of the packets at the base station. In Figures 5.18a



(a) Total number of transmissions (b) Ratio of received to requested. (c) Ratio of innovative to received. (d) Ratio of collision to transmits for large d.v..

Figure 5.17: Operational metrics for different data volumes.



(a) Without feedback/redundancy. (b) With feedback/redundancy. (c) Large data vol. (10KB/s).

Figure 5.18: Temporal profile of packet arrival at base station for various configurations.

and 5.18b, we compare the network traffic over time from the temporal distribution of packet reception at the base station for a configuration without and with reliability mechanisms respectively. The inclusion of end-to-end reliability spreads packet transmission throughout the entire generation duration. Finally, in Figure 5.18c, the impact of a larger data volume is observed, with a large ratio of innovative packets arriving still at the end of the generation interval (t_b =3s).

5.3.3 Benchmark and Route Lifetime Analysis

5.3.3.1 Benchmark Against CTP

We now compare the performance of the baseline configuration of the network coding protocol against a benchmark protocol, the Collection Tree Protocol. Figure 5.19 presents the packet delivery ratio for the baseline protocol configuration and CTP. We observe that the CTP and the network coding protocol have very distinct performance profiles, with the network coding protocol being consistently better than CTP. The downside is the smaller efficiency of the network coding protocols, observable in the ECDF of energy efficiency in Figure 5.20.

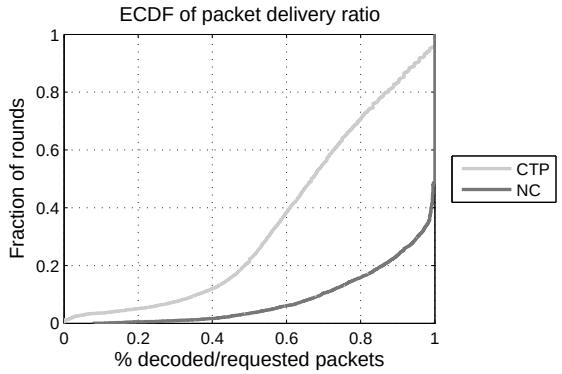


Figure 5.19: PDR comparison of CTP and NC for t_b =3 seconds

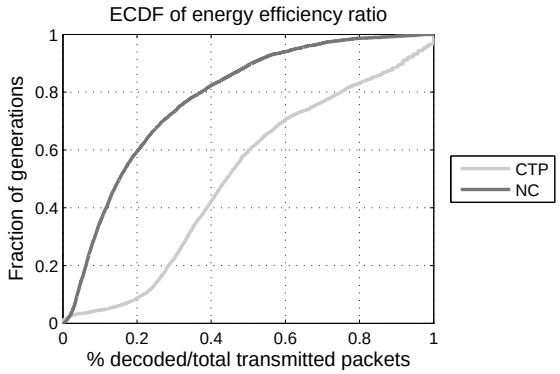


Figure 5.20: Energy efficiency comparison of CTP and NC for t_b =3 seconds

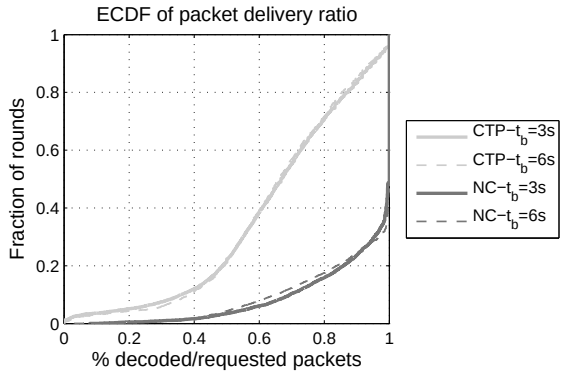


Figure 5.21: PDR comparison of CTP and NC for different t_b under small requested volumes

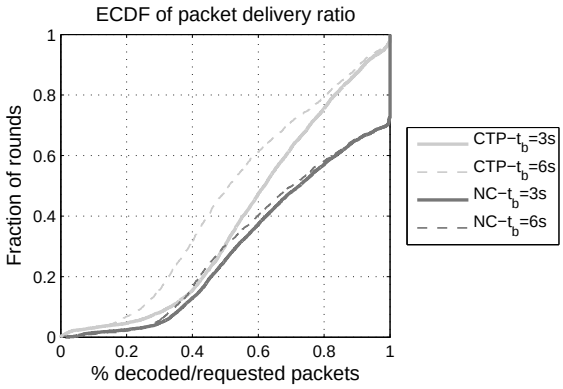


Figure 5.22: PDR comparison of CTP and NC for different t_b under large requested volumes.

5.3.3.2 Route Lifetime

Finally, we evaluate the initial premise or our work: the resilience of network coding protocol to changes in topology occur is better than that of CTP, as t_b increases. Figure 5.21 presents the PDR for the baseline configuration of the network coding protocol and CTP for the two values of t_b and a small requested data volume. We observe that CTP does not suffer significantly with the increase in t_b. It is only when a heavier load is requested that a visible impact can be observed, as seen in Figure 5.22. The packet delivery ratio of CTP degrades as t_b increases from 3 to 6 seconds, while the network coding protocol incurs in a smaller performance penalty. In conclusion, we observe that the performance of the network coding protocol hinges on a trade-off between the requested data volume and the inter-beacon interval.

5.3.4 Impact of Topology Characteristics

We also studied the relationship between topological characteristics and packet delivery ratio, as some topologies may be more prone to induce additional load than others. Figure 5.23 shows the PDR of the network coding baseline configuration with respect to a topology metric computed per

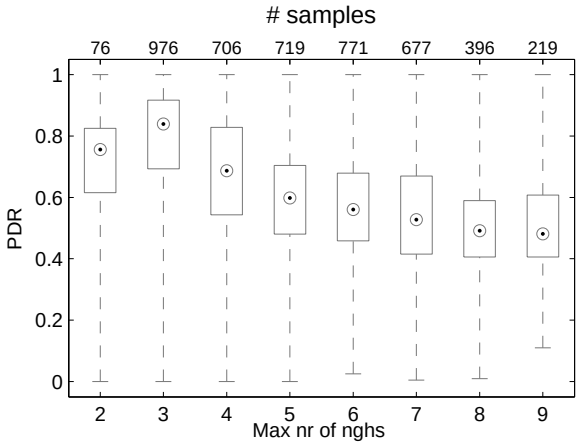


Figure 5.23: PDR for different topology types.

generation (specifically t_b =3 seconds), namely the maximum number of neighbours of all nodes. It can be seen that the performance degrades as the metric value increases, and we are currently investigating if denser topologies exert excess traffic load in the network.

5.4 Final Remarks

In this chapter we addressed the challenge of designing higher layer protocols, namely routing and transport, for the purpose of data collection in M/VANET scenarios. Given the limitations of MST-based protocols such as CTP in base station-centric data collection applications in scenarios with mobility, we developed a protocol based on opportunistic forwarding and network coding for this task. A thorough exploration of the design space and a performance comparison against a benchmark protocol, using simulations over real-world connectivity traces, is carried out.

Our results support the following conclusions. CTP suffers a considerable toll in packet delivery ratio as larger inter-beacon intervals are used. As for our design space exploration, our results show that:

- small performance variation among different forwarding policies;
- performance is similar whether the forwarding decision is based on larger PDR or smaller number of hops to base station;
- the use of packet deferral alleviates congestion induced by a broadcast protocol;
- large packets and small generations perform best;
- a large Galois Field size fares slightly better;
- redundancy and feedback mechanisms are essential for a competitive performance;
- redundancy packets should be injected at sources.

Comparing to CTP, performance varies considerably with respect to an application-driven parameter: the requested data volume per node. Finally, a smaller inter-beacon interval t_b affects negligibly the packet deliver ratio of the network coding protocol, therefore outperforming CTP as initially hypothesized.

Ongoing work focus on understanding the different performance profiles with respect to topology aspects (e.g. dense, highly-connected topologies versus sparse, small topologies; rate of variation among topologies over time; degree of connectivity of the base station; etc). The study of local and global strategies to find the optimal trade-off between relaying ratios, redundancy packet creation and paths to explore is also being explored, although it can be foreseen that the complexity of the protocol may increase.

Chapter 6

Conclusions

6.1 Contributions

In this thesis, we developed a range of solutions for data collection over ad hoc networks with mobile and static nodes. Our contributions fall in the scope of network design for this application, which we further broke down into three aspects: (i) scenario characterization; (ii) infrastructure planning; and (iii) network operation.

We contributed to improve accurate scenario characterization on the particular field of propagation modelling for device-to-device channels by proposing better methodologies to collect and process experimental data. In the process of collecting RSSI-distance measurement pairs between mobile devices or vehicles, GPS is one of the most widely used technologies to obtain position estimates, but it is subject to numerous sources of errors that affect its position estimates. We carried out field measurements to collect exact and erroneous position estimates, and evaluate the difference in the estimated model parameters if one or other dataset are used. We formulate a model for the error of distances extracted from GPS position estimates. From this study, we propose best-practice guidelines for measurement campaigns, specifically that measurements should be taken at distance higher than $\sqrt{2}\sigma_{\text{GPS}}$. We also proposed an *a posteriori* parameter correction strategy. This work was carried out in collaboration with Dr. Traian E. Abrudan and results have been published at the IEEE Transactions on Wireless Communications [6].

In the perspective of infrastructural planning, the use of measurement data and datasets from the scenario can improve the placement of intrafructural nodes and save resources. We addressed the problem of placing road-side communication hubs that serve nearby sensor nodes by sending their data to the cloud via urban wireless backhauls. In this problem, a number of practical, end-system and communication constraints must be observed while seeking the goal of minimizing the necessary resources. A particular aspect of this process is estimating data transfers from the road-side nodes to vehicular nodes, for which we contributed with a novel approach for city-wide estimation of data transfers in I2V links. This model builds on an experimental characterization of the links between mobile and static terminals. We undertook a campaign to understand I2V links between a road-side node and a large-scale vehicular network, and how local features of a

particular location (bus stops, traffic lights) influence the measured values. The solutions produced by our proposed decision support framework for placement of communication hubs were validated against an actual platform. Regarding service by infrastructural backhaul, close to 60% of the deployed DCUs are located up to 100 meters of a framework placement, and 87.5% had good or sufficient WiFi service. For vehicular backhaul-served DCUs, our model for I2V data transfers estimation can predict/estimate the actual data volumes up to one order and a half of magnitude. The initial characterization experiments at a prototype DCU were published at ACM MobiCom 2015 Workshop on Challenged Networks [7], and the subsequent placement strategy has been through a first round of reviews in a submission to ACM Transactions on Sensor Networks [8].

Our work on the third facet of network design, network operation, addresses the development and test of protocols for gathering data from an ad hoc network of vehicular and static nodes at an infrastructural base station. Existing base station-driven collection protocols harness beaconing to set up a minimum spanning routing tree. In a dynamic network, this strategy is subject to degradation of the routing information at the nodes as the time since the last beacon passes. Our contribution is the evaluation of an opportunistic forwarding protocol paired with network coding strategies and identification of the application and design conditions in which it may outperform existing protocols. We developed a testbed to carry out an extensive exploration of the design space of the framework protocol over connectivity traces obtained from a real-world scenario. Our results support a number of practical insights regarding the design of network coding protocols, such as: performance may vary with load on network, which in turn depends on requested data volume and other mechanisms such as feedback; feedback and reliability are crucial for competitive delivery rates; and redundancy should be injected at the source nodes instead of being performed during transit to the base station. Comparing against a state-of-the-art beaconing-based structured protocol (CTP), the network coding protocol fared better against a longer period between beacons in terms of PDR. This work has been carried out in collaboration with Professor Daniel E. Lucani and a publication is being completed.

In addition to the core contributions of this thesis, some of the work carried out during this thesis was developed within the scope of several projects, from where a number of additional publications resulted but did not fit into this thesis.

- I co-authored an article describing the sensing platform UrbanSense alongside Yunior Luis, Tiago Lourenço, Carlos Pérez-Penichet, Tânia Calcçada and Ana Aguiar, and that was published in the IEEE flagship conference on smart cities, IEEE International Smart Cities Conference [102], and that won the Best Student Paper award.
- Starting with M.Sc. student André Sá and terminating with M.Sc. student Diogo Guimarães, alongside Tiago Condeixa from VENIAM and Prof. Dr. Susana Sargento from the University of Aveiro, we developed a data collection solution based on delay tolerant networking that was showcased as a Demo at the ACM MobiCom 2015, in Paris [125].
- A high-level description of the PortoLivingLab [126] was co-authored by me and João Rodrigues, Susana Cruz, Tiago Lourenço, Pedro M. D'Orey, Yunior Luis, Susana Sargento,

Ana Aguiar and João Barros, and is currently under review at the IEEE Internet of Things Journal.

- I also co-authored an article describing the work carried out by M.Sc. Leonid Kholkine, in collaboration with André Cardote from VENIAM and Ana Aguiar, regarding a solution to the problem of undesired WiFi connections from smartphones to the vehicular access points of *BusNet* and that was published in IEEE Vehicular Networking Conference 2016 [127].
- With M.Sc. student Fábio Cunha, we started a prediction model of the transferable data volumes in I2V contacts from location features such as the number of stopping opportunities (zebra crossings, traffic lights, bus stops) and range and number of lines-of-sight.
- Finally, the work of Chapters 3 and 5 of this thesis was carried out within the context of the Vital Responder project. In this context, I developed an algorithm for dynamic building evacuation system for indoor scenarios and built a demonstrator of this application [128] that was showcased to the Minister of Science Dr. José Mariano Gago. An advanced version of this system and a scheme for fast on-the-fly deployment was accepted at a conference with Luís Pinto, Sérgio Crisóstomo, Traian E. Abrudan, and João Barros [129].

6.2 Limitations

There are a number of limitations that may apply to the contributions of this thesis and reduce their scope of application.

In Chapter 3, the impact of GPS position estimates in parameter estimation can be alleviated if the GPS receiver is capable of using historical position estimates (i.e. taken over some time up to the current instant) for position enhancement. This is not considered within the scope of our GPS error model, but this behaviour could be modelled by introducing a temporal dimension to the standard deviation of the GPS error affecting position estimates, σ_{GPS}. While the overall contribution of the work remains, the impact of the position estimates may be diminished as time progresses up to a point that it no longer causes a noticeable impact.

The work about communication hub placement, presented in Chapter 4, uses a generic deployment range to abstract from a particular technology for the link between sensor unit and communication hub. In our particular application scenario, increasing this range is accompanied by a growth in the number of served sensors, but in practice a subset of locations may not be able to support such distances. This issue can be addressed by introducing different deployment ranges and costs that apply within areas with an adequate resolution. For example, considering the micro-cell as the atomic spatial element, this strategy would imply that each micro-cell would be assigned a deployment cost for each one of the pre-placed sensors that are located up to a maximum range. Defining a maximum range per micro-cell will require incorporating a description of the cityscape (or any other target environment) in our framework.

Finally, the performance of the structured (MST-based) and opportunistic protocols in Chapter 5 can vary considerably as the scale of the target scenario varies, particularly as the number

of nodes grows substantially. In the first case, routing information inconsistency is more prone to occur as the number of nodes and hops increases, as the chances of link break up also increase. In the second case, given the opportunistic nature of the protocol, congestion mitigation strategies will play an important role to avoid saturation. A possible solution is to increase the number of road-side units, in order to minimize the number of hops from any node to a base station. An alternative protocol design is to eliminate the route setup phase (in our case provided by the beaconing mechanism) and apply geographical routing for packet forwarding, e.g. that of [61]. However, geographical routing has been shown to underperform in some conditions [72].

6.3 Future Work

There are lines of potential evolution in each topic of this thesis.

Regarding the work of Chapter 3, the improvement of position estimates, particularly from GPS data or other sources, is a thriving research area at this point. Substantial research is going into the problem of mitigating GPS errors, either by design or post-processing on collected data solution. It could be evaluated if the knowledge of the actual distances between multiple inexpensive GPS terminals can be used to improve the location estimates from the terminals, following the approach of [130]. Another research line may seek to understand the impact of erroneous path loss models on ranging strategies.

The placement procedure for road-side client nodes discussed in Chapter 4 can be improved by including more accurate datasets and models. There are currently in the literature some relevant approaches to large-scale modelling of wireless channels and estimation of propagation parameters, for example from crowdsensed datasets [35, 11]. It would be an exciting challenge to improve the accuracy of our procedure for city-scale estimation of I2V data volumes and compare the results against the measurements taken by the UrbanSense platform. Also open for further research are the validation of the number of connections as best predictor of I2V transferred data volumes, and a more thorough study of association times in I2V connections.

The work described in Section 5, the extensive design space of a opportunistic forwarding/network coding protocol, has shown us the wide range performance profiles that can be expected from such protocol design. There is room to discuss if the additional complexity of a network coding protocol is worth with respect to simpler protocols. Some recent works have addressed the performance of network coding in time-varying networks [131], that may add to this discussion. The co-existence of multiple RSUs poses challenges both for structured and opportunistic protocols about which routing tree or information should each node use.

As a global future work, we will strive to achieve a full vertical integration of all three aspects of network design, as depicted in Figure 1.4. In this thesis, the aspect of network operation, explored in the third part (*Data Collection Protocols*), relied on simulation over connectivity traces. In the future, we aim to carry out protocol simulation over mobility traces (as these are more general than connectivity traces), incorporate the accurate channel models obtained with our methodology, and simulate I2V collection service from our framework-recommended placements.

Appendix A

Reduction and Proof of Min-Hub Problem

Reduction Procedure of Set Cover into Min-Hub Problem

We provide a polynomially time-bounded procedure to reduce the set cover problem to the Min-Hub problem, and prove the validity of the procedure.

Let us first revisit the Min-Hub and set cover problems.

– **Min-Hub problem:** A set of communication hubs is to be placed to serve a set of pre-placed sensor units S. The hubs can serve an arbitrary number of sensor units, as long as hubs are within r_d meters of the sensor units. Each potential hub can be uniquely identified by the set of sensor units that share that hub. We refer to D as the collection of distinct sets of sensor units that can share a hub, and we seek as a minimal collection D' of sets of D that covers all elements in S.

– **Set cover problem:** Given a universe of elements U and a collection C of subsets of U, we seek a subset C' of C that contains all elements in U and is minimal in size.

A strategy to reduce the set cover problem to the Min-Hub problem is now presented.

Procedure for reduction:

Let U=S. Construct C as follows:

1. Set all elements of U (sensor units in S) as elements of C (i.e. we allow hubs to serve a single sensor unit);

2. Determine all connected components of sensors distanced to any other sensor by $2r_d$ or less. For each connected component, find the subsets q in which all elements are within less than $2r_d$ from all other elements in the same subset. Add subsets q to C.

3. Repeat step 2 over the subsets q output by the previous step, until no output subset has more than 2 elements. ■

Proof of NP-hardness of Min-Hub Problem

We now provide a proof of the NP-hardness of the Min-Hub problem. The proof can be broken down into two steps: (1) showing that the set cover problem is NP-hard; (2) proving that the set cover problem can be reduced to the Min-Hub problem. Regarding the first point, the set cover problem was proven to be NP-complete by Karp [106]. We proceed to address the second point.

The premise to be proven is that, if and only if the Min-Hub Problem instance we created is a 'yes' instance, the original instance of the set cover problem is a 'yes' instance. The following proof is loosely based on Section 8.1 of Kleinberg and Tardos [132].

Proof:

We start by constructing an instance of the Min-Hub problem from an instance of the set cover problem. Let $U = S$. We create a collection D of sets of S as follows: label sets of C from 1 to n, and assign all sets $C_i \in C$ as sets $D_i \in D$. Note that $D_i \subseteq S$ for all i. This construction can be made in polynomial time.

Assuming we have a black box for the Min-Hub problem, the same black box can be run for this Min-Hub problem instance created from an instance of a set cover problem. We need to show that, if the created Min-Hub problem instance is a yes instance, then the original set cover problem instance is also a yes instance. We proceed to do so.

– Let C' be a collection of sets that constitutes a set cover for U. From our construction, C' corresponds to a collection D' of groups of sensor units. We claim that the sets listed in D' cover S. Note that any element u of U is an element s in S. Given that C' is a set cover for U, it follows that all elements u must be contained in C'. Thus, knowing that C' corresponds to a collection D' of subsets of S, D' contains all $s \in S$.

– Now, let us assume that a hub assignment D' that associates hubs and sensor units minimally exists. Since each set in C' is naturally associated with a set in D, let C' be the collection of these sets. Thus, $|D'|=|C'|$. We claim that at least one set in D' contains u, for any u. By construction, $U=S$ and any element u is an element s. The elements of D' are individual s or sets of s, and thus correspond to u or sets of u. Thus, D' must contain at least one instance of each u, for any u.

■

Appendix B

Performance of Network Coding Protocol over Design Space

Baseline configuration

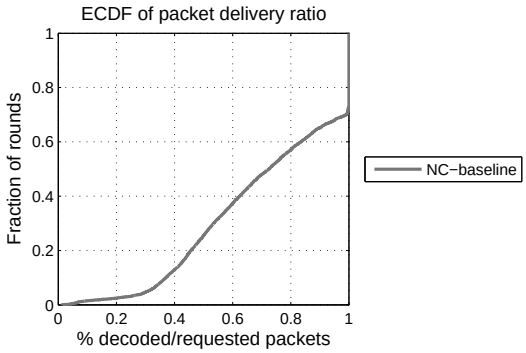


Figure B.1: Packet delivery ratio baseline configuration for large requested data volume (10KB/s/node).

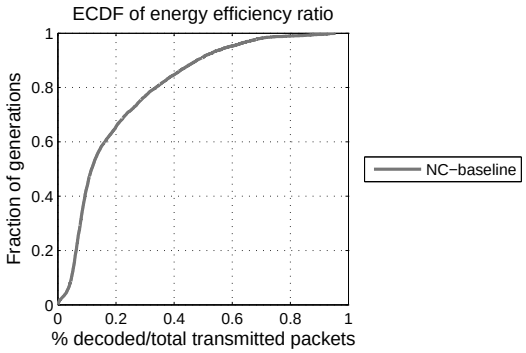


Figure B.2: Energy efficiency ratio baseline configuration for large requested data volume (10KB/s/node).

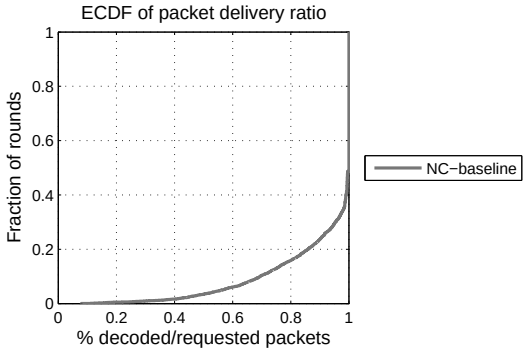


Figure B.3: Packet delivery ratio baseline configuration for small requested data volume (3KB/s/node).

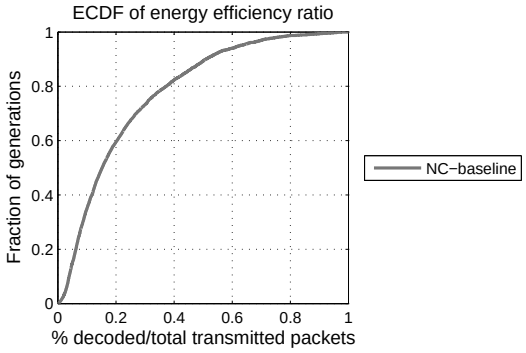


Figure B.4: Energy efficiency ratio baseline configuration for small requested data volume (3KB/s/node).

Forwarding Policies

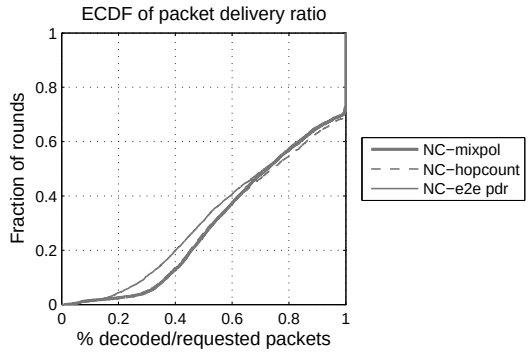


Figure B.5: Packet delivery ratio with different forwarding policies for large requested data volume (10KB/s/node).

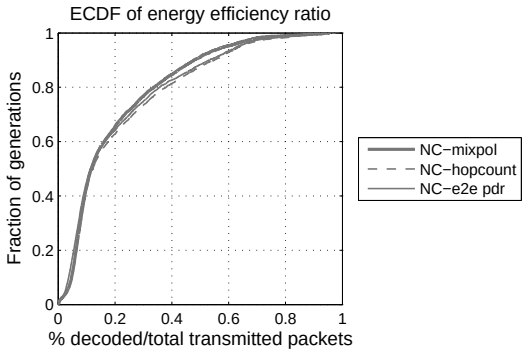


Figure B.6: Energy efficiency ratio with different forwarding policies for large requested data volume (10KB/s/node).

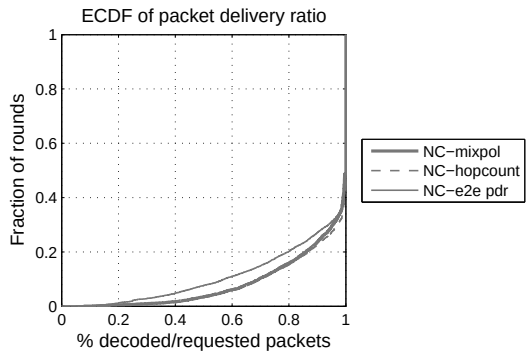


Figure B.7: Packet delivery ratio with different forwarding policies for small requested data volume (3KB/s/node).

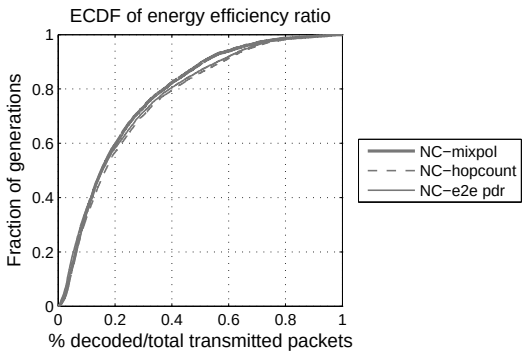


Figure B.8: Energy efficiency ratio with different forwarding policies c for small requested data volume (3KB/s/node).

Congestion mitigation

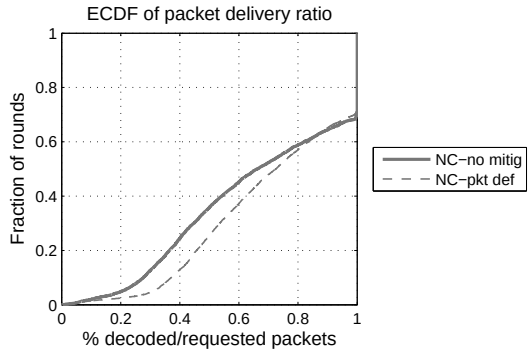


Figure B.9: Packet delivery ratio with and without congestion mitigation for large requested data volume (10KB/s/node).

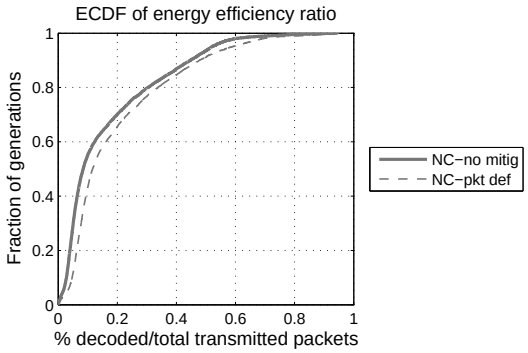


Figure B.10: Energy efficiency ratio with and without congestion mitigation for large requested data volume (10KB/s/node).

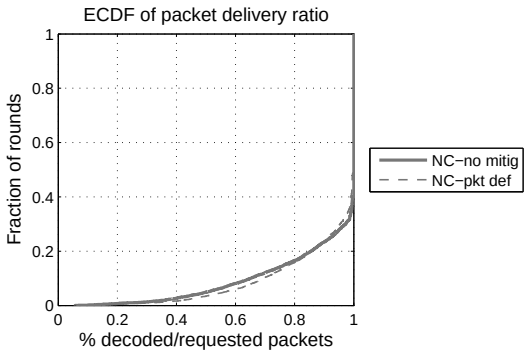


Figure B.11: Packet delivery ratio with and without congestion mitigation for small requested data volume (3KB/s/node).

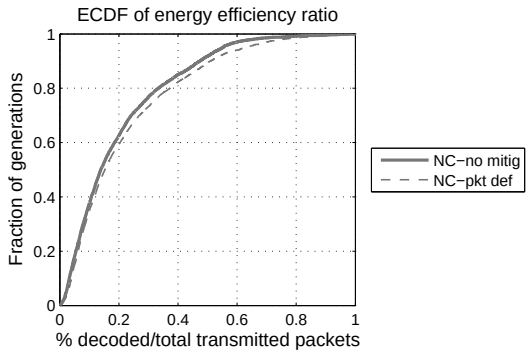


Figure B.12: Energy efficiency ratio with and without congestion mitigation for small requested data volume (3KB/s/node).

Reliability mechanisms

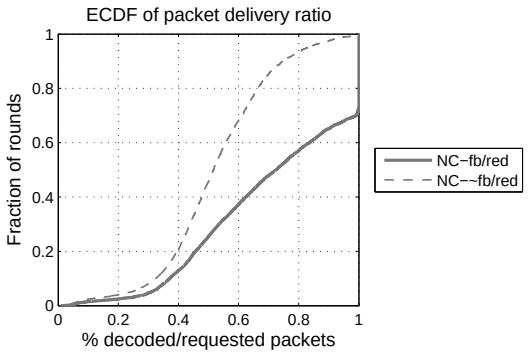


Figure B.13: Packet delivery ratio with and without reliability mechanisms for large requested data volume (10KB/s/node).

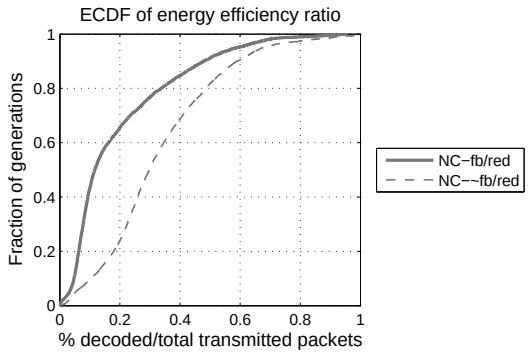


Figure B.14: Energy efficiency ratio with and without reliability mechanisms for large requested data volume (10KB/s/node).

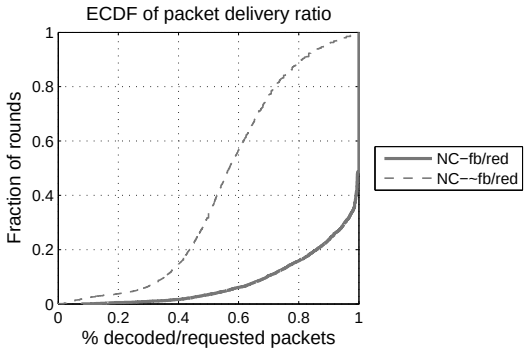


Figure B.15: Packet delivery ratio with and without reliability mechanisms for small requested data volume (3KB/s/node).

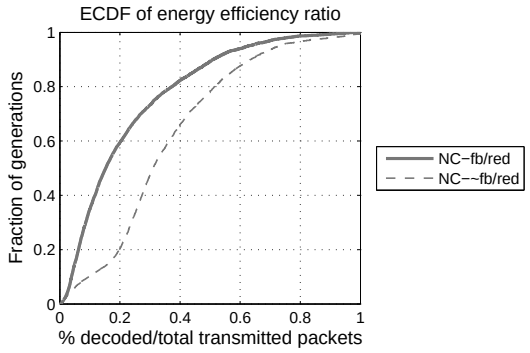


Figure B.16: Energy efficiency ratio with and without reliability mechanisms for small requested data volume (3KB/s/node).

Redundancy injection policies

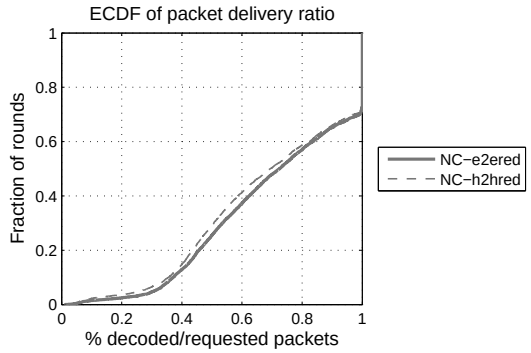


Figure B.17: Packet delivery ratio with different redundancy injection mechanisms for large requested data volume (10KB/s/node).

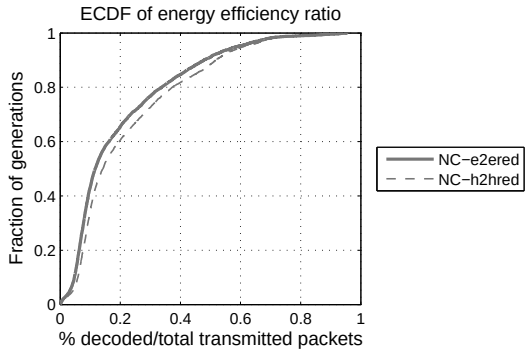


Figure B.18: Energy efficiency ratio with different redundancy injection mechanisms for large requested data volume (10KB/s/node).

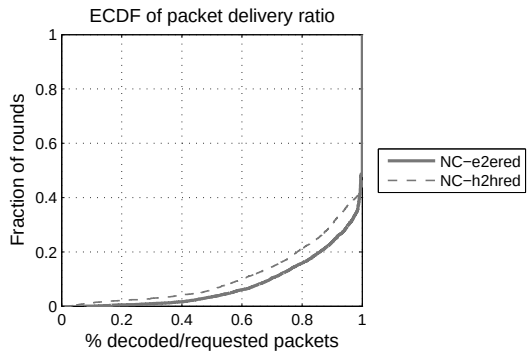


Figure B.19: Packet delivery ratio with different redundancy injection mechanisms for small requested data volume (3KB/s/node).

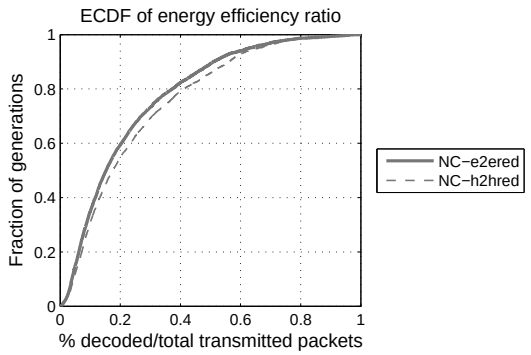


Figure B.20: Energy efficiency ratio with different redundancy injection mechanisms for small requested data volume (3KB/s/node).

Galois Field size

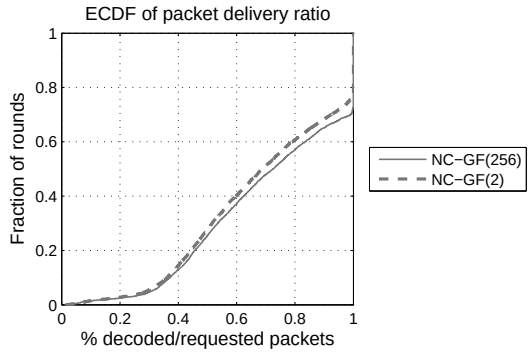


Figure B.21: Packet delivery ratio with different Galois Field sizes for large requested data volume (10KB/s/node).

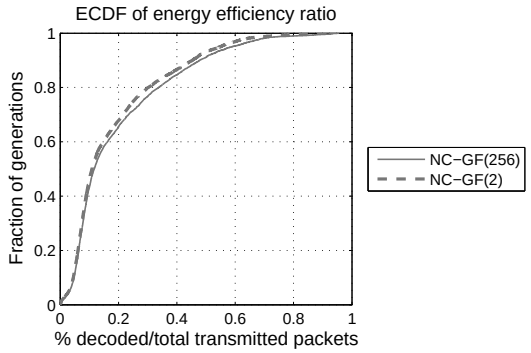


Figure B.22: Energy efficiency ratio with different Galois Field sizes for large requested data volume (10KB/s/node).

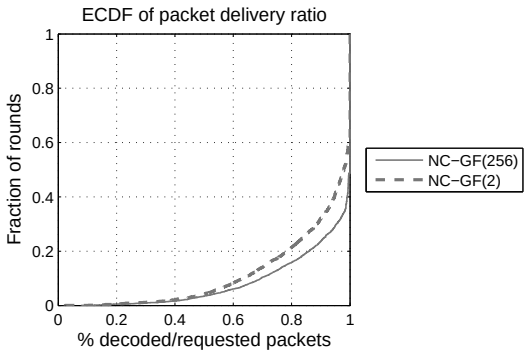


Figure B.23: Packet delivery ratio with different Galois Field sizes for small requested data volume (3KB/s/node).

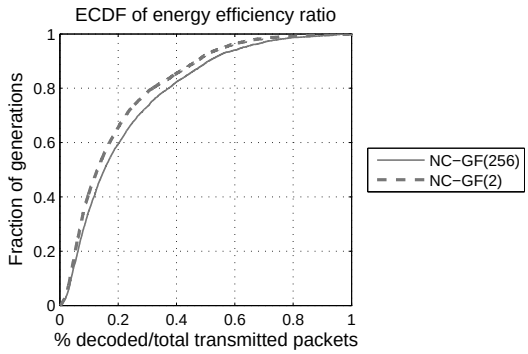


Figure B.24: Energy efficiency ratio with different Galois Field sizes for small requested data volume (3KB/s/node).

Coding Breadth

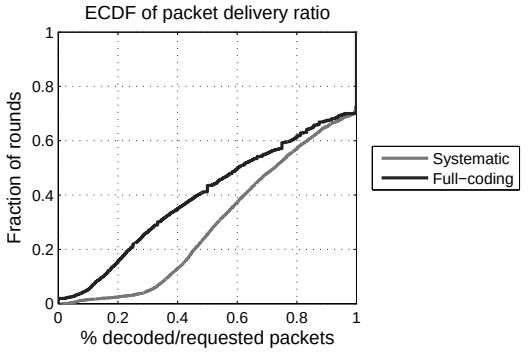


Figure B.25: Packet delivery ratio with different coding breadths for large requested data volume (10KB/s/node).

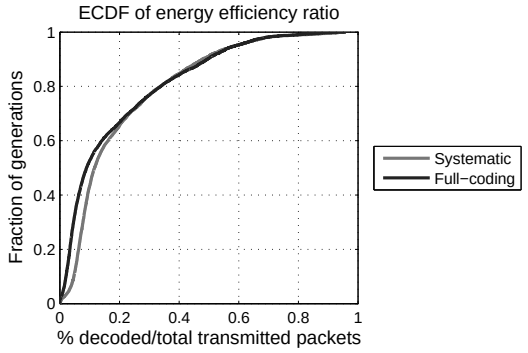


Figure B.26: Energy efficiency ratio with different coding breadths for large requested data volume (10KB/s/node).

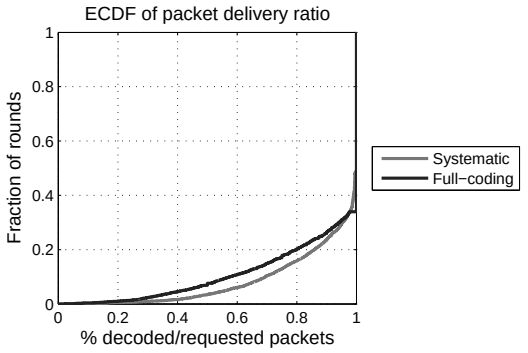


Figure B.27: Packet delivery ratio with different coding breadths for small requested data volume (3KB/s/node).

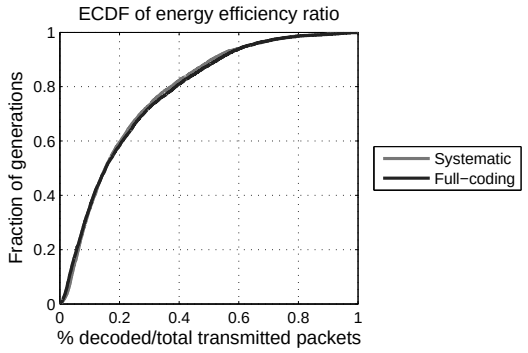


Figure B.28: Energy efficiency ratio with different coding breadths for small requested data volume (3KB/s/node).

References

[1] Amadeu Araújo. 61 bombeiros mortos em serviço numa década. Online. http://www.dn.pt/portugal/interior/61-bombeiros-mortos-em-servico-numa-decada–1637705.html (last visited: 2017-01-10).

[2] HP Enterprise. White paper: Smart cities and the internet of things. Technical report, HP, 2016. http://h20195.www2.hpe.com/v2/getpdf.aspx/4AA6-5129ENW.pdf?ver=1.0 (last visited: 2017-01-11).

[3] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, and Jameela Al-Jaroodi. Applications of big data to smart cities. *Journal of Internet Services and Applications*, 6(1):25, 2015.

[4] Ricardo J Sánchez, Jan Hoffmann, Alejandro Micco, Georgina V Pizzolitto, Martín Sgut, and Gordon Wilmsmeier. Port efficiency and international trade: Port efficiency as a determinant of maritime transport costs. *Maritime Economics & Logistics*, 5(2):199–218, 2003.

[5] Kap Hwan Kim and Hans-Otto Günther. *Container terminals and terminal operations*, pages 3–12. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[6] P. M. Santos, T. E. Abrudan, A. Aguiar, and J. Barros. Impact of position errors on path loss model estimation for device-to-device channels. *IEEE Transactions on Wireless Communications*, 13(5):2353–2361, May 2014.

[7] Pedro M. Santos, Tania Calçada, Susana Sargento, Ana Aguiar, and João Barros. Experimental characterization of i2v wi-fi connections in an urban testbed. In *Proceedings of the 10th ACM MobiCom Workshop on Challenged Networks*, CHANTS '15, pages 5–8, New York, NY, USA, 2015. ACM.

[8] Pedro M. Santos, Tânia Calçada, Ana Aguiar, Daniel Moura, and João Barros. Data collector placement for urban sensing platforms with wireless backhauls. *ACM Transactions on Sensor Networks*, 2015. Submitted to; revised version under preparation.

[9] Theodore S Rappaport et al. *Wireless communications: principles and practice*, volume 2. 1996.

[10] A. J. Rustako, N. Amitay, G. J. Owens, and R. S. Roman. Propagation measurements at microwave frequencies for microcellular mobile and personal communications. In *IEEE 39th Vehicular Technology Conference*, pages 316–320 vol.1, May 1989.

[11] M. Boban, J. Barros, and O. K. Tonguz. Geometry-based vehicle-to-vehicle channel modeling for large-scale simulation. *IEEE Transactions on Vehicular Technology*, 63(9):4146–4164, Nov 2014.

[12] I. Sen and D. W. Matolak. Vehicle-vehicle channel models for the 5-ghz band. *IEEE Transactions on Intelligent Transportation Systems*, 9(2):235–245, June 2008.

[13] L. Cheng, B. E. Henty, D. D. Stancil, F. Bai, and P. Mudalige. Mobile vehicle-to-vehicle narrow-band channel measurement and characterization of the 5.9 ghz dedicated short range communication (dsrc) frequency band. *IEEE Journal on Selected Areas in Communications*, 25(8):1501–1516, Oct 2007.

[14] S.Y. Seidel, T.S. Rappaport, S. Jain, M.L. Lord, and R. Singh. Path loss, scattering and multipath delay statistics in four european cities for digital cellular and microcellular radiotelephone. *IEEE Transactions on Vehicular Technology*, 40(4):721–730, 1991.

[15] Xiongwen Zhao, Jarmo Kivinen, P. Vainikainen, and K. Skog. Propagation characteristics for wideband outdoor mobile communications at 5.3 GHz. *IEEE Journal on Selected Areas in Communications*, 20(3):507–514, 2002.

[16] W. Viriyasitavat, M. Boban, H. M. Tsai, and A. Vasilakos. Vehicular communications: Survey and challenges of channel and propagation models. *IEEE Vehicular Technology Magazine*, 10(2):55–66, June 2015.

[17] Jürgen Kunisch and Jörg Pamp. Wideband car-to-car radio channel measurements and model at 5.9 ghz. In *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*, pages 1–5. IEEE, 2008.

[18] Ian Tan, Wanbin Tang, Ken Laberteaux, and Ahmad Bahai. Measurement and analysis of wireless channel impairments in dsrc vehicular communications. In *Communications, 2008. ICC'08. IEEE International Conference on*, pages 4882–4888. IEEE, 2008.

[19] Johan Karedal, Nicolai Czink, Alexander Paier, Fredrik Tufvesson, and Andreas F Molisch. Path loss modeling for vehicle-to-vehicle communications. *Vehicular Technology, IEEE Transactions on*, 60(1):323–328, 2011.

[20] C. Sommer, S. Joerer, and F. Dressler. On the applicability of two-ray path loss models for vehicular network simulation. In *Proceedings of IEEE Vehicular Networking Conference (VNC)*, pages 64–69, Seoul, Korea, November 2012.

[21] Thomas Mangel, Oliver Klemp, and H. Hartenstein. 5.9 GHz inter-vehicle communication at intersections: a validated non-line-of-sight path-loss and fading model. *EURASIP Journal on Wireless Communications and Networking*, 2011(1):182, November 2011.

[22] R. Meireles, M. Boban, P. Steenkiste, O. Tonguz, and J. Barros. Experimental study on the impact of vehicular obstructions in vanets. In *2010 IEEE Vehicular Networking Conference*, pages 338–345, Dec 2010.

[23] M. Boban, T. T. V. Vinhoza, M. Ferreira, J. Barros, and O. K. Tonguz. Impact of vehicles as obstacles in vehicular ad hoc networks. *IEEE Journal on Selected Areas in Communications*, 29(1):15–28, January 2011.

[24] National Coordination Office for Space-Based Positioning, Navigation, and Timing. Official U.S. Government information about the Global Positioning System (GPS) and related topics. Url: http://www.gps.gov/.

[25] Michael G Wing, Aaron Eklund, and Loren D Kellogg. Consumer-grade global positioning system (gps) accuracy and reliability. *Journal of Forestry*, 103(4):169–173, 2005.

[26] Michael G Wing and Aaron Eklund. Performance comparison of a low-cost mapping grade global positioning systems (GPS) receiver and consumer grade GPS receiver under dense forest canopy. *Journal of Forestry*, 105(1):9–14, 2007.

[27] Paul A Zandbergen. Accuracy of iphone locations: A comparison of assisted gps, wifi and cellular positioning. *Transactions in GIS*, 13:5–25, 2009.

[28] H. Braham, S. B. Jemaa, G. Fort, E. Moulines, and B. Sayrac. Spatial prediction under location uncertainty in cellular networks. *IEEE Transactions on Wireless Communications*, 15(11):7633–7643, Nov 2016.

[29] Matthias Wellens, Burkhard Westphal, and Petri Mahonen. Performance evalu- ation of ieee 802.11-based wlans in vehicular scenarios. In *65th IEEE Vehicular Technology Conference*, pages 1167–71, 2007.

[30] J. Ott and D. Kutscher. Drive-thru internet: Ieee 802.11b for "automobile" users. In *23rd Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2004*, volume 1, pages –373, March 2004.

[31] Pierpaolo Bergamo, M Cesana, D Maniezzo, G Pau, Kung Yao, D Whiteman, and M Gerla. Ieee 802.11 wireless network under aggressive mobility scenario. In *Proc. International Teletraffic Congress (ITC), Las Vegas, NV*, 2003.

[32] R. Gass, J. Scott, and C. Diot. Measurements of in-motion 802.11 networking. In *7th IEEE Workshop on Mobile Computing Systems and Applications*, pages 69–74, Aug 2006.

[33] Vladimir Bychkovsky, Bret Hull, Allen Miu, Hari Balakrishnan, and Samuel Madden. A measurement study of vehicular internet access using in situ wi-fi networks. In *Proceedings of the 12th annual international conference on Mobile computing and networking*, pages 50–61. ACM, 2006.

[34] P. Belanovic, D. Valerio, A. Paier, T. Zemen, F. Ricciato, and C. F. Mecklenbrauker. On wireless links for vehicle-to-infrastructure communications. *IEEE Transactions on Vehicular Technology*, 59(1):269–282, Jan 2010.

[35] C. Xiang, P. Yang, C. Tian, L. Zhang, H. Lin, F. Xiao, M. Zhang, and Y. Liu. Carm: Crowd-sensing accurate outdoor rss maps with error-prone smartphone measurements. *IEEE Transactions on Mobile Computing*, 15(11):2669–2681, Nov 2016.

[36] Bang Wang. Coverage problems in sensor networks: A survey. *ACM Comput. Surv.*, 43(4):32:1–32:53, October 2011.

[37] Krishnendu Chakrabarty, S. Sitharama Iyengar, Hairong Qi, and Eungchun Cho. Grid coverage for surveillance and target location in distributed sensor networks. *Computers, IEEE Transactions on*, 51(12):1448–1453, Dec 2002.

[38] Jie Wang and Ning Zhong. Efficient point coverage in wireless sensor networks. *Journal of Combinatorial Optimization*, 11(3):291–304, 2006.

[39] H. Paul Williams. *Model Building in Mathematical Programming*. John Wiley & Sons, 4th edition, 1978.

[40] Santpal S. Dhillon, Kisbnendu Chakrabarty, and S. Sitharama Iyengar. Sensor placement for grid coverage under imprecise detections. In *Information Fusion, 2002. Proceedings of the Fifth International Conference on*, volume 2, pages 1581–1587. IEEE, 2002.

[41] Santpal S. Dhillon and Krishnendu Chakrabarty. Sensor placement for effective coverage and surveillance in distributed sensor networks. In *Proc. of IEEE Wireless Communications and Networking Conference*, pages 1609–1614, 2003.

[42] Xiaochun Xu and Sartaj Sahni. Approximation algorithms for sensor deployment. *IEEE Trans. Comput.*, 56(12):1681–1695, December 2007.

[43] Yongping Xiong, Jian Ma, Wendong Wang, and Dengbiao Tu. Roadgate: mobility-centric roadside units deployment for vehicular networks. *International Journal of Distributed Sensor Networks*, 2013, 2013.

[44] Oscar Trullols, Marco Fiore, Claudio Casetti, C.F. Chiasserini, and Jose M. Barcelo Ordinas. Planning roadside infrastructure for information dissemination in intelligent transportation systems. *Computer Communications*, 33(4):432 – 442, 2010.

[45] Cristiano M. Silva, Andre L.L. Aquino, and Wagner Meira. Deployment of roadside units based on partial mobility information. *Comput. Commun.*, 60(C):28–39, April 2015.

[46] Baber Aslam, Faisal Amjad, and Cliff Changchun Zou. Optimal roadside units placement in urban areas for vehicular networks. In *Computers and Communications (ISCC), 2012 IEEE Symposium on*, pages 000423–000429. IEEE, 2012.

[47] Javier Barrachina, Piedad Garrido, Manuel Fogue, Francisco J. Martinez, Juan-Carlos Cano, Carlos T. Calafate, and Pietro Manzoni. Road side unit deployment: A density-based approach. *Intelligent Transportation Systems Magazine, IEEE*, 5(3):30–39, Fall 2013.

[48] Christian Lochert, Björn Scheuermann, Christian Wewetzer, Andreas Luebke, and Martin Mauve. Data aggregation and roadside unit placement for a vanet traffic information system. In *Proceedings of the Fifth ACM International Workshop on VehiculAr Inter-NETworking*, VANET '08, pages 58–65, New York, NY, USA, 2008. ACM.

[49] Amine Kchiche and Farouk Kamoun. Centrality-based access-points deployment for vehicular networks. In *2010 IEEE 17th International Conference on Telecommunications (ICT)*, pages 700–706, April 2010.

[50] E. M. Royer and Chai-Keong Toh. A review of current routing protocols for ad hoc mobile wireless networks. *IEEE Personal Communications*, 6(2):46–55, Apr 1999.

[51] Azzedine Boukerche. *Algorithms and protocols for wireless, mobile Ad Hoc networks*, volume 77. John Wiley & Sons, 2008.

[52] Charles E Perkins and Elizabeth M Royer. Ad-hoc on-demand distance vector routing. In *Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA'99. Second IEEE Workshop on*, pages 90–100. IEEE, 1999.

[53] David B Johnson, David A Maltz, Josh Broch, et al. Dsr: The dynamic source routing protocol for multi-hop wireless ad hoc networks. *Ad hoc networking*, 5:139–172, 2001.

[54] Vincent Douglas Park and M Scott Corson. A highly adaptive distributed routing algorithm for mobile wireless networks. In *INFOCOM'97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution., Proceedings IEEE*, volume 3, pages 1405–1413. IEEE, 1997.

[55] Charles E Perkins and Pravin Bhagwat. Highly dynamic destination-sequenced distance-vector routing (dsdv) for mobile computers. In *ACM SIGCOMM Computer Communication Review*, volume 24, pages 234–244. ACM, 1994.

[56] Josh Broch, David A Maltz, David B Johnson, Yih-Chun Hu, and Jorjeta Jetcheva. A performance comparison of multi-hop wireless ad hoc network routing protocols. In *Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, pages 85–97. ACM, 1998.

[57] Shree Murthy and Jose Joaquin Garcia-Luna-Aceves. An efficient routing protocol for wireless networks. *Mobile Networks and Applications*, 1(2):183–197, 1996.

[58] Uichin Lee and Mario Gerla. A survey of urban vehicular sensing platforms. *Computer Networks*, 54(4):527–544, 2010.

[59] 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.

[60] Yoann Dieudonné, Bertrand Ducourthial, and Sidi-Mohammed Senouci. Col: A data collection protocol for vanet. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 711–716. IEEE, 2012.

[61] Mohammad Nozari Zarmehri and Ana Aguiar. Data gathering for sensing applications in vehicular networks (poster). In *Vehicular Networking Conference (VNC), 2011 IEEE*, pages 222–229. IEEE, 2011.

[62] Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. Wireless sensor networks: a survey. *Computer networks*, 38(4):393–422, 2002.

[63] MJ Handy, Marc Haase, and Dirk Timmermann. Low energy adaptive clustering hierarchy with deterministic cluster-head selection. In *Mobile and Wireless Communications Network, 2002. 4th International Workshop on*, pages 368–372. IEEE, 2002.

[64] Stephanie Lindsey and Cauligi S Raghavendra. Pegasis: Power-efficient gathering in sensor information systems. In *Aerospace conference proceedings, 2002. IEEE*, volume 3, pages 3–1125. IEEE, 2002.

[65] Arati Manjeshwar and Dharma P Agrawal. Teen: Arouting protocol for enhanced efficiency in wireless sensor networks. In *IPDPS*, volume 1, page 189, 2001.

[66] Wendi Rabiner Heinzelman, Anantha Chandrakasan, and Hari Balakrishnan. Energy-efficient communication protocol for wireless microsensor networks. In *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*, pages 10–pp. IEEE, 2000.

[67] Sinchan Roychowdhury and Chiranjib Patra. Geographic adaptive fidelity and geographic energy aware routing in ad hoc routing. In *International Conference*, volume 1, pages 309–313, 2010.

[68] Yan Yu, Ramesh Govindan, and Deborah Estrin. Geographical and energy aware routing: A recursive data dissemination protocol for wireless sensor networks. Technical report, Technical report ucla/csd-tr-01-0023, UCLA Computer Science Department, 2001.

[69] Wendi Rabiner Heinzelman, Joanna Kulik, and Hari Balakrishnan. Adaptive protocols for information dissemination in wireless sensor networks. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, pages 174–185. ACM, 1999.

[70] Chalermek Intanagonwiwat, Ramesh Govindan, and Deborah Estrin. Directed diffusion: a scalable and robust communication paradigm for sensor networks. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 56–67. ACM, 2000.

[71] Omprakash Gnawali, Rodrigo Fonseca, Kyle Jamieson, David Moss, and Philip Levis. Collection tree protocol. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, pages 1–14. ACM, 2009.

[72] Rui Meireles, Peter Steenkiste, Jo ao Barros, and Daniel C. Moura. Lasp: Look-ahead spatial protocol for vehicular multi-hop communication. In *IEEE Vehicular Networking Conference*, 2016.

[73] Lorenzo Keller, Emre Atsan, Katerina Argyraki, and Christina Fragouli. Sensecode: Network coding for reliable sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, 9(2):25, 2013.

[74] R. Prior, D.E. Lucani, Y. Phulpin, M. Nistor, and J. Barros. Network coding protocols for smart grid communications. *Smart Grid, IEEE Transactions on*, 5(3):1523–1531, May 2014.

[75] K. C. Lee and M. Gerla. Opportunistic vehicular routing. In *Wireless Conference (EW), 2010 European*, pages 873–880, April 2010.

[76] Sanjit Biswas and Robert Morris. Opportunistic routing in multi-hop wireless networks. *ACM SIGCOMM Computer Communication Review*, 34(1):69–74, 2004.

[77] Cédric Westphal. Opportunistic routing in dynamic ad hoc networks: the oprah protocol. In *Mobile Adhoc and Sensor Systems (MASS), 2006 IEEE International Conference on*, pages 570–573. IEEE, 2006.

[78] Yuan Yuan, Hao Yang, Starsky HY Wong, Songwu Lu, and William Arbaugh. Romer: resilient opportunistic mesh routing for wireless mesh networks. In *IEEE workshop on wireless mesh networks (WiMesh)*, volume 6, 2005.

[79] Sachin Katti, Hariharan Rahul, Wenjun Hu, Dina Katabi, Muriel Médard, and Jon Crowcroft. Xors in the air: practical wireless network coding. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 243–254. ACM, 2006.

[80] Szymon Chachulski, Michael Jennings, Sachin Katti, and Dina Katabi. Trading structure for randomness in wireless opportunistic routing. *SIGCOMM Comput. Commun. Rev.*, 37(4):169–180, August 2007.

[81] Yan Yan, Baoxian Zhang, Jun Zheng, and Jian Ma. Core: a coding-aware opportunistic routing mechanism for wireless mesh networks [accepted from open call]. *Wireless Communications, IEEE*, 17(3):96–103, 2010.

[82] Rudolf Ahlswede, Ning Cai, S-YR Li, and Raymond W Yeung. Network information flow. *Information Theory, IEEE Transactions on*, 46(4):1204–1216, 2000.

[83] S-YR Li, Raymond W Yeung, and Ning Cai. Linear network coding. *Information Theory, IEEE Transactions on*, 49(2):371–381, 2003.

[84] Ralf Koetter and Muriel Médard. Beyond routing: An algebraic approach to network coding. In *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 1, pages 122–130. IEEE, 2002.

[85] Tracey Ho, Ralf Koetter, Muriel Medard, David R Karger, and Michelle Effros. The benefits of coding over routing in a randomized setting. 2003.

[86] Tracey Ho, Muriel Médard, Ralf Koetter, David R Karger, Michelle Effros, Jun Shi, and Ben Leong. A random linear network coding approach to multicast. *Information Theory, IEEE Transactions on*, 52(10):4413–4430, 2006.

[87] Philip A. Chou, Yunnan Wu, and Kamal Jain. Practical network coding, 2003.

[88] Christina Fragouli, Jörg Widmer, and Jean-Yves Le Boudec. A network coding approach to energy efficient broadcasting: From theory to practice. In *INFOCOM*, 2006.

[89] D. C. Teles, M. F. M. Colunas, J. M. Fernandes, I. C. Oliveira, and J. P. S. Cunha. iVital: A real time monitoring mobile system for first responder teams. In *MONAMI 2011*, Aveiro, Portugal, 2011.

[90] Yu Song Meng and Yee Hui Lee. Investigations of foliage effect on modern wireless communication systems: A review. *Progress In Electromagnetics Research*, 105:313–332, 2010.

[91] J. D. Parsons. *The Mobile Radio Propagation Channel*. Wiley, Baffins Lane, Chichester, England, 2 edition, November 2000.

[92] MA Weissberger. An initial critical summary of models for predicting the attenuation of radio waves by trees. *Final Report Electromagnetic Compatibility Analysis Center*, 1, 1982.

[93] A. Seville and K. H. Craig. Semi-empirical model for millimetre-wave vegetation attenuation rates. *Electronics Letters*, 31(17):1507–1508, 1995.

[94] J. A R Azevedo and F.E.S. Santos. An empirical propagation model for forest environments at tree trunk level. *Antennas and Propagation, IEEE Transactions on*, 59(6):2357–2367, 2011.

[95] CCIR. Influences of terrain irregularities and vegetation on troposphere propagation, 1986.

[96] M.P.M. Hall. COST 235 activities on radiowave propagation effects on next-generation fixed-service terrestrial telecommunication systems. In *Proceedings of the 8th International Conference on Antennas and Propagation*, volume 2, pages 655–659, Edinburgh, Scotland, UK, March 1993.

[97] M.O. Al-Nuaimi and A.M. Hammoudeh. Measurements and predictions of attenuation and scatter of microwave signals by trees. *Microwaves, Antennas and Propagation, IEE Proceedings*, 141(2):70–76, 1994.

[98] G.G. Joshi, C.B. Dietrich, C.R. Anderson, W.G. Newhall, W.A. Davis, J. Isaacs, and G. Barnett. Near-ground channel measurements over line-of-sight and forested paths. *IEE Proceedings - Microwaves, Antennas and Propagation*, 152(6):589–596, December 2005.

[99] C. Oestges, B.M. Villacieros, and D. Vanhoenacker-Janvier. Radio channel characterization for moderate antenna heights in forest areas. *IEEE Transactions on Vehicular Technology*, 58(8):4031 –4035, October 2009.

[100] J.A.G. Fernandez, I. Cuiñandas, and M.G. Sánchez. Radioelectric propagation in a deciduous tree forest at wireless networks frequency bands. In *Proceedings of the 5th European Conference on Antennas and Propagation (EUCAP)*, pages 3274–3278, Rome, Italy, April 2011.

[101] J. V. Candy. *Bayesian Signal Processing: Classical, Modern and Particle Filtering Methods*. Wiley-Interscience, New York, NY, USA, 2009.

[102] Yunior Luis, Pedro M. Santos, Tiago Lourenco, Carlos Pérez-Penichet, Tânia Calcada, and Ana Aguiar. Urbansense: an urban-scale sensing platform for the internet of things. Accepted for publication at *2016 IEEE Smart Cities Conference (ISC2)*, 2012. Currently available at [https://feupload.fe.up.pt/get/J7QSAOLn1JbSw9p].

[103] Veniam. Smart city case study: Creating the world's largest network of connected vehicles for smart cities. Online, 2015. https://veniam.com/wp-content/uploads/2015/10/PortoCaseStudy.pdf.

[104] C. Ameixieira, A. Cardote, F. Neves, R. Meireles, S. Sargento, L. Coelho, J. Afonso, B. Areias, E. Mota, R. Costa, R. Matos, and J. Barros. Harbornet: a real-world testbed for vehicular networks. *Communications Magazine, IEEE*, 52(9):108–114, September 2014.

[105] NLANR/DAST. Iperf. https://iperf.fr/.

[106] Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, editors, *Complexity of Computer Computations*, The IBM Research Symposia Series, pages 85–103. Springer US, 1972.

[107] Vasek Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979.

[108] OpenStreetMap Foundation OSMF. Openstreetmap, 2016. https://www.openstreetmap.org.

[109] M. Boban, J. Barros, and O. Tonguz. Geometry-based vehicle-to-vehicle channel modeling for large-scale simulation. *IEEE Transactions on Vehicular Technology*, 63(9):4146–4164, Nov 2014.

[110] Pedro M. Santos, Tania Calçada, Susana Sargento, Ana Aguiar, and João Barros. Experimental characterization of i2v wi-fi connections in an urban testbed. In *Proceedings of the 10th ACM MobiCom Workshop on Challenged Networks*, CHANTS '15, pages 5–8, New York, NY, USA, 2015. ACM.

[111] DAVID S. JOHNSON. Approximation algorithms for combinatorial problems. *JOURNAL OF COMPUTER AND SYSTEM SCIENCES*, 9:256–278, 1973.

[112] Carlos Ameixieira, André Cardote, Filipe Neves, Rui Meireles, Susana Sargento, Luís Coelho, João Afonso, Bruno Areias, Eduardo Mota, Rui A. Costa, Ricardo Matos, and João Barros. Harbornet: A real-world testbed for vehicular networks. *CoRR*, abs/1312.1920, 2013.

[113] Alec Woo, Terence Tong, and David Culler. Taming the underlying challenges of reliable multihop routing in sensor networks. In *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems*, SenSys '03, pages 14–27, New York, NY, USA, 2003. ACM.

[114] Sze-Yao Ni, Yu-Chee Tseng, Yuh-Shyan Chen, and Jang-Ping Sheu. The broadcast storm problem in a mobile ad hoc network. In *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking*, MobiCom '99, pages 151–162, New York, NY, USA, 1999. ACM.

[115] C. F. Chiasserini, E. Viterbo, and C. Casetti. Decoding probability in random linear network coding with packet losses. *IEEE Communications Letters*, 17(11):1–4, November 2013.

[116] Ming Xiao, T. Aulin, and M. Medard. Systematic binary deterministic rateless codes. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2066–2070, July 2008.

[117] April Rasala Lehman and Eric Lehman. Complexity classification of network information flow problems. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '04, pages 142–150, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics.

[118] S. Feizi, D.E. Lucani, C.W. Sorensen, A. Makhdoumi, and M. Medard. Tunable sparse network coding. In *Proc. Int. Zurich Seminar Commun.*, pages 107–110, March 2012.

[119] P. Pakzad, C. Fragouli, and A. Shokrollahi. Coding schemes for line networks. In *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, pages 1853–1857, Sept 2005.

[120] D.S. Lun, P. Pakzad, C. Fragouli, M. Medard, and R. Koetter. An analysis of finite-memory random linear coding on packet streams. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2006 4th International Symposium on*, pages 1–6, April 2006.

[121] B. Haeupler and M. Medard. One packet suffices - highly efficient packetized network coding with finite memory. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 1151–1155, July 2011.

[122] D. S. Lun, M. Medard, and R. Koetter. Network coding for efficient wireless unicast. In *Communications, 2006 International Zurich Seminar on*, pages 74–77, 2006.

[123] Daniel Enrique Lucani, Milica Stojanovic, and Muriel Médard. Random linear network coding for time division duplexing: When to stop talking and start listening. In *INFOCOM 2009, IEEE*, pages 1800–1808. IEEE, 2009.

[124] J. K. Sundararajan, D. Shah, and M. Medard. Arq for network coding. In *2008 IEEE International Symposium on Information Theory*, pages 1651–1655, July 2008.

[125] Pedro M. Santos, Tânia Calçada, Diogo Guimarães, Tiago Condeixa, Susana Sargento, Ana Aguiar, and João Barros. Demo: Platform for collecting data from urban sensors using vehicular networking. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, pages 167–169, New York, NY, USA, 2015. ACM.

[126] Pedro M. Santos, João Rodrigues, Susana Cruz, Tiago Lourenço, Pedro M. D'Orey, Yunior Luis, Susana Sargento, Ana Aguiar, and João Barros. Portolivinglab: an iot-based sensing platform for smart city. *IEEE Internet of Things Journal*, 2017. Under review.

[127] L. Kholkine, P. M. Santos, A. Cardote, and A. Aguiar. Detecting relative position of user devices and mobile access points. In *IEEE Vehicular Networking Conference VNC*, 2016.

[128] Pedro M. Santos, Fausto Vieira, and João Barros. Demo: Dynamic building evacuation system. Online, unpublished, 2010. https://www.youtube.com/watch?v=PJSqnq6ZfOE.

[129] Luis Pinto, Pedro M. Santos, Sérgio Crisóstomo, Traian E. Abrudan, and João Barros. On-the-fly deployment of wireless sensor networks for indoor assisted guidance. In *2013 IEEE International Conference on Cyber-Physical Systems, Networks and Applications*, 2013.

[130] Daniel K. Schrader, Byung-Cheol Min, Eric T. Matson, and J. Eric Dietz. Real-time averaging of position data from multiple {GPS} receivers. *Measurement*, 90:329 – 337, 2016.

[131] l. Hai, J. Wang, p. wang, H. Wang, and T. Yang. High throughput network coding aware routing in time-varying multi-hop networks. *IEEE Transactions on Vehicular Technology*, PP(99):1–1, 2016.

[132] Jon Kleinberg and Eva Tardos. *Algorithm Design.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.