THE POWER OF QUASI-SHORTEST PATHS AND THE IMPACT OF NODE MOBILITY ON DYNAMIC NETWORKS

Dianne Scherly Varela de Medeiros

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Miguel Elias Mitre Campista

Rio de Janeiro
Setembro de 2017

THE POWER OF QUASI-SHORTEST PATHS AND THE IMPACT OF NODE MOBILITY ON DYNAMIC NETWORKS

Dianne Scherly Varela de Medeiros

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Miguel Elias Mitre Campista, D.Sc.

Prof. Guy Pujolle, Dr. d'État

Prof. Luiz Fernando Bittencourt, D.Sc.

Prof. Igor Monteiro Moraes, D.Sc.

Prof. Luís Henrique Maciel Kosmalski Costa, Dr.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2017

Medeiros, Dianne Scherly Varela de

The power of quasi-shortest paths and the impact of node mobility on dynamic networks/Dianne Scherly Varela de Medeiros. – Rio de Janeiro: UFRJ/COPPE, 2017.

XVI, 105 p.: il.; 29, 7cm.

Orientador: Miguel Elias Mitre Campista

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2017.

Referências Bibliográficas: p. 98 – 105.

1. Multihop communication. 2. (κ, λ)-vicinity. 3. Quasi-shortest paths. 4. ρ-geodesic betweenness. 5. Spreadness factor. I. Campista, Miguel Elias Mitre. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

To my parents, my professors,
and my fiancé.

Acknowledgments

I would like to thank, first and foremost, God, who gave me strength to achieve my goals. I also thank my parents Antonio Roberto and Shirley, for all the support they gave me during all my life. Thank you for making me who I am.

I thank my advisor Miguel Elias Mitre Campista, for all his guidance and help during this work. Thank you for all the suggestions and advice you gave me. You have contributed immensely to my professional and personal growth. I hope to be as good an advisor as you are.

I thank the suggestions from the members of the jury of my qualifying exam, Marcelo Dias de Amorim, Guy Pujolle and Luís Henrique Maciel Kosmalski Costa. I also thank the members of the jury of my thesis defense, Guy Pujolle, Luís Henrique Maciel Kosmalski Costa, Igor Monteiro Moraes and Luiz Fernando Bittencourt, for being available to evaluate my work.

I thank all the professors from the laboratory Grupo de Teleinformática e Automação (GTA) at Universidade Federal do Rio de Janeiro (UFRJ), Miguel Elias Mitre Campista, Luís Henrique Maciel Kosmalski Costa, Pedro Braconnot Velloso, Aloysio de Castro Pinto Pedroza and Otto Carlos Muniz Bandeira Duarte. I acknowledge you, Miguel and Otto, and also Professor Guy Pujolle, for providing me a remarkable experience. I thank you Guy, for all the help and attention, and I appreciate the time spent in your team. Thank you for receiving me at Université Pierre et Marie Curie (UPMC) and providing me great opportunities. I also thank Marcelo Dias de Amorim for sharing his great ideas with me. Thank you for your time and support. I further thank Nathalie Mitton for receiving me at Inria Lille Nord Europe and introducing me to the world of emulated wireless sensor networks. I also would like to thank professors Jefferson Elbert Simões and Celina Herrera de Figueiredo for their technical help.

I thank the friends from GTA and Équipe PHARE/UPMC, for their help and friendship. I acknowledge you all for contributing to the development of this work. Specially, thank you Borges, Rodrigo and Sadok for teaching me and helping me every time I needed. Thank you Diogo, for the funny moments and for your participation in my experience abroad, at the Laboratoire d'Informatique Paris 6 (LIP6/UPMC). We are never alone when we have friends. I thank each and every

friend from GTA for making work at this laboratory a wonderful and fun experience.

I also would like to thank the staff of the Programa de Engenharia Elétrica from COPPE/UFRJ, specially Daniele and Maurício for their dedication and agility to solve our academic problems.

I thank all the professors that participated in my formation and everyone that contributed in one way or another to my personal formation. Finally I thank the financial support provided by the funding agencies CNPq, FAPERJ and CAPES.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

O PODER DOS CAMINHOS QUASE MAIS CURTOS E O IMPACTO DA MOBILIDADE DOS NÓS EM REDES DINÂMICAS

Dianne Scherly Varela de Medeiros

Setembro/2017

Orientador: Miguel Elias Mitre Campista

Programa: Engenharia Elétrica

O objetivo desta tese é investigar três aspectos importantes das redes dinâmicas: o impacto da mobilidade dos nós na transmissão de dados em múltiplos saltos, o efeito do uso de caminhos mais longos na importância relativa dos nós, e o desempenho da rede na presença de falha em nós centrais. Para analisar o primeiro aspecto, este trabalho propõe a (κ, λ)-vizinhança, que estende a vizinhança tradicional para considerar como vizinhos nós a múltiplos saltos de distância e restringe o estabelecimento de enlaces de acordo com a velocidade relativa entre os nós. Essa proposta é usada posteriormente no desenvolvimento de três estratégias de encaminhamento. A restrição de velocidade relativa imposta nessas estratégias resulta em uma redução significativa do consumo de recursos, sem que ocorra impacto significativo na taxa média de entrega de pacotes. Para analisar o segundo aspecto, propõe-se a centralidade de intermediação ρ-geodésica, que usa caminhos mais curtos e quase mais curtos para quantificar a importância relativa dos nos. Os caminhos quase mais curtos são limitados por um fator de espalhamento ρ. O uso de caminhos não-ótimos provoca o reranqueamento de diversos nós e tem como principal efeito uma menor ocupação de posições mais centrais por pontos de articulação. Por fim, o desempenho da rede em presença de falha é investigado através de simulações nas quais as falhas atingem nós definidos como os mais centrais de acordo com métricas de centralidade distintas. O resultado é uma redução brusca da vazão média da rede, independentemente da métrica usada para determinar quais são os nós mais centrais. O grande trunfo da métrica proposta é que, apesar da severa redução na vazão, é grande a probabilidade de manter a rede conectada após a falha, uma vez que é pouco provável que um nó em falha nas posições mais centrais seja também um ponto de articulação.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

THE POWER OF QUASI-SHORTEST PATHS AND THE IMPACT OF NODE MOBILITY ON DYNAMIC NETWORKS

Dianne Scherly Varela de Medeiros

September/2017

Advisor: Miguel Elias Mitre Campista

Department: Electrical Engineering

The objective of this thesis is to investigate three important aspects of dynamic networks: the impact of node mobility on multihop data transmission, the effect of the use of longer paths on the relative importance of nodes and the performance of the network in the presence of failure on central nodes. To analyze the first aspect, this work proposes the (κ, λ)-vicinity, which extends the traditional vicinity to consider as neighbors nodes at multihop distance and restricts the link establishment according to the relative speed between nodes. This proposal is used later on the development of three forwarding strategies. The relative speed restriction imposed on these strategies results in significant reduction of resources consumption, without incurring significant impact on the average packet delivery ratio. To analyze the second aspect, we propose the ρ-geodesic betweenness centrality, which uses shortest and *quasi*-shortest paths to quantify the relative importance of a node. The *quasi*-shortest paths are limited by a spreadness factor, ρ. The use of non-optimal paths causes the reranking of several nodes and its main effect is a reduced occupation of the most central positions by articulation points. Lastly, the network performenace in presence of failures is investigated through simulations, in which failures happen on nodes defined as the most central according to distinct centrality metrics. The result is a severe reduction of the average network throughput, and it is independent of the metric used to determine which nodes are the most central. The major strength of the proposed metric, then, is that, despite the severe reduction of the throughput, there is a high probability of maintaining the network connected after a failure, because it is unlikely that a failing node in the most central position is also an articulation point.

Contents

List of Figures **xii**

List of Tables **xvi**

1 Introduction **1**

1.1 Objectives . . . 3
1.2 Contributions . . . 7
1.3 Organization . . . 9

2 Network Model, Concepts and Definitions **10**

2.1 Network model . . . 10
2.2 Paths and costs . . . 11
2.3 Taking nodes on quasi-shortest-path into account . . . 11
2.4 Connectivity and articulation points . . . 13
2.5 Node vicinity . . . 13
2.6 State of a node . . . 15
2.7 Vicinity timeline . . . 16

3 Datasets Description and Characterization **18**

3.1 Dynamic network datasets . . . 18
3.1.1 Mobility Dataset: Taxi scenario . . . 21
3.1.2 Ad Hoc City Dataset: Bus scenario . . . 21
3.1.3 TAPASCologne Dataset: Cologne synthetic scenario . . . 22
3.2 Static network datasets . . . 23
3.2.1 Freeman's EIES . . . 23
3.2.2 Doubtful Sound Dolphins . . . 24
3.2.3 PhD Students . . . 24
3.2.4 Snapshots from the TAPASCologne dataset . . . 24
3.2.5 Randomly generated static network datasets . . . 26

I Multihop Communications in Dynamic Networks 28

4 Multihop Network Connectivity 29
4.1 Background . . . 29
4.2 Problem statement . . . 31
4.3 Vicinity analysis methodology . . . 34
4.3.1 Mobility trace parsing . . . 35
4.3.2 Discovery of $(0, 1)$-vicinity . . . 36
4.3.3 Computation of (κ, λ)-vicinity . . . 36
4.3.4 Generation of (κ, λ)-vicinity timeline . . . 37

5 Vicinity Analysis: Results and Discussion 38
5.1 Results . . . 38
5.1.1 Behavior of nodes' relative speed . . . 38
5.1.2 Influence of relative speeds on 1-hop contacts duration . . . 40
5.1.3 Behavior of number and duration of contacts per (κ, λ)-vicinity 43
5.1.4 Average time spent by nodes in each State σ . . . 45
5.1.5 Number of useful contacts according to the (κ, λ)-vicinity . . . 46
5.2 Discussion . . . 48
5.2.1 Applications . . . 49
5.2.2 Relative-speed-aware packet forwarding . . . 51

II Assessment of Node Importance Using Quasi-Shortest Paths 57

6 The ρ-Geodesic Betweenness Centrality 58
6.1 Background . . . 58
6.1.1 Betweenness centrality . . . 59
6.1.2 Accounting path length: bounded-distance, distance-scaled and linearly-scaled betweenness . . . 60
6.1.3 Analyzing weighted networks: flow and Opsahl's betweenness . 61
6.1.4 Including longer paths: current flow and random walk betweenness . . . 62
6.2 The ρ-geodesic betweenness . . . 64
6.2.1 Metric overview . . . 64
6.2.2 Metric formalization . . . 66
6.2.3 Properties . . . 67
6.2.4 Algorithm for computing the ρ-geodesic centrality . . . 69

7 ρ-Geodesic Betweenness: Characterization and Discussion 72
7.1 Analysis guidelines . . . 72
7.2 Metrics correlation . . . 73
7.3 Metrics reranking ability . . . 75
7.4 Intermediation ability . . . 79
7.5 Fault tolerance analysis . . . 81
7.6 Discussion . . . 84
7.6.1 Random walk and ρ-geodesic betweenness centralities in scale-free networks . . . 86
7.6.2 Applications . . . 87

8 Conclusions and Future Work 92

Bibliography 98

List of Figures

2.1 The shortest path between v_i and v_j is $\lambda^*_{i,j} = 3$ hops long. If $\rho = 1$, the *quasi*-shortest path of length $\lambda^*_{i,j} = 4$ through v_k can be considered too. 12
2.2 Definition of v_i's (κ, λ)-vicinity, disregarding vehicles relative speeds ($m = 1$). Subgraph $\mathcal{G}_{R_0}$ coincides with $\mathcal{G}$, including all nodes in the network. 14
2.3 Example of $\mathcal{G}_{R_\kappa}$ for nodes with relative speed in R_κ, the links connecting them, and the (κ, λ)-vicinity of node v_i. 15

3.1 Normalized frequency of updates in each scenario. We divide the cities in small areas of 100 $\times$ 100 m^2 and we analyze the sampled period of each dataset. We consider that areas where we can find frequent updates are also denser areas. 19
3.2 Cumulative distribution function of absolute speeds for each scenario. (c) Cars in the synthetic scenario register the highest absolute speeds, followed by the (a) taxis and the (b) buses. The latter registers the highest number of very low absolute speeds and the most uniform distribution among the analyzed datasets. 20
3.3 The visualization of the network topology provided by each dataset highlights the differences between them. The traditional betweenness is represented in the node size while the degree, in the node color, such that larger nodes have higher traditional betweenness and more bluish nodes have higher degree. 25
3.4 Comparison between a star network and sample random networks with power law degree distribution for different α. All networks have 100 nodes and it is clear that the structure of the network changes with α, becoming more similar to a star as α increases. 26

4.1 Example of (κ, λ)-vicinity timelines, considering different values for m to show the influence of relative speeds. Nodes at lower relative speeds can reach up to 8-hop distance, whereas at higher relative speeds the maximum hop distance drops to $\lambda = 3$. 32
4.2 State transition probabilities (p_{nm}) in percentage, considering $m = 1$ and $m = 5$. The states and transition probabilities considering only the (b) subgraph where nodes are in contact at lower relative speeds is very similar to the distribution presented by the (a) complete graph. Nodes at (c) higher relative speeds are not able to reach upper states and have higher probability of transitioning to State ∞. 33

5.1 Cumulative distribution function of relative speeds for contacts at 1-hop distance. In the (a) Taxi and (b) Bus scenarios, the relative speeds of 1-hop contacts are usually low. In the (c) Synthetic scenario we observe a wider range of relative speeds. 39
5.2 Contact duration for 1-hop contacts as a function of relative speed for each scenario. 40
5.3 Total number of contacts as a function of the (κ, λ)-vicinity for each scenario, using $c = \{100, 200\}$ m. In (a) Taxi and (b) Bus scenarios there is a significant amount of contacts communicating at hop distances lower than 3. In (c) Synthetic scenario we can find several nodes communicating even at 6-hop distance. 43
5.4 Sum of all contacts duration as a function of the (κ, λ)-vicinity for each scenario, using $c = \{100, 200\}$ m. In the (a) Taxi and (b) Bus scenarios, the most significant contribution to the total contact duration happen at very low relative speeds (< 10 km/h) and up to 2-hop distance. In the (c) Synthetic scenario the most significant contribution is obtained for nodes communicating up to 6-hop distance and at relative speeds between 40 and 50 km/h, followed by very low relative speeds (< 10 km/h). 45
5.5 Cumulative distribution function of the duration of contacts for the first vicinity and the last vicinity able to successfully communicate using packet bundles. In all scenarios both the hop distance and the relative speeds harshly influence the communication. In the (a) Taxi and (b) Bus scenarios there are more nodes able to fully transfer big packet bundles. 47
5.6 Modified format of the `OLSR` Hello header. The `AbsSpeedX` and `AbsSpeedY` (marked in gray) fields replace a previously 16-bit `Reserved` field. 53

5.7 Comparison of the performance metrics using OLSR and each relative-speed-aware forwarding strategy based on the OLSR. The average packet delivery ratio for all strategies is statistically equal, but VicR stands out because it significantly reduces resource consumption. RelSpeedVicR is also able to reduce resource consumption, even though less than VicR. 55

6.1 Example of network topology where betweenness centrality metrics can fail to capture the importance of a node on a quasi-shortest path. The clouds represent any type of network topology, as long as it is connected. 64

6.2 The nodes in the example can be divided in 3 sets according to their ability to intermediate flows: $\{v_a, v_e\}$ will never intermediate them, $\{v_b, v_d\}$ intermediate the great majority, and $\{v_c\}$ intermediates if necessary. Nevertheless, shortest-path-based centralities, such as the traditional and the distance-scaled betweenness, classify v_c in the same set of v_a and v_e. 68

7.1 The correlation with the traditional betweenness is clearly strong for all metrics, being stronger for the distance-scaled betweenness. The random walk and ρ-geodesic betweenness show more capability to identify nodes that could receive a different value for the betweenness. Note that the axes are normalized by $\big(|\mathcal{V}| - 1\big) \cdot \big(|\mathcal{V}| - 2\big)$ if the graph is directed, and by $0.5 \cdot \big(|\mathcal{V}| - 1\big) \cdot \big(|\mathcal{V}| - 2\big)$ otherwise. 74

7.2 The Kendall's W coefficients plotted to each pairwise combination of betweenness centrality metrics show a high level of agreement between them. The lowest concordance happens between the traditional and the ρ-geodesic betweenness for $\rho = 5$. This is due to the potentially numerous *quasi*-shortest paths considered on the metric computation, which vary more significantly the importance of nodes and, consequently, their rank positions. 76

7.3 The distance-scaled, random walk and ρ-geodesic betweenness are able to redistribute the node ranking to different extents, compared to the traditional betweenness. 77

7.4 Compared to the traditional betweenness, the distance-scaled and ρ-geodesic betweenness are able to spread the classification rank, giving room to more positions. Hence, we find less nodes tied in the same position. The random walk betweenness surprisingly increases the number of tied nodes in the Dolphins dataset. 78

7.5 The proposed ρ-geodesic betweenness is able to reduce the number of times that nodes lose their ability to intermediate flows in the network compared to the other metrics, even for the most important nodes. The ρ-geodesic betweenness can prevent up to 3 reallocations more than the traditional and distance-scaled betweenness. 80
7.6 In the highly dynamic scenario of the Cologne vehicular network, the ρ-geodesic betweenness is the metric that changes less frequently the ranking of nodes. Of all nodes in the network, 1.2% remain in the same rank position for up to 10 seconds. 81
7.7 The number of articulation points existing on each snapshot of the dynamic Cologne network considering the top #5 positions. In general, this number is lower for the distance scaled (B_{dist}) and ρ-geodesic (B_ρ) betweenness, when compared to the traditional betweenness (B_{trad}). . 82
7.8 Compared to the traditional (B_{trad}) and distance scaled (B_{dist}) betweenness, a single failure on one of the top #5 nodes of the ρ-geodesic betweenness (B_ρ) is generally less harsh to the average network throughput. 84
7.9 Nodes classified in the first position according to the distance scaled and ρ-geodesic betweenness are frequently the same as the traditional betweenness, considering all samples of the Cologne dataset. For lower ranks, the frequency is significantly reduced. 85
7.10 Averaged results for the maximum and minimum absolute differences between the random walk and the ρ-geodesic betweenness, for $\rho = \{1, 3\}$. The minimum difference is independent of ρ, while the maximum difference becomes more significant for higher ρ. 87
7.11 Example topology to model the delivery of goods. Three types of truck drivers want to travel from the factory (v_a) to the store (v_b) to deliver some products. Driver type `I` (ordinary) always follow paths with the most central nodes; driver type `II` (impatient) does not want to risk wasting time on a possibly congested shortest path; and driver type `III` (altruist) does not want to congest the shortest paths and, thus, follow slightly longer paths. 90

List of Tables

2.1 Summarized notation used in this work. 17

3.1 Dynamic network datasets main characteristics. 18
3.2 Static network datasets main characteristics. 23

5.1 Relation between contact duration, relative speeds and physical distance between nodes. Non-existent contacts are represented as "Ø" and existing contacts as "+". We use "++" for each combination of relative speed and duration where contacts exist to indicate at which distance we observe more contacts. 42
5.2 Average time spent in each state, in seconds. 46
5.3 Source files in NS-3 that implement the modified classes related to `OLSR`'s operation. 52
5.4 Relative speed thresholds according to the hop distance to the destination. 54

6.1 Summary of betweenness centrality metrics discussed in this work. . . 63
6.2 Comparison of traditional (B_{trad}), distance-scaled (B_{dist}) and random walk (B_{rnd}) betweenness metrics for nodes in Figure 6.1. 66
6.3 Comparison of the betweenness of nodes in Figure 6.2. Both the random walk (B_{rnd}) and ρ-geodesic betweenness (B_{ρ}) are able to broaden the node ranking. The distance-scaled (B_{dist}) and traditional (B_{trad}) betweenness cannot capture the importance of node v_c. 68

7.1 Number of components in each snapshot of the Cologne network before and after failure on a single articulation point. 82
7.2 The probability that a failed node is also an articulation point is higher in the majority of cases for the traditional betweenness, compared to the other metrics. 83

Chapter 1

Introduction

Network science is an ascending interdisciplinary field that gained much importance in the last few decades [1]. It is applied to the study of several topics in different research areas, such as physics, biology, sociology, economics and engineering, covering a wide range of networks, e.g., trading, semantic, information, terrorist, social, computer, genetic, infrastructure, and protein networks, among many others. Thanks to network science, researchers are able to better understand the interactions and associations among network elements, so that they can discover the fundamental principles that govern the network behavior, structure and functionalities [1]. As a consequence, network science allows researchers to model real world phenomena, such as the spread of viruses and the "rich gets richer" effect, or even determine the role of each element within the network. To this end, it is necessary to use several tools, such as graph theory, data mining, inferential modeling, and social structures. As a final step, network science provides the necessary knowledge to control or, at least, predict the behavior of real systems [2].

Real systems are usually put together under the domain of complex networks, which are characterized by an irregular and complex structure that dynamically evolves over time [2]. In addition, they represent systems with thousands or millions of nodes, such as neural, genetic, transportation, vehicular, computer, electrical, and telecommunications networks, as well as the Internet and the World Wide Web, among many others [2]. Nowadays, people that study such networks are mainly interested in understanding their dynamical behavior. Particularly, researchers aim to investigate how the network structure (i.e., network topology) affects the system properties over time. The performance of wireless networks, for instance, is strongly governed by the network structure. Such networks remain in spotlight even after years of extensive investigation, because new challenges frequently arise.

The dynamics of wireless networks structure is related to two properties: the link quality and the node mobility. The latter is intrinsic to dynamic networks, which are characterized for being prone to frequent topology changes. The result

is the removal and addition of link as nodes move around, causing intermittent connectivity. The link quality can change the network structure even when nodes cannot move, as in static networks. This happens because any modification on the network surroundings, such as weather fluctuations, can improve or degrade the link quality, consequently changing the link state. Hence, to study the influence of the network structure on system properties, we need to analyze both dynamic and static networks.

In this thesis we focus on Vehicular Ad Hoc Networks (VANETs), which are a special case of wireless networks, where node mobility can be very intense. Nevertheless, we also use static networks, derived from social and randomly generated networks. The dynamic nature of VANETs adds even more challenges to the wireless paradigm. Handling mobility in such networks remains an open research issue, especially when communications occur through multiple hops. Multihop communications are jeopardized by a number of obstacles, such as the ability of nodes to move around and the intra-flow interference. As a consequence, these networks often face intermittent connectivity, lack of end-to-end paths, and frequent changes of intermediary nodes on a path, preventing efficient data transfers [3].

what action the node should perform with the packet, for instance, it could instantly forward it or drop it if no neighbors are within its radio range.

Routing in VANETs cannot be designed as in static wireless networks, because, in the latter, lack of end-to-end paths is transitory. In VANETs, there is a high probability of not having a fully connected network and, consequently, end-to-end paths can be rare. Therefore, nodes need to have information about their surroundings to decide the action they should perform on the packet, e.g., drop it or forward it to the next hop. To make this decision, nodes need to know how to make the most of contact opportunities. When developing routing protocols for VANETs, it is also important to consider that in these networks, there is a reach of interest that depends on the application, i.e., communication happens between nodes within a region and, usually, it does not involve nodes in distant areas of a city. The performance of multihop communications in such scenarios is highly dependent of nodes ability to establish efficient routes according to the current network conditions [4]. Researchers tackle this problem by proposing prediction mechanisms to anticipate contact availability [5] and disruption [6].

Network science also provides tools to discover the roles played by nodes within the networks. To this end, it is necessary to evaluate nodes relative importance, improving the decision making process. For instance, based on the acquired knowledge one can find nodes that play the role of brokers and decide to prevent their failure and protect them from attacks at any cost. This is because brokers are responsible for everything that travels between two communities and a failure on such

nodes can interrupt the flow between these communities. In computer networks, a broker is a bridge, and the communities could be two different Local Area Networks (LANs). Many flows are processed by the broker, thus, one could decide to install flow analyzers on such nodes. Hence, important nodes are good candidates to run a number of control functions, or to help with content dissemination [7–9], depending on their roles within the network. Such important nodes are also said to be central nodes and we use both terms interchangeably in this thesis.

The definition of central node may change from one application to the other and the identification of central nodes is not trivial, particularly in large and dynamic networks. This identification is fundamental in several networks [10–14]. The most usual metrics to assess node importance according to its structural position in the network are the centrality metrics [15–17]. There are plenty of them and some will be discussed in this thesis. Meanwhile, it is only necessary to know that the main centrality metrics are degree, closeness, betweenness and eigenvector centralities. The other existing centralities are usually variants based on them. The *degree centrality* relates to how popular a node is. The *closeness centrality* is related to how quickly a flow can spread from a node to all other nodes in the network, i.e., a node with high closeness is close to all other nodes in the network. The *eigenvector centrality* relates to how well connected a node is, so that a node is important if its acquaintances are also important. Lastly, the *betweenness centrality* relates the importance of a node with the number of shortest paths (geodesics) it belongs to [15]. Nodes that play the role of brokers have high betweenness, because they are in-between many other nodes (along the shortest path). Hence, they can control when and what flows between other pairs of nodes, if it flows along shortest paths.

Investigating the structural importance of nodes in both dynamic and static networks is fundamental to make better decisions. Particularly, in dynamic networks, network science can also help to provide insights on the influence of node mobility, so that we can maximize the exploitation of contact opportunities. We investigate both aspects in this thesis. In the following section we highlight our specific objectives.

1.1 Objectives

In this thesis, we investigate the influence of the network structure on the system behavior. More specifically, we study how this structure affects (i) data forwarding and (ii) nodes relative importance. We also (iii) introduce an analysis that focuses on network resilience. To this end, we use dynamic wireless networks, but we also rely on static networks to lay the basis of our analysis. Hence, the objectives of this work are three-fold:

1. To investigate the impact of node mobility on the establishment of multihop communications, through the analysis of node vicinity;

2. To analyze the influence of longer paths on the assessment of node centrality; and

3. To verify the impact of single failures of the most central nodes on the network performance when flows follow shortest paths.

The vicinity analysis is essential for vehicular networking, because it allows a better comprehension of contact opportunities. The idea is to maximize the communicability between pairs of vehicles, i.e., increase the number of opportunities to successfully transfer data. This is important if vehicle-to-vehicle communications are used to extend the communication range of a vehicle. For instance, in drive assistant applications it is necessary that vehicles have an extended horizon of awareness, beyond their local surroundings. This is only possible if vehicles can communicate successfully with other vehicles at multihop distances. There is a limit, however, for the necessary reachability of the communication. Depending on the application, there is no need for a vehicle in the south of a city to communicate with a group of vehicles in the north of the same city. In addition, considering other types of applications, such as entertainment (or infotainment), it is necessary to maximize data transfer between vehicles, which also requires to study and deeply understand the behavior of contact opportunities. Several studies concerning mobility patterns and connectivity in Mobile Ad Hoc Networks (MANETs) already exist and many of them are compiled in several surveys [18–23]. Efforts are also made to address these issues in VANETs and solutions have already been proposed to partially overcome the intermittent connectivity problem [5, 6, 24–26]. Other works analyze contact opportunities through the study of node vicinity [27–29]. Typically, a contact happens when two nodes are within mutual radio range, restricting node vicinity to directly reachable nodes. This restriction is not detrimental to a number of applications for which the 1-hop vicinity is sufficient, such as the detection of congestion in urban scenarios [30]. Nevertheless, applications that rely on the communication between nodes separated by longer hop distances waste several contact opportunities due to the vicinity restriction. Besides that, node vicinity changes frequently due to the intense node mobility. Phe-Neau et al. [31, 32] use an unusual approach to exploit contact opportunities. They extend the concept of node vicinity in terms of hops to also consider nodes reachable at longer distances. The resulting extended vicinity incorporates nodes even if they are out of mutual radio range and, as a consequence, nodes find more contact opportunities, through multihop contacts. We know, however, that VANETs suffer with intermittent connectivity, notably in multihop communications.

In this work, we analyze contact opportunities, taking into account the relationship between nodes extended vicinity and their relative speeds. The goal is to quantitatively evaluate the expected notion of "better connectivity at lower relative speeds" to shed more light into multihop communications in typical vehicular scenarios. To this end, we propose a methodology to group nodes according to their relative speeds, i.e., we consider links only between nodes at a certain interval of relative speeds. This restriction can influence the contact duration, but it helps to identify conditions for multihop communications, which in a broader sense, depend on whether opportunistic contacts appear for long enough to be considered useful. We first (i) study the influence of nodes' relative speed on the vicinity behavior. Then, we (ii) further extend the concept of node vicinity to also include nodes' relative speed. We consider that relative speeds are more suitable than absolute speeds because they determine contact duration. The idea is to identify the feasibility of multihop communications in typical vehicular scenarios. We also (iii) analyze the usefulness of such communications through the investigation of how much data a node could transfer to its peer during the available contact duration. Finally, (iv) we propose and evaluate simple forwarding mechanisms that use the outcomes of the vicinity analysis.

In addition to the vicinity analysis, a new rationale behind the definition of node centrality is also proposed in this thesis. Centrality metrics are used to determine the importance of a node to the network. The focus is on the betweenness centrality because it finds nodes that can potentially intermediate more flows between other nodes in the network. Note that, in network science, "flow" is a broad term that can represent anything that travels across a network, such as packets, gossips, electrical or chemical signals, vehicles, among other entities. Thus, its definition will depend on the type of network studied. For instance, if we analyze a computer network, a flow will represent packets exchanged between communicating nodes. Such networks can benefit from the traditional concept of betweenness, which uses shortest paths to determine node importance. In computer networks, e.g., it can be used to estimate the monitoring and control capabilities of a node [33], to design protocols to elect nodes as cluster heads [34], to detect the location of vulnerabilities in a network [34], to design routing protocols in delay tolerant [10, 16, 35, 36] and wireless sensor [37] networks, among other applications.

Several works question the use of shortest paths as the sole parameter to quantify the importance of nodes [38–42]. We also question this approach and we argue that it may underestimate important nodes — in particular, those in the close vicinity of shortest paths but that do not belong to them. This happens when a node that falls on many paths slightly longer than the shortest path is ignored by the betweenness centrality. We inquire why these nodes are neglected, if they are good

candidates, in practice, to maintain the network connected in case of failure of more important nodes, or to reduce the load on such nodes. Hence, we propose a weighted betweenness centrality, the ρ-geodesic betweenness, which extends the traditional betweenness proposed by Freeman [43] to also consider slightly longer paths when assessing node importance. Such paths are defined herein as *quasi*-shortest paths. In a nutshell, the ρ-geodesic betweenness of a node v_k is computed using the proportion of shortest and *quasi*-shortest paths that v_k falls on between all possible pairs of nodes in the network. This proportion is weighted by the ratio between the cost of the shortest path connecting a pair of nodes and the cost of the *quasi*-shortest path between the same pair of nodes passing through v_k. The search for *quasi*-shortest paths is limited by a parameter ρ, which defines the maximum extra path cost that the proposed ρ-geodesic betweenness can take into account. We show in this thesis that a small ρ is enough to capture well the idea of *quasi*-shortest paths while keeping the computational load low. The metric proposed in this work can be used as part of the modelling of situations where the management of flows try to escape from the common-sense, aiming to avoid unwanted consequences that are expected to happen.

We evaluate the proposed metric by comparing it with other existing betweenness centrality metrics. We (i) verify if the metrics are capable of pinpointing nodes that should receive a different value for their centralities compared with the traditional betweenness. This means that the set of most central nodes can change from one metric to the other. The nodes that improve their position in the rank can be more suitable to be used, depending on the application. Then, we (ii) compare the coefficient of concordance of the rankings obtained for each metric and (iii) investigate if they can break ties between nodes classified in the same position. The goal is to verify if the metrics are measuring similar characteristics to determine the node ranking and to find if the metrics can provide a broadened ranking, which widens the number of options to choose from a more fine-grained node ranking. Following, we (iv) verify the influence of the parameter ρ on the variation of node positioning on the rank. Then we (v) analyze the behavior of the rank over time to verify the impact of the metric on the ability to intermediate paths. Finally, we study the performance of a dynamic network in presence of single failure. To this end, we (vi) evaluate the number of critical nodes elected as the most central by each centrality metric. Then, we consider a shortest-path-based packet forwarding strategy to (vii) analyze the impact on the network throughput when a failure happens on nodes classified by each metric as the most central, when flows follow shortest paths.

1.2 Contributions

The contributions of this work are summarized hereafter.

1. We study the effect of network structure and node mobility on multihop communications and on the evaluation of node centrality. We identify the importance of quantifying the influence of relative speeds on communications in multihop networks, and the need for a betweenness metric that better captures the importance of nodes that participate on paths slightly longer than the shortest one.

2. We propose an extended definition of node vicinity to include both nodes at multihop distances from an ego node (central node in the vicinity) and their relative speeds, instead of only focusing on the node adjacent vicinity. Moreover, we propose the ρ-geodesic betweenness, a weighted centrality metric that better evaluates the importance of nodes that do not necessarily fall on shortest paths but frequently participate in paths almost as short as the shortest ones;

3. We characterize the extended vicinity in three distinct scenarios, to analyze its behavior under distinct conditions, such as varied node density, and we analyze the usefulness of a contact opportunity to transfer large files, which could be required by entertainment applications in VANETs. Then, we demonstrate through simulations that we can potentially reduce network resource consumption, without reducing the average packet delivery ratio, using the relation between relative speeds, hop distance, and contact duration to make forwarding decisions;

4. We characterize the proposed ρ-geodesic betweenness centrality and we analyze the connectivity of a dynamic network through a comparative investigation, where we verify the number of critical nodes in central positions and the impact on network throughput when central nodes fail and flows follow shortest paths.

These contributions are reported in the following papers, in order of publication:

- "Uma avaliação da Influência da Velocidade dos Nós no Estabelecimento de Caminhos em Redes Ad Hoc Veiculares", accepted in the Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2015);

- "Intermediação por Espalhamento: Caminhos Quase Mais Curtos Também Importam", accepted in the Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2016);

- "Weighted Betweenness for Multipath Networks", accepted in the Global Information Infrastructure and Networking Symposium (GIIS 2016);
- "Eficiência dos Caminhos Quase Mais Curtos em Redes Dinâmicas", accepted in the Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2017);
- "The Power of Quasi-Shortest Paths: ρ-Geodesic Betweenness", accepted in the IEEE Transactions on Network Science and Engineering (TNSE 2017);

In the case scenarios studied in this work, results considering different radio ranges confirm that, indeed, the best contact opportunities happen at few hop distances and low relative speeds. Nevertheless, we found that a significant number of useful contacts can happen even between nodes at high relative speeds, separated by multihop distances. Even in such conditions, we show that nodes can transfer MB-size messages according to the contact duration. Besides more general results, we also observe that contacts with longer duration become less frequent for relative speeds higher than 40 km/h and most likely happen between nodes less than 3 hops away, in sparser scenarios. On the other hand, even considering lower relative speeds, results show that contacts between nodes separated by more than 6 hops are not frequent. We also note that high relative speeds can potentially degrade the number of useful contacts more severely than the hop distance. Finally, comparing our forwarding strategies with the Optimized Link State Routing Protocol (`OLSR`) [44], we show that it is possible to reduce the waste of resources, without decreasing the average packet delivery ratio, if we restrict multihop communications considering the relation between the reachability of nodes and their relative speeds. This happens even when the forwarding decision only takes into account local information. Although the `OLSR` is not the most suitable routing protocol for VANETs, we use it because in the current state of the work, we need a routing protocol for wireless networks that has information about the global network structure.

The vicinity of a node must also include relative speeds both in more theoretical evaluations and in practical settings [21], and the results obtained herein can be used as a step forward to develop more sophisticated message dissemination schemes in vehicular networks. All these results are discussed in our technical report "Impact of Relative Speed on Node Vicinity Dynamics in VANETs" [45], submitted to the Wireless Networks Journal (Springer).

The comparisons between our proposed centrality metric and other related metrics showed that the ρ-geodesic betweenness can rerank several nodes, even though it is strongly correlated to the traditional definition of betweenness, already using low values for the spreadness factor ρ. It is also useful to provide a wider range

of rank positions, presenting a more fine-grained classification. Yet, the ρ-geodesic betweenness reduces the number of resources reallocations, when node centrality is used to place such resources. In addition, our metric is able to keep nodes on the same rank position for longer time spans in networks with dynamic topology. The number of articulation points elected by the ρ-geodesic betweenness as the most central nodes are always less or equal than the number elected by other betweenness centralities. In a network where flows follow shortest paths, the throughput suffers a great reduction when a central node fails, which is similar to all the metrics. Even though the throughput is reduced, the probability that the network is split into several connected components is also reduced when failures happen on the most central nodes elected by our metric.

1.3 Organization

We organize this work as follows. We first introduce in Chapter 2 some definitions necessary to lay the basis of our work. Chapter 3 describes and characterizes the datasets used in this thesis. We then proceed to our first analysis, in Chapter 4, where we propose an extension of node vicinity and a methodology to analyze it. Chapter 5 discusses the results of our vicinity study, including the analysis of three proposed forwarding schemes based on our results. Following we begin our second analysis, starting with Chapter 6, where we propose a novel weighted betweenness centrality metric. Chapter 7 presents the characterization of the proposed metric and discusses some possible applications. Finally, Chapter 8 concludes this work and presents future research directions.

Chapter 2

Network Model, Concepts and Definitions

In this chapter, we explain the network model and we formalize the main definitions and concepts necessary to establish the basis of this work.

2.1 Network model

We consider that networks can be modelled as weighted graphs, $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$. The set of vertices $\mathcal{V}$ contais all nodes in the network, and the set of edges $\mathcal{E}$ comprises all links between nodes in $\mathcal{V}$. Each edge has a cost that belongs to the set of weights $\mathcal{W}$. Two neighbor nodes υ_i and υ_j are connected by an edge $\varepsilon_{i,j}$ whose cost is $\omega_{i,j} \in \mathbb{R}^*_+$. The edge $\varepsilon_{j,i}$ automatically exists if the graph is undirected (or symmetric). Otherwise, it will exist only if υ_j is also neighbor of υ_i, in which case we say the graph is directed (or asymmetric). The graph $\mathcal{G}$ is connected if all nodes in $\mathcal{V}$ are reachable, and not connected, otherwise. If all nodes are reachable, it is certain that there exists an edge between each and every pair of adjacent nodes. Connectivity is further discussed in Section 2.4.

We further consider that each node υ_i can move at speed $\vec{s_i}$, where $\{|\vec{s_i}| \in \mathbb{R}^+ \mid s_{min} \leq |\vec{s_i}| < s_{max}\}$, and s_{min} and s_{max} are the minimum and the maximum absolute speeds allowed, respectively. The relative speed of nodes υ_i and υ_j is, thus, given by $\vec{r_{i,j}} = \vec{s_i} - \vec{s_j}$, where $|\vec{r_{i,j}}| = |\vec{s_i} - \vec{s_j}|$ and $|\vec{r_{i,j}}| \in [0, 2 \times s_{max}]$. For the sake of simplicity, we use the notation s_i and $r_{i,j}$ to represent, respectively, $|\vec{s_i}|$ and $|\vec{r_{i,j}}|$, whenever possible.

We can divide the set of all relative speeds into m consecutive subsets, in which each relative speed $r_{i,j} \in \bigcup_{\kappa=0}^{m-1} R_\kappa$, $\forall \upsilon_i, \upsilon_j \in \mathcal{V}$, where $R_\kappa = [\kappa \times s_\delta, (\kappa + 1) \times s_\delta[$ and $s_\delta = \frac{2 \times s_{max}}{m}$. In this case, we can group all pairs of nodes υ_i, υ_j with $r_{i,j} \in R_\kappa$ in a subset of nodes $\mathcal{V}(R_\kappa) \subseteq \mathcal{V}$. Consequently, we can obtain the subgraph

$\mathcal{G}_{R_\kappa}(\mathcal{V}(R_\kappa), \mathcal{E}(R_\kappa)) \subseteq \mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{E}(R_\kappa)$ is the set of existing links connecting adjacent nodes in $\mathcal{G}_{R_\kappa}$ with $r_{i,j} \in R_\kappa$. Thus, although two adjacent nodes υ_i, υ_j may be included in $\mathcal{V}(R_\kappa)$ due to their relative speed, a link connecting them will only exist in $\mathcal{E}(R_\kappa)$ if $r_{i,j} \in R_\kappa$ and they are within mutual radio range. Otherwise, υ_i, υ_j may still be mutually reachable if they are interconnected by *a sequence of adjacent links* between pairs of nodes also in $\mathcal{V}(R_\kappa)$. Hence, according to our definition, if $\upsilon_i, \upsilon_k, \upsilon_w, \upsilon_j$ are in $\mathcal{V}(R_\kappa)$, and if $\{r_{i,k}, r_{w,j}\} \in R_\kappa$, then there is a link between υ_i, υ_k and another between υ_w, υ_j. A path from υ_i to υ_j will exist in $\mathcal{G}_{R_\kappa}$ only if $r_{k,w} \in R_\kappa$.

As a corollary, if $m = 1$, all relative speeds are within the same subset $R_0 = [0, s_\delta[= [0, 2 \times s_{max}[$. Analogously to relative speeds, we divide the set of absolute speeds into consecutive subsets, represented by S_κ.

2.2 Paths and costs

A *path* $p_{1,L}$ between source υ_1 and destination υ_L is an ordered sequence of distinct nodes in which any consecutive pair of nodes is connected by a link. A path does not contain any loops and any change in the sequence of nodes, either by switching or by shifting a node, originates a new path. We denote the length of path $p_{1,L}$ as $\lambda_{1,L} = L - 1$, with $L \in \mathbb{N}^*$. The cost of this path is denoted by $\delta_{1,L}$, with $\delta_{1,L} \in \mathbb{R}^*_+$, and it is given by the sum of the individual costs of all links composing the path.

The *shortest path* $p^*_{1,L}$ between υ_1 and υ_L is the one with the smallest cost, denoted by $\delta^*_{1,L}$. This path is also known in the literature as the *least cost path*. In this work, we use both *shortest path* and *least cost path* interchangeably. We also consider, without loss of generality, the number of hops as the cost of a path, such that $\delta_{1,L} = \lambda_{1,L}$ and, as a consequence, $\delta_{1,L} \in \mathbb{N}^*$. In this case, the cost of the shortest path is given by $\delta^*_{1,L} = \lambda^*_{1,L}$. Note that more than one shortest path (geodesic) may exist between the same pair of nodes. We denote the number of shortest paths between υ_i, υ_j as $n^*_{i,j}$. Yet, we denote the number of shortest paths between υ_i, υ_j passing through υ_k as $n^*_{i,j}(\upsilon_k)$.

2.3 Taking nodes on quasi-shortest-path into account

In some networks, flows do not follow shortest paths. In other networks, it is interesting to have other paths that are a little bit longer. In a computer network, for instance, the use of shortest paths can lead to the overload of nodes on such paths. We could reduce this load by splitting the flow between alternative paths. Ideally, these paths will be as short as the shortest one, and we will be able to increase the

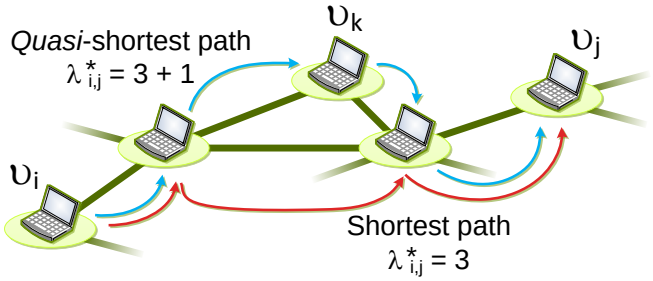


Figure 2.1: The shortest path between v_i and v_j is $\lambda^*_{i,j} = 3$ hops long. If $\rho = 1$, the *quasi*-shortest path of length $\lambda^*_{i,j} = 4$ through v_k can be considered too.

overall communication throughput and end-to-end cost. Nevertheless, occasionally, another shortest path is not available, but many other paths are. We could, then, use a *slightly longer* path to reduce the load on the shortest path, at the cost of a *small* increase on the end-to-end cost. Very long paths, however, have low or none contribution to the network operation and should not be used [41]. Hence, we introduce two conjugated concepts to consider such important alternative paths when analyzing a network.

Definition 1. Spreadness: *The spreadness ρ is the maximum tolerable difference between the costs $\delta_{1,L}$ and $\delta^*_{1,L}$, i.e., $\rho = \delta_{1,L} - \delta^*_{1,L}$, with $\rho \in \mathbb{R}_+$.*

Definition 2. *Quasi*-shortest path: *The quasi-shortest path is a path $p_{1,L}$ for which $\delta_{1,L} - \delta^*_{1,L} \leq \rho$, where ρ is the spreadness factor.*

The *quasi*-shortest path is the most important concept of this work. The idea behind it is illustrated in Figure 2.1, where $\rho = 1$. Such *quasi*-shortest paths are able to increase the importance of nodes that are ignored or underestimated when we consider only the shortest paths – this is the case, for example, of node v_k (that does not fall on any shortest paths). Nevertheless, this node is very close to all shortest paths between both sides of the network, as represented by nodes v_i and v_j, respectively. Paths going through v_k differ from the shortest path by only one hop. Note that more than one *quasi*-shortest path with the same cost can exist between two nodes and more than one of these paths can pass through the same intermediary node. Therefore, we represent the number of *quasi*-shortest paths between v_i, v_j as $n_{i,j}$ and, among those, the ones passing through v_k as $n_{i,j}(v_k)$.

The spreadness ρ defines the extra cost we can add to the shortest path and, as a consequence, it determines the maximum cost of the *quasi*-shortest path. This limitation avoids the explosion of the number of possible paths. Although we defined $\rho \in \mathbb{R}_+$, in this work we consider the number of hops as cost metric and, thus, $\rho \in \mathbb{N}$. The spreadness limits the search depth to look only for *quasi*-shortest paths that are slightly longer than the shortest path. The idea is based on the fact that the

throughput of information traveling through paths for which $\delta_{1,L} \gg \delta^*_{1,L}$ is expected to be low [40]. Note that if $\rho = 0$, $\delta_{1,L} = \delta^*_{1,L}$, and only the shortest paths are considered.

2.4 Connectivity and articulation points

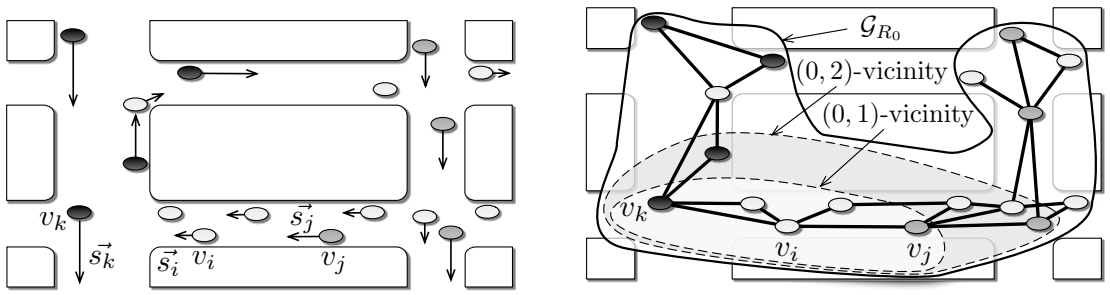
The use of graphs can reveal interesting properties from networks, such as network connectivity. A network is connected if there is at least one path between all pairs of nodes, and it is bi-connected if there are at least two node-disjoint paths between all pairs of nodes. The path redundancy present in bi-connected networks excludes the possibility of finding nodes that can split the network into one or more connected components in case of failure. Nodes that can potentially disconnect the network are said to be articulation points and they represent critical vulnerabilities. Formally, v_a is an articulation point if there exist two nodes $v_i, v_j \in \mathcal{V}$ with $v_i \neq v_j \neq v_a$ and $v_i \neq v_a$, such that v_a is part of all paths $p_{i,j}$.

2.5 Node vicinity

The typical vicinity of node v_i is composed of directly reachable nodes, i.e., all nodes $v_j \in \mathcal{V}$ within mutual radio range of v_i. In this work, we refer to the central node of the vicinity, v_i, also as the "ego node". When any node v_j is within mutual range with v_i, the link $\varepsilon_{i,j}$ exists and we say that nodes v_i, v_j are in contact. Hence, all nodes v_j in v_i's vicinity are in contact with v_i. Using the typical definition of node vicinity, a fraction of nodes can remain nearby the ego node without ever entering mutual radio range. As a consequence, the ego has a limited view of its contact opportunities. Additionally, nodes can frequently enter and exit the mutual radio range, which incurs several vicinity changes over time. Phe-neau et al. [31, 32] extend the concept of contact to consider also nodes reachable via multiple hops. Consequently, they also extend the concept of vicinity, incorporating nodes even if they are out of mutual radio range. Therefore, nodes can potentially find more contact opportunities. The relative speed of nodes can greatly influence these opportunities, because it determines the link existence and contact duration [45], which is equal to the path duration. Hence, we extend the vicinity proposed by Phe-neau et al. to also consider the relative speed of nodes.

Definition 3. (κ, λ)**-vicinity**: *The* (κ, λ)*-vicinity of a node* $v_i \in \mathcal{G}_{R_\kappa}$ *is the set of all nodes also in* $\mathcal{G}_{R_\kappa}$ *for which the shortest path from* v_i *is* λ *hops long at most.*

The vicinity of a node $v_i \in \mathcal{G}_{R_\kappa}$ can be characterized only by parameters κ and λ, where λ defines the maximum number of hops from v_i, while κ defines the range of



(a) Vehicles v_i, v_j, and v_k moving at speeds, s_i, s_j, and s_k, respectively.

(b) $\mathcal{G}_{R_0}$, $(0,1)$- and $(0,2)$-vicinity of node v_i.

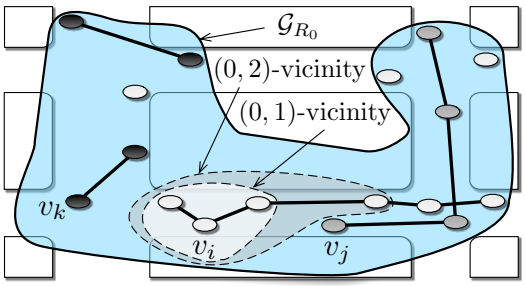
Figure 2.2: Definition of v_i's (κ, λ)-vicinity, disregarding vehicles relative speeds ($m = 1$). Subgraph $\mathcal{G}_{R_0}$ coincides with $\mathcal{G}$, including all nodes in the network.

relative speeds considered and, consequently, which subgraph must be used. Hence, nodes in $\mathcal{G}_{R_\kappa}$ may not belong to the same (κ, λ)-vicinity of v_i, according to the number of hops (λ) of the shortest path interconnecting them.

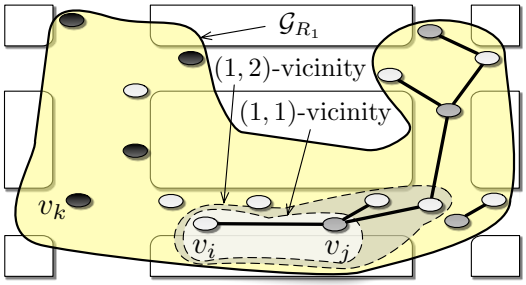
Figure 2.2(a) depicts a network of nodes moving at speeds within $[0, 45$ km/h] (arrows starting at nodes indicate their absolute speed). In this figure, nodes v_i, v_j, and v_k move, respectively, at absolute speeds s_i, s_j, and s_k, where $s_i \in [0, 15$ km/h[, $s_j \in [15, 30$ km/h[, and $s_k \in [30, 45$ km/h]. If $m = 1$, the relative speed between all pairs of nodes in the network lies within $R_0 = [0, 90$ km/h]. Thus, all nodes are within the $(0, \lambda)$-vicinity of v_i. Figure 2.2(b) shows v_i's $(0, 1)$- and $(0, 2)$-vicinity, and the subgraph $\mathcal{G}_{R_0}$ obtained from the subset $\mathcal{V}(R_0)$. Note that with $m = 1$, $\mathcal{V}(R_0) = \mathcal{V}$. Hence, all the links connecting nodes in the network do exist and can be used to compute paths. As a consequence, $\mathcal{G}_{R_0} = \mathcal{G}$ and the delimitation of v_i's vicinity does not change according to the different relative speeds, similarly to [32].

If we consider $m = 3$, we have three different subsets of relative speeds: $R_0 = [0, 30$ km/h[, $R_1 = [30, 60$ km/h[, and $R_2 = [60, 90$ km/h]. In this case, we can separate the pairs of nodes within $\mathcal{V}$ in subsets, according to their relative speeds: $\mathcal{V}(R_0)$, $\mathcal{V}(R_1)$, and $\mathcal{V}(R_2)$. From these subsets we obtain the subgraphs illustrated in Figure 2.3, $\mathcal{G}_{R_0}$, $\mathcal{G}_{R_1}$, and $\mathcal{G}_{R_2}$. Note that, we can compute the shortest paths to obtain the (κ, λ)-vicinity of node v_i only after finding $\mathcal{G}_{R_\kappa}$.

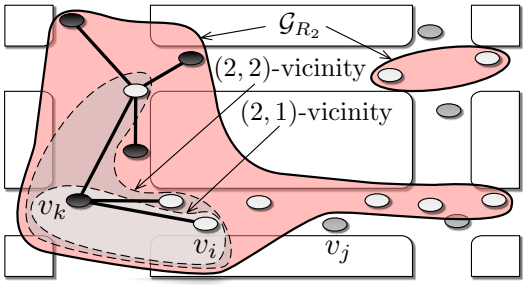
Figure 2.3(a) shows $\mathcal{G}_{R_0}$ and the $(0, 1)$- and $(0, 2)$-vicinity of v_i. We observe that although the $(0, 2)$-vicinity includes all nodes in the $(0, 1)$-vicinity, it does not include all nodes in $\mathcal{G}_{R_0}$. Therefore, v_i requires more than two hops to reach a node which is not in its $(0, 2)$-vicinity. In the worst case, no paths connecting v_i to these nodes exist in $\mathcal{G}_{R_0}$, which means that $\lambda \to \infty$ for the subset $\mathcal{V}(R_0)$. Figures 2.3(b) and 2.3(c) show, respectively, the $(1, 1)$- and $(1, 2)$-vicinity, and the $(2, 1)$- and $(2, 2)$-vicinity of node v_i, as well as $\mathcal{G}_{R_1}$ and $\mathcal{G}_{R_2}$.



(a) $\mathcal{G}_{R_0}$, and $(0,1)$- and $(0,2)$-vicinity.



(b) $\mathcal{G}_{R_1}$, and $(1,1)$- and $(1,2)$-vicinity.



(c) $\mathcal{G}_{R_2}$, and $(2,1)$- and $(2,2)$-vicinity.

Figure 2.3: Example of $\mathcal{G}_{R_\kappa}$ for nodes with relative speed in R_κ, the links connecting them, and the (κ, λ)-vicinity of node v_i.

2.6 State of a node

Each node v_i in direct contact with a peer node v_j in $\mathcal{G}_{R_\kappa}$ is considered to be in State σ, where $\sigma = 1$. If they need one more node $v_k \in \mathcal{G}_{R_\kappa}$ to reach each other, than v_i, v_j are in State 2. If no path exists between them in $\mathcal{G}_{R_\kappa}$, this pair of nodes is in State ∞, which only represents the absence of intermediary nodes in the same $\mathcal{G}_{R_\kappa}$ to set up a path between v_i and v_j. This does not necessarily mean that v_i and v_j are out of reach, because they can be in contact in another $\mathcal{G}_{R_\kappa}$. Hence, we can define the state of a node as follows.

Definition 4. State σ: The state of a node corresponds to the shortest hop distance connecting it to another node.

Note that, as the state of a node depends, in fact, of the distance to a peer node, each node can be in more than one state simultaneously, depending on how many peers it has.

The definition of node state is important to analyze transition probabilities between the states, according to the following model. We model the vicinity dynamics of a node pair as a continuous time Markovian process. This means that the current markovian state summarizes the past history of the process [46] and the transition to another state can happen at any instant of time. The memoryless aspect of this process implies that the duration of each state follows an exponential distribution and

each modification is independent of the sampled time. Thus, we maintain our analysis agnostic to the duration of a given state to be independent of the time sampling frequency of events. These properties are convenient since we are most interested in capturing the vicinity changes of nodes at different relative speeds. Hence, in our model, for a given pair of nodes υ_i, υ_j, the hop distance between them in a given epoch e is represented by a random variable $X^e_{i,j}$, which is stored in a State σ. Each pair υ_i, υ_j may change its state only once per epoch e and the number of states is equal to the maximum number of hops interconnecting a pair of nodes plus the State ∞. The current State σ of a node is independent of previous states. Consequently, if υ_i and υ_j are λ-hops distant in e, there is a probability that the distance between them will be d in $e+1$. Hence, we have that $p_{ab} = \mathbb{P}(X^{e+1}_{ij} : \sigma = b \mid X^e_{ij} : \sigma = a) \geq 0$.

2.7 Vicinity timeline

The vicinity timeline is a pairwise component of the vicinity analysis and it can be defined as follows.

Definition 5. (κ, λ)-vicinity timeline: The (κ, λ)-vicinity timeline of a pair of nodes in $\mathcal{G}_{R_\kappa}$ is the sequence of states for the pair of nodes in analysis, overtime.

Each entry in the vicinity timeline is an event represented by a tuple $\langle t_i, t_f, \upsilon_i, \upsilon_j, \lambda, r_{i,j} \rangle$, where t_i and t_f are the initial and final instants of time of the event, υ_i, υ_j is the pair of nodes, λ is the shortest hop distance between them, and $r_{i,j}$ is their relative speed. State ∞ is represented by $\lambda = 0$ in the (κ, λ)-vicinity timeline. Time intervals are atomic, i.e., there is no other event in the whole timeline starting or finishing at an instant of time t, where $t_i < t < t_f$. This is important to better understand concurrent events. The state transitions in the (κ, λ)-vicinity of the node is stored in such timelines, allowing to determine the state transition probabilities, which details how nodes move relative to each other.

Table 2.1 summarizes the notation described in this chapter, in order of appearance, to facilitate future reference.

Table 2.1: Summarized notation used in this work.

Notation	**Comment**
$\mathcal{V}$	Set of vertices
$\mathcal{E}$	Set of edges
$\omega_{i,j}$	Cost of edge between nodes i and j
υ_i	Node i
$\varepsilon_{i,j}$	Edge between nodes i and j
$\vec{s_i}$ or s_i	Absolute speed of node i
$\vec{r_{i,j}}$ or $r_{i,j}$	Relative speed between nodes i and j
R_κ	Range of relative speeds
κ	Index of relative speed range
$\mathcal{V}(R_\kappa)$	Set of nodes moving at relative speeds within R_κ
$\mathcal{E}(R_\kappa)$	Set of edges between nodes moving at relative speeds within R_κ
$\mathcal{G}_{R_\kappa}$	Subgraph of nodes moving at relative speeds within R_κ
S_κ	Range of absolute speeds
$p_{i,j}$	Path between nodes i and j
$\lambda_{i,j}$	Hop distance between two nodes, i.e., the path length
$\delta_{i,j}$	Cost of the path between nodes i an j
$p^*_{i,j}$	Shortest path between nodes i and j
$\delta^*_{i,j}$	Cost of the shortest path between nodes i and j
$\lambda^*_{i,j}$	Length of the shortest path between two nodes
$n^*_{i,j}$	Number of shortest paths between nodes i and j
$n^*_{i,j}(\upsilon_k)$	Number of shortest paths between nodes i and j passing through node k
ρ	Spreadness factor
$n_{i,j}$	Number of *quasi*-shortest paths between nodes i and j
$n_{i,j}(\upsilon_k)$	Number of *quasi*-shortest paths between nodes i and j passing through node k
σ	State of a node

Chapter 3

Datasets Description and Characterization

In this work, we use different datasets available on the Internet, as well as several randomly generated static networks. Three of the publicly available datasets represent dynamic networks, and they are used for the vicinity analysis proposed in Chapter 4. The remaining four are static networks and we use them to assess the centrality metric proposed in Chapter 6, as well as the randomly generated static networks and snapshots from one of the dynamic network datasets. We organize this chapter in three sections, where we discuss the datasets separately, according to the type of network they represent.

3.1 Dynamic network datasets

In our analyses, we use three different dynamic network datasets, which are summarized in Table 3.1: Mobility Dataset [47], Ad Hoc City Dataset [48], and TAPAS-Cologne Dataset [49]. They represent, respectively, medium, sparse and high density urban scenarios. Each one of them is generated by capturing vehicle mobility, which

Table 3.1: Dynamic network datasets main characteristics.

Feature	**Mobility Dataset** [47]	**Ad Hoc City Dataset** [48]	**TAPASCologne Project** [49]
Type of trace	Real	Real	Hybrid
Location method	GPS	GPS	–
Number of vehicles	536	1,200	121,140
Type of vehicle	Taxi	Bus	Car
Total duration	30 days	30 days	2 hours
Sampling frequency	1 sample / 10 seconds	525 samples / 1 day	1 sample / 1 second
Analyzed period	1 day	1 day	10 minutes
Location	San Francisco, CA – USA	Seattle, WA – USA	Cologne – Germany

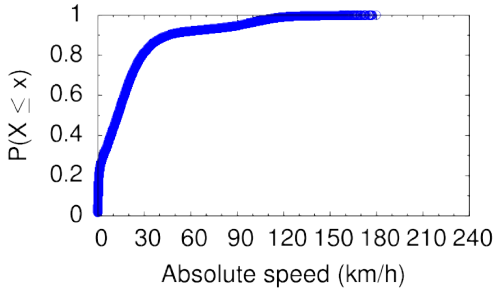
(a) Taxi scenario.

(b) Bus scenario.

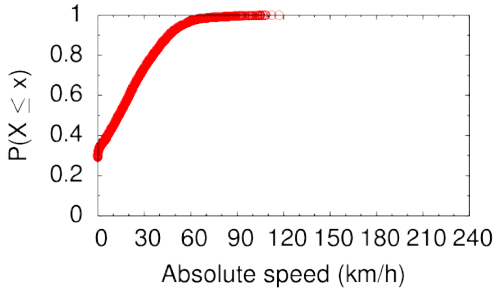
(c) Synthetic scenario.

Figure 3.1: Normalized frequency of updates in each scenario. We divide the cities in small areas of 100 $\times$ 100 m^2 and we analyze the sampled period of each dataset. We consider that areas where we can find frequent updates are also denser areas.

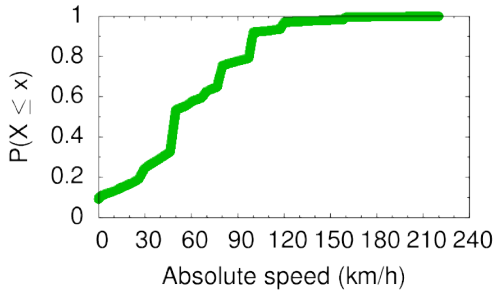
can be plotted to obtain the city map for each scenario, as shown in Figure 3.1. This figure represents, in fact, a heatmap of the normalized frequency of updates in small areas of 100 $\times$ 100 m^2 for each dataset. We normalize the values by the highest frequency in each scenario, considering the sampled period. The color range shows that the more frequent the updates, the more dark red the area. Frequent updates can happen in an area due to the number of updates sent by a set of vehicles or due to the number of vehicles sending updates. Knowing that vehicles periodically send updates, we consider that an area with frequent updates much likely has several nodes. Hence, we can infer that the frequency of updates is directly related to the density of vehicles in the area. It is important to observe that each dataset represents distinct scenarios, where different types of vehicles move across the cities, with a diversified range of absolute speeds. To characterize the scenarios in relation to these speeds, we plot the Cumulative Distribution Function (CDF) of absolute speeds for each dataset, as shown in Figure 3.2. We discuss the characterization of each scenario in the following subsections.



(a) Taxi scenario.



(b) Bus scenario.



(c) Synthetic scenario.

Figure 3.2: Cumulative distribution function of absolute speeds for each scenario. (c) Cars in the synthetic scenario register the highest absolute speeds, followed by the (a) taxis and the (b) buses. The latter registers the highest number of very low absolute speeds and the most uniform distribution among the analyzed datasets.

3.1.1 Mobility Dataset: Taxi scenario

The Mobility Dataset [47] represents the movement of 536 taxis in San Francisco, California – USA, over 30 days. In fact, the dataset also involves part of the outskirts of San Francisco. Taxi location is obtained through GPS and updates are sent each 10 seconds. Update events are represented by a tuple $\langle lat, long, o\_flag, t\rangle$, where lat and $long$ are the latitude and longitude of the vehicle, o_flag is the occupancy flag for the taxi (not used in this work), and t is the event time. Note that the existence of periodic updates does not necessarily mean that every taxi send its location at the end of each period. From the provided 30-day dataset, we analyze 1 day.

Figure 3.1(a) shows the normalized frequency of updates for the Taxi scenario, considering the analyzed 1-day trace. We observe that, although taxis do not have predefined routes or time schedules, updates are more frequent in the northern area of the city, which is the city center, according to Google Maps. We also note that a great number of updates are sent by vehicles traveling through a specific route to the south, which leads to the city airport, according to Google Maps. We observe some very straight lines in this figure that do not match any existing roads in reality. This happens as a consequence of lack of periodic updates for each and every taxi in the scenario, which leads to approximation errors related to the taxi position, when computing the absolute speeds.

Figure 3.2(a) shows the CDF of absolute speeds for the Taxi scenario. We observe that 1% of the absolute speeds is equal to 0 km/h and almost 30% of the absolute speeds lie within $[0, 4$ km/h$]$. Such percentage of very low absolute speeds can be a consequence of waiting for passengers at taxi stands, in addition to stops due to traffic lights and street intersections. Traffic jam also contributes to this percentage and its presence in this scenario is highly plausible, as approximately 90% of the registered absolute speeds lie within $[0, 45$ km/h$]$.

3.1.2 Ad Hoc City Dataset: Bus scenario

The Ad Hoc City Dataset [48] registers the mobility of the fleet of city buses in Seattle, WA – USA, over 30 days. GPS devices are embedded in 1,200 buses and location information is updated 525 times per day for each bus. Events are represented by tuples of the type $\langle d, t, \upsilon_i, rt, x, y\rangle$, where d is the day for that event, t is the event time, υ_i is the bus identification, rt is the route followed by the vehicle and x, y are the Cartesian coordinates for the vehicle position. Each bus sends consecutive updates at different time intervals. In average, we should expect one update for each bus at, approximately, each 3 min. The following characterization refers to 1 day from the 30-day dataset.

In the Bus scenario, we expect to find less vehicles distributed across the city

when compared to the Taxi scenario. We further expect that buses have predefined routes and time schedules. Similarly to the Taxi scenario, we plot the heatmap for the frequency of updates in each region of 100×100 m^2. Figure 3.1(b) shows the resulting map. We observe that frequent updates are concentrated on a small region, which is the city center, according to Google Maps, where we believe the bus density is higher.

Figure 3.2(b) shows the CDF of absolute speeds for the chosen day. We observe that approximately 30% of the absolute speeds in this scenario are equal to 0 km/h, indicating that buses stop more than taxis. Indeed, this is expected because, besides the influence of traffic jams, the number of bus stops in the city is usually higher than the number of taxi stands. In addition, the time spent to pick up passengers at bus stations is usually higher than at taxi stands. Yet, the number of buses parked at the bus garage but that continue to send location updates can also influence the CDF, resulting in a great percentage of null absolute speeds. Similarly to the Taxi scenario, 90% of the registered absolute speeds lie within $[0, 45$ km/h$]$.

3.1.3 TAPASCologne Dataset: Cologne synthetic scenario

The TAPASCologne Dataset [49] was produced by the Institute of Transportation Systems at the German Aerospace Center (ITS-DLR). The goal is to model the car traffic in Cologne city, Germany, with the highest possible level of accuracy compared to the real traffic. The dataset is a hybrid model, built with a set of tools to simulate vehicular mobility, such as the software Simulation of Urban Mobility (SUMO) and the Travel and Activity Patterns Simulation (TAPAS) methodology, among others. Location updates are sent each 1 second, but not by every single car. Each update event is represented by a tuple $\langle t, v_i, x, y, s_i \rangle$, where t is the event time, v_i is the vehicle identification, x, y are the Cartesian coordinates for the position, and s_i is the absolute speed. In this work we refer to this scenario, interchangeably, as Synthetic or Cologne scenario.

The complete dataset covers an area of approximately 400 km^2 and comprises more than 700,000 individual trips of regular people cars during a 24-hour period. Routes and time schedules are not predefined, although they usually follow a pattern for each person, and very high absolute speeds are registered due to the presence of highways crossing the city. At the time the analysis in this work was carried out, the TAPASCologne project provided a 2-hour subset of the dataset. We divide this dataset in smaller subsets to find the one with the least number of vehicles. The goal is to ensure timely convergence of our analyses. Hence, we select the first 10 minutes, with almost 9,000 vehicles. Analogously to the Taxi and Bus scenarios, in Figure 3.1(c) we plot the frequency of updates for the Synthetic scenario, and we

observe that updates are frequent in several areas of the city.

Figure 3.2(c) shows the CDF of absolute speeds in the 10-minute subset trace. Absolute speeds are more distributed compared to the other traces and we find significant number of registers with high absolute speed. Approximately 10% of absolute speeds are equal to 0 km/h and 30% lie within $[0, 40$ km/h$]$. Further, 90% of the registered speeds are under 100 km/h. This indicates that cars tend to move faster than buses and taxis, which is expected. The presence of a wide range of absolute speeds in this dataset, achieving speeds higher than 200 km/h, is probably a consequence of the coexistence, in the same scenario, of roads and streets with different achievable speeds, including unlimited speed highways (autobahns).

3.2 Static network datasets

In addition to the dynamic datasets, we use snapshots from the Cologne synthetic dataset, and three more static datasets derived from social networks. All of them have distinct characteristics and they are summarized in Table 3.2. Figure 3.3 shows the graph obtained from each static dataset, and one snapshot sample from the Cologne dataset. In this illustration, the importance of the node is depicted according to its topological position. The scheme goes as follows: smaller nodes have smaller traditional betweenness, hence, the less flows they intermediate using shortest paths (geodesics); more reddish nodes have lower degree, hence, they have fewer neighbors; more bluish nodes have higher degree, hence, they have more neighbors. We further use several randomly generated static networks, which are discussed in a separate section at the end of this chapter.

3.2.1 Freeman's EIES

The Freeman's EIES dataset presents the communication relationships between people in a group of 32 academics [50] interested in interdisciplinary research. The data consists of all messages sent plus acquaintance relationships. The graph of relationships provided by this dataset is shown in Figure 3.3(a). A directed edge between

Table 3.2: Static network datasets main characteristics.

Feature	**Freeman's EIES** [50]	**Doubtful Sound Dolphins** [51]	**PhD Students** [52]	**Cologne snapshots** [49]
Number of nodes	32	62	1,025	1,584–1,916
Number of edges	460	159	1,043	1,573–2,044
Symmetry	Asymmetric	Symmetric	Asymmetric	Symmetric
Density	0.464	0.084	0.001	0.001
Number of samples	1	1	1	10

two nodes $[v_i, v_j]$ exists only if v_i has sent a message to v_j, totaling 460 links, with a density of 0.464. Note that only few nodes in this dataset have high traditional betweenness, and they are also the nodes with highest degree (large bluish nodes). Although we can also find few nodes with high degree and low traditional betweenness (small bluish nodes), degree and traditional betweenness are closely related, i.e., nodes with high traditional betweenness tend to have high degree.

3.2.2 Doubtful Sound Dolphins

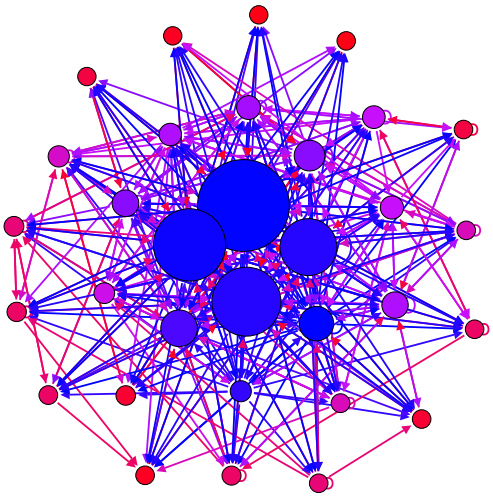
The Dolphins dataset provides the association relationship between 62 dolphins in Doubtful Sound, New Zealand [51]. Each node corresponds to a dolphin and the interaction between them is represented by an undirected edge, totaling 159 links. The density of this network is 0.084 and the graph provided by the frequent associations between dolphins is shown in Figure 3.3(b). In this figure, we observe several nodes with high traditional betweenness and average degree (large purplish nodes).

3.2.3 PhD Students

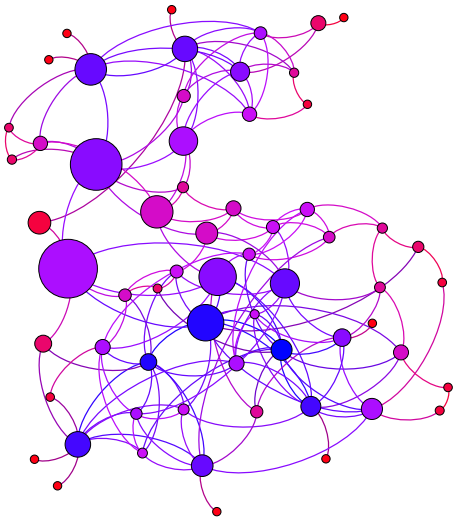
The PhD Students dataset is a very low density network (0.001) representing the relationships between 1,025 PhD students and advisors [52]. This is a directed network, where a link exists from v_i to v_j only if v_i is the supervisor of v_j, totaling 1,043 links. The graph obtained from these relationships is shown in Figure 3.3(c). We observe that this network has a peculiar structure, where many nodes behave as islands (i.e., roots) to which many other leaf nodes are attached. This is an important characteristic that must be remembered when analyzing the centrality of nodes in this network. We can spot a single node with high degree (bluish node), and the majority of nodes have low traditional betweenness (small nodes).

3.2.4 Snapshots from the TAPASCologne dataset

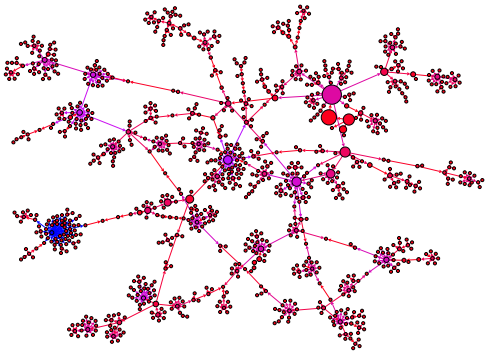
We use 10 samples of the original dataset presented in Subsection 3.1.3, containing from 1,584 to 1,916 nodes and from 1,573 to 2,044 undirected links, depending on the snapshot sample. Snapshots with more nodes do not necessarily have more edges, and vice-versa. Each node in this network represents a vehicle and an edge exists between them if they are less than 50 meters away from each other. The density of all samples is 0.001 and one of the sample graphs obtained from this dataset is shown in Figure 3.3(d). We observe, in this figure, that the majority of nodes are put together in small groups, in which nodes have low traditional betweenness and low degree (small reddish nodes). We also note other larger components, among which



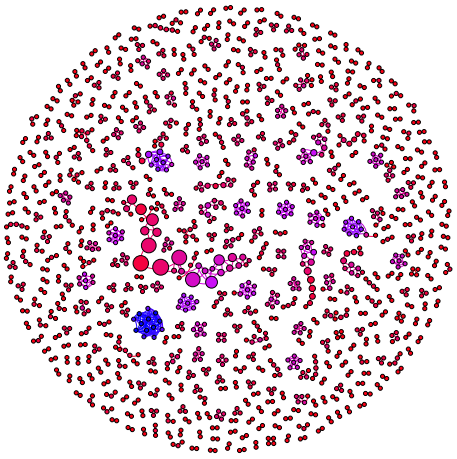
(a) Freeman.



(b) Dolphins.



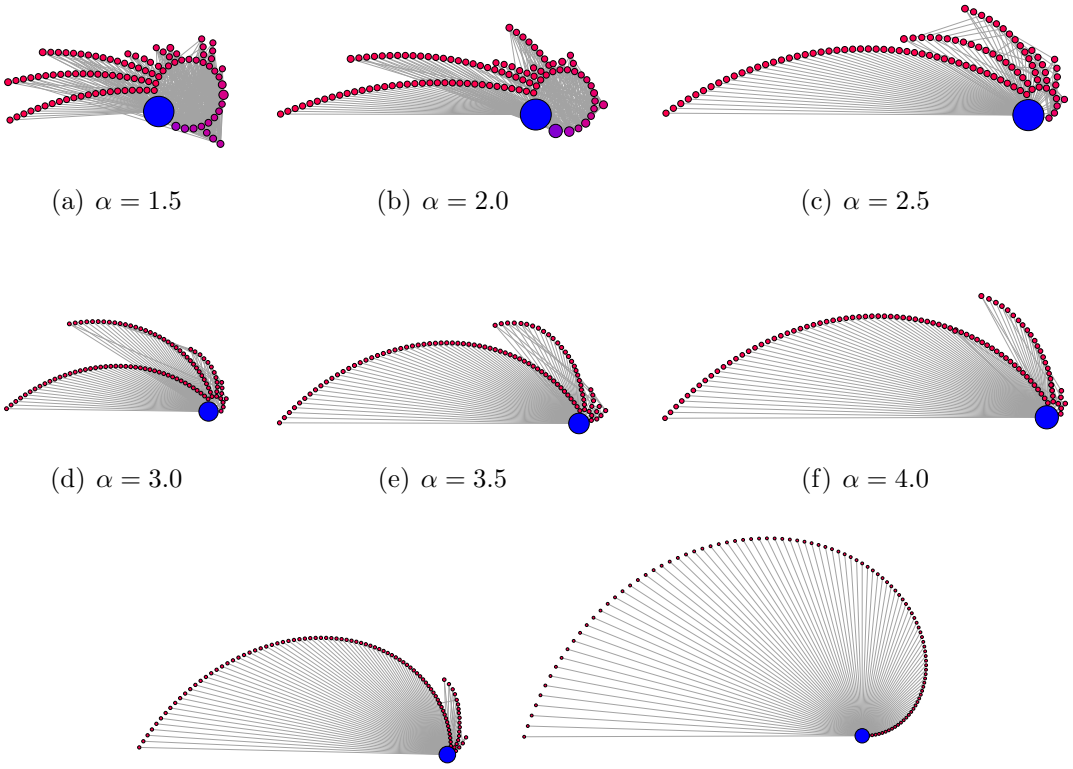
(c) PhD Sutdents.



(d) Cologne #1.

Figure 3.3: The visualization of the network topology provided by each dataset highlights the differences between them. The traditional betweenness is represented in the node size while the degree, in the node color, such that larger nodes have higher traditional betweenness and more bluish nodes have higher degree.

we can easily spot the giant component and few very well connected components. The giant component is composed by several nodes with high betweenness and low degree (large reddish nodes), whereas the well-connected components are composed by nodes with high degree and low betweenness (bluish small nodes).



(a) $\alpha = 1.5$ (b) $\alpha = 2.0$ (c) $\alpha = 2.5$

(d) $\alpha = 3.0$ (e) $\alpha = 3.5$ (f) $\alpha = 4.0$

(g) $\alpha = 4.5$ (h) Star, $|V| = 100$

Figure 3.4: Comparison between a star network and sample random networks with power law degree distribution for different α. All networks have 100 nodes and it is clear that the structure of the network changes with α, becoming more similar to a star as α increases.

3.2.5 Randomly generated static network datasets

In addition to the datasets described in Section 3.2, we use several random static networks. The randomly generated networks share two characteristics: they all have 100 nodes and they are scale-free, i.e., they follow, approximately, a power-law degree distribution ($P \propto degree(v_i)^{-\alpha}$). We choose such distribution because it is widely known in the literature that most real networks are scale-free [53].

We use Python's NetworkX 1.10 module to generate the random scale-free static networks. This module implements several graph generators, among which we find the Barabási-Albert [54], Holme-Kim [55], and Havel-Hakimi [56, 57] algorithms. The first one generates scale-free graphs with scaling factor $\alpha \approx 3$ [53]. The second is essentially an extension of the Barabási-Albert algorithm with a tunable clustering coefficient that allows achieving higher clustering. The third is able to construct simple graphs given a valid degree sequence. Hence, if the input is a power-law sequence, the resulting simple graph will be scale-free.

We want to use scale-free networks with several scaling factors. Thus, we can-

not use the Barabási-Albert nor the Holme-Kim algorithms, because they produce networks with α very close to 3. In addition, we want connected simple graphs, i.e., networks where there is at least one path between any pair of nodes and that do not have multiple edges between adjacent nodes. Hence, the Havel-Hakimi algorithm perfectly fits our purpose. We arbitrarily choose a number of nodes equal to 100 and, using the Havel-Hakimi algorithm, we timely generate 326 undirected and connected simple graphs with scaling factor $1.5 \leq \alpha \leq 4.9$ (step of 0.1). Outside this range, the algorithm was not able to timely generate graphs that comply with our restrictions: simple, connected, and undirected.

We tried to generate at least 10 graphs for each α but, for some scaling factors, it takes too long to find random graphs that comply with our restrictions. Hence, we generated 10 graphs for each scaling factor within 1.5 and 4.6, and only 2 graphs for each $\alpha \in \{4.7, 4.8, 4.9\}$. Note that increasing α means that many more nodes will have very low degree. This is the reason why it is hard to find *connected* simple graphs for $\alpha \geq 4.7$. Particularly, the Havel-Hakimi algorithm originates star-like graphs for higher α. This effect is shown in Figure 3.4, where the corresponding star graph is depicted in Figure 3.4(h) for comparison. In Figure 3.4, nodes are colored by degree, and node size represents the traditional betweenness, similarly to Figure 3.3. Nodes are also grouped by degree in each radial axis. The nodes with highest traditional betweenness in these graphs also have high degree (large bluish nodes). Even though we did not achieve a very wide range of scaling factors, we can use the obtained networks in our analysis without loss of realism. This is possible because researchers claim that for most real networks α falls approximately between 2 and 3 [53, 58]. Hence, in the following chapters we only show the results obtained this range of scaling factor.

Part I

Multihop Communications in Dynamic Networks

Chapter 4

Multihop Network Connectivity

This chapter discusses the problematic of node mobility on the establishment of multihop paths on wireless dynamic networks, presenting related works. It also describes the analysis proposed to investigate node vicinity behavior, aiming to maximize the exploitation of contact opportunities.

4.1 Background

The characterization of mobility patterns in wireless networks remains an open research issue. This is particularly true for challenged networks, where contact information is not known *a priori* and there is no infrastructure to provide connectivity. A good example of such scenario is Vehicular Ad Hoc Networks (VANETs) [59], especially if they rely only on Vehicle-to-Vehicle (V2V) communications. VANETs are highly dynamic and the intense node mobility contributes to the intermittent connectivity and lack of end-to-end paths. This can hinder communications during contact opportunities, increasing the difficulty to achieve efficient data transfer [3]. Several studies concerning mobility patterns and connectivity in VANETs already exist and many of them are compiled in a large number of surveys [18–23]. These works are valuable to provide better insights on routing protocol development for message dissemination in networks with intermittent connectivity.

The main premise of routing protocols designed for static wireless networks is that lack of end-to-end paths are transitory. In dynamic networks, where nodes present intense mobility, this assumption may not be valid, as the probability of not having a connected network is high. In such environments, it is important to know when contact opportunities happen, how long they last and how large the available bandwidth during contact is. Routing protocols designed for dynamic networks must consider that end-to-end paths are, in fact, rare. The knowledge of each node about its surroundings must be used at each encounter to determine what action the node should perform with the packet, for instance, it could instantly

forward it or drop it if no neighbors are within its radio range. Protocols do not have information, a priori, about contacts and there is no infrastructure to provide connectivity. To make better decisions and maximize the exploitation of contact opportunities, protocols rely on node vicinity [27–29] and mobility history [60, 61]. This increases the probability of successfully forwarding a message to the destination. Nodes for which this probability is high are said to have high utility [62]. The relative speed of nodes could be used in the computation of this metric. For instance, if the relative speed between v_i and v_j is too high, v_i should not try to forward a message through v_j. Hence the overhead on the network would be reduced, becauses useless packets would not be forwarded.

Focusing on node vicinity, typically, two nodes are considered to be in *contact* if they are within mutual radio range, limiting node vicinity to directly reachable neighbors. Nevertheless, depending on the network topology, a significant fraction of nodes can remain in the nearby vicinity of a node, within a few hops, when not in direct contact. Therefore, the typical definition of contact incurs a limited view of contact opportunities. As an alternative to circumvent this restriction and better exploit the (potentially) few opportunities, some works propose complex prediction mechanisms to anticipate contact availability [5] and disruption [6]. These works consider that contacts happen at 1-hop distance, leaving aside several contact opportunities in the nearby neighborhood.

To exploit more contact opportunities, Phe-Neau et al. [31, 32] propose to extend the concept of *contact* to also include nodes reachable via multiple hops. Hence, the resulting extended vicinity incorporates nodes even if they are out of mutual radio range, increasing the number of contact opportunities. Hoque et al. [63] use the idea of extended vicinity to develop an algorithm to analyze multihop connectivity and network partitioning in VANETs. Forwarding strategies can be tuned to exploit such opportunities by sending Hello packets (beacons) to discover the nearby nodes forming the neighborhood. The use of Hello packets introduces an overhead that depends on the size of the Hello message, the number of neighbors and the frequency of updates. Nevertheless, nodes can still benefit from multihop contact opportunities, even if the control overhead increases, because they do not need to wait for a direct contact to establish communication with other nodes. Hence, the use of such multihop contact opportunities can even reduce end-to-end delays.

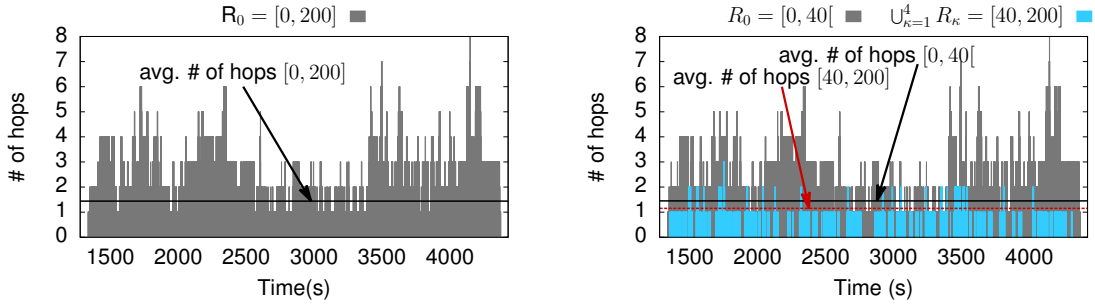
None of the works mentioned in this section studies the influence of relative speeds on path establishment. This quantity is important to better understand node mobility and maximize contacts exploitation. Shelly et al. [61], for instance, use the relative speed of nodes to predict the residual lifetime of a link, using the node attached to the longest-lasting link to forward the message. This reduces the probability of link breakage during the communication. In the best case, 90% of the

predictions have an inaccuracy of less than 30% for a small Hello interval. Several packets can be dropped because of the failed prediction. We carried out a study to investigate the influence of nodes' relative speeds on the establishment of contacts using the extended vicinity proposed by Phe-Neau et al. The study revealed that contacts are concentrated at few hops [45, 64], as expected, and flooding procedures aiming path discovery tend to be inefficient in highly dynamic mobile wireless networks [64]. The number of works in the field of mobility pattern analysis is extensive and the majority state that assessing node mobility is fundamental to understand the network connectivity and design better routing protocols. Independent whether the extended vicinity is considered, all of the aforementioned works use node mobility information, but they leave aside the impact of the relation between the number of hops separating a pair of nodes and their relative speed.

4.2 Problem statement

Multihop communications are able to extend the coverage of a dynamic network, but path management and routing in such networks is complex due to node mobility. Hence, the investigation of mobility patterns is essential to gather information about suitable communication opportunities. This is particularly important for VANETs (Vehicular Ad Hoc Networks), because nodes in these networks usually experiment short contacts due to the intense node mobility. To maximize the exploitation of contact opportunities, we must analyze node vicinity behavior over time, which is basically defined in terms of achievable states, time spent in each state, and state transitions. Such behavior is highly influenced by nodes' relative speed, as we report in our work [64]. Hence, in this thesis, we extend this work by aggregating nodes' relative speed to the concept of node vicinity. As such, protocols could use information about relative speeds to predict contact duration and to prioritize information exchange between nodes according to their relative speeds. Note that, although absolute speeds are obtained more easily in a vehicular network, we consider that relative speeds are more suitable than absolute speeds because it determines contact duration. Relative speeds between nodes within a predetermined radio range can be estimated in real-time, as shown by Wang et al. [65].

Aiming to demonstrate the influence of nodes' relative speed on multihop contacts, we take as example a real scenario composed of buses, detailed in Chapter 3. Figure 4.1 shows the (κ, λ)-vicinity timeline, as well as the average number of hops, for all pairs of nodes in the Bus scenario. The X-axis represents the number of hops, λ, separating two nodes. Note that $\lambda = 0$ represents State ∞, in which nodes are out of contact. We consider that buses can exchange packets with each other using a wireless medium, where only nodes within 100 m radio range are able to commu-



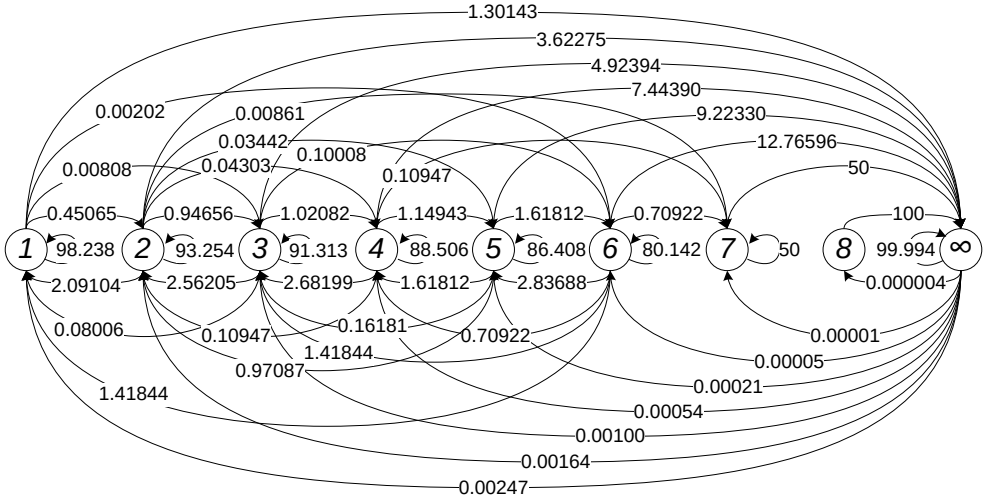
(a) $m = 1$. (b) $m = 5$.

Figure 4.1: Example of (κ, λ)-vicinity timelines, considering different values for m to show the influence of relative speeds. Nodes at lower relative speeds can reach up to 8-hop distance, whereas at higher relative speeds the maximum hop distance drops to $\lambda = 3$.

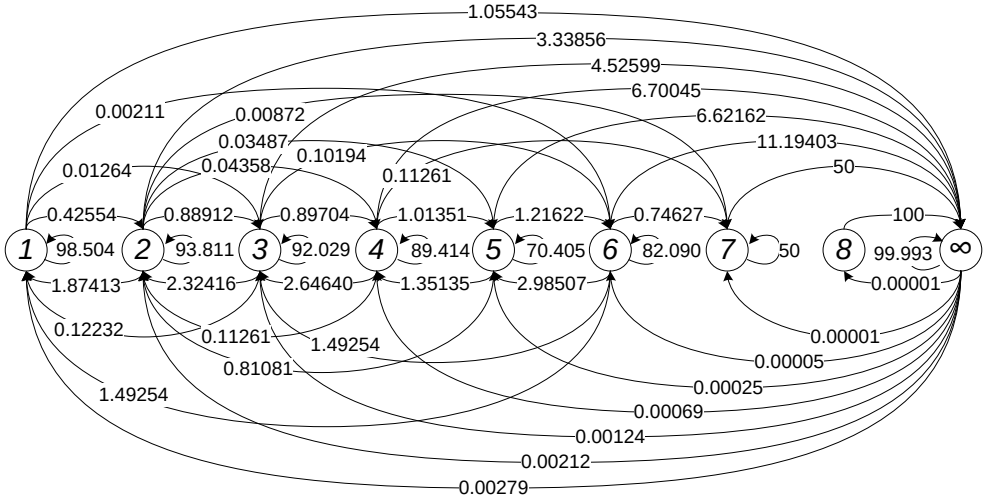
nicate without packet losses. We set $m = 1$ and $m = 5$ to demonstrate the influence of relative speeds on the vicinity timeline. If $m = 1$, $r_{i,j} \in R_0 = [0, 200$ km/h$]$, $\forall v_i, v_j$; whereas for $m = 5$, we have five intervals of 40 km/h each. Figure 4.1(a) shows that $\lambda_{max} = 8$ is the maximum number of hops, between pairs of nodes with $r_{i,j} \in [0, 200$ km/h$]$. In addition, as the average number of hops (λ_{avg}) is approximately equal to 1.44, we conclude that the number of contacts at 1-hop distance is much higher than at any other hop distance.

Figure 4.1(b) shows the results for pairs of nodes with $r_{i,j} \in [0, 40$ km/h$[$ and $r_{i,j} \in [40, 200$ km/h$]$. For the first range, we quickly note the significant similarity with Figure 4.1(a). It is clear that both the maximum ($\lambda_{max} = 8$) and average number of hops ($\lambda_{avg} = 1.45$) remain very similar. For $r_{i,j} \in [40, 200$ km/h$]$, graph bars become sparser and λ_{max} drops to 3 and $\lambda_{avg} = 1.15$. Hence, most contacts in the Bus scenario happen for relative speeds lower than 40 km/h and contacts at higher relative speeds happen mainly at small λ. As a consequence, we shall not consider long hop distances for communications if nodes are moving at higher relative speeds. All scenarios are further investigated in Chapter 5.

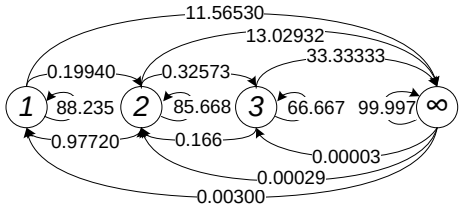
Figure 4.2 shows the state transition probabilities, p_{nm}, considering all nodes in the Bus scenario. Note that values are given in percentage (%) and each State σ represents nodes that are at λ-hop distance from another node. In this scenario, contacts are rare and tend to happen for low relative speeds and states. In Figure 4.2(a) we consider a single range of relative speeds and, consequently, all nodes are in the same set of nodes $\mathcal{V}_0$. Figures 4.2(b) and 4.2(c), in turn, are obtained when we use $m = 5$ to split the range of relative speeds, separating nodes into different subsets $\mathcal{V}$. Figure 4.2(b) represents the subset of nodes with relative speeds within $[0, 40$ km/h$[$ ($\mathcal{V}_0$), while Figure 4.2(c) shows the transition probabilities considering all nodes at relative speeds within the other 4 subsets ($[40, 200$ km/h$]$). On the one



(a) $m = 1$ and $\mathcal{V}(R_0)$.



(b) $m = 5$ and $\mathcal{V}(R_0)$.



(c) $m = 5$ and $\bigcup_{\kappa=1}^{4} \mathcal{V}(R_\kappa)$.

Figure 4.2: State transition probabilities (p_{nm}) in percentage, considering $m = 1$ and $m = 5$. The states and transition probabilities considering only the (b) subgraph where nodes are in contact at lower relative speeds is very similar to the distribution presented by the (a) complete graph. Nodes at (c) higher relative speeds are not able to reach upper states and have higher probability of transitioning to State ∞.

hand, we observe that Figures 4.2(a) and 4.2(b) are very similar, maintaining the same upper state (σ_{max}) and the great majority of state transitions. On the other hand, Figure 4.2(c) shows that the upper state for nodes within $[40, 200$ km/h] is significantly lower, confirming our observations from Figure 4.1. Moreover, most state transitions disappear, especially for $\sigma > 3$. This indicates that the highest contribution for multihop contacts in this scenario is concentrated at speeds lower than 40 km/h.

We observe in Figure 4.2 that, independently of the relative speed interval, transitions from one state to itself are most likely to happen. This probability usually decreases for upper states, indicating that longer hop contacts are more difficult to maintain for longer intervals of time. The highest probabilities are found for the transitions $\infty \rightarrow \infty$ or $8 \rightarrow \infty$, while the lowest probabilities are found for $\infty \rightarrow \sigma$, where σ is any other state. Yet, upper states have higher probability of transitioning to infinity. We also observe that any pair of nodes in State σ presents higher probability to go from its current state to State $\sigma \pm 1$ rather than to State $\sigma \pm n$, with $n > 1$. Moreover, nodes are more likely to return to State $\sigma - 1$ than to go forward to State $\sigma + 1$.

Comparing Figures 4.2(b) and 4.2(c), we further observe that the probability of going to State ∞ from any other state is much higher for higher relative speeds, reinforcing the instability of multihop contacts and, moreover, the inconvenient influence of high relative speeds on these contacts. In fact, the instability increases even for direct contacts (State 1). For instance, a pair of nodes at relative speed within $[0, 40$ km/h[has 1.055% probability of disconnecting, against a 11.565% probability if they were at higher relative speeds. These results can be easily obtained for other datasets and we generalize them for the other scenarios investigated in this work.

As expected, indeed, node mobility influences multihop communication. Therefore, we should consider it when developing protocols and applications for mobile wireless networks, especially vehicular networks. To further evaluate the impact of such speeds on multihop communications we analyze the (κ, λ)-vicinity, which includes nodes' relative speed. In the end of Chapter 5 we use the results obtained from this evaluation to analyze the performance of forwarding strategies that take into account the influence of nodes' relative speed on multihop wireless communications.

4.3 Vicinity analysis methodology

In order to analyze the (κ, λ)-vicinity, this work proposes the following procedure:

1. Mobility trace parsing;

2. Discovery of $(0, 1)$-vicinity;

3. Computation of (κ, λ)-vicinity;

4. Generation of (κ, λ)-vicinity timeline.

4.3.1 Mobility trace parsing

We first parse an input mobility trace to generate an output file as required by the next step. The input file must include, at least, the position of nodes (x, y) at each time (t) and the node identification (υ_i). The output file will contain, additionally, the absolute (s_i) and the decomposed vector (s_{ix}, s_{iy}) of speeds of each node. If s_i is not provided, we compute it, cleaning the inconsistencies, i.e., duplicates and incoherent absolute speeds. Duplicated data is simply ignored, while incoherent absolute speeds are fixed through linear interpolation and use of a threshold, based on what is intuitively expected for the scenario.

When absolute speeds are not provided, we compute them for each υ_i at each position over time, as shown in Algorithm 1. We consider that a node moves at a constant speed s_i between consecutive events, i.e., between t and $t + \Delta t$, and in the last register υ_i stops. Then, we verify for each time t at each position $x_i^{(t)}, y_i^{(t)}$, if the provided or calculated absolute speed is higher than the threshold for the scenario. Case positive, the speed is considered inconsistent, and we assume that a more consistent value must be found for the time interval $[t, t + \Delta t]$. To this end, we interpolate the current event $\left(t, x_i^{(t)}, y_i^{(t)}\right)$ with the event after the immediately consecutive one, $\left(t + 2\Delta t, x_i^{(t+2\Delta t)}, y_i^{(t+2\Delta t)}\right)$. Then, we assume that the node moves with the new calculated absolute speed from $\left(x_i^{(t)}, y_i^{(t)}\right)$ to $\left(x_i^{(t+\Delta t)}, y_i^{(t+\Delta t)}\right)$. If the speed is still above the threshold, we ignore the register. We also use the following boundary condition to compute the absolute speeds. When there is no event at $t+2\Delta t$ we cannot use the INTERPOLATE function, hence, we assign the last absolute speed (from $t - \Delta t$) to the current event (t) and consider that the node stopped at $t+\Delta t$, assigning null absolute speed to this event. As a consequence, less than 0.3% of the data is ignored due to inconsistencies for the traces we used, which does not affect our final conclusions.

Each trace samples events at a different rate. Thus, after obtaining absolute speeds, we format the results with a uniform granularity to facilitate the computation of the $(0, 1)$-vicinity. Nodes positions are updated considering that they maintain a constant speed between two adjacent points. At the end of this step, the size of the dataset increases significantly, e.g., the 10-minute subset selected from the TAPASCologne dataset increases from approximately 250 MB to 700 GB, which makes it difficult to analyze.

Algorithm 1 COMPUTATION OF ABSOLUTE SPEEDS

Input: File composed of tuples $\langle t,\ \upsilon_i,\ x,\ y,$ (optional) $s_i \rangle$
Output: File composed of tuples $\langle t,\ \upsilon_i,\ x,\ y,\ s_i,\ s_{ix},\ s_{iy} \rangle$

1: **while** υ_i **do**
2: **if** s_i does not exist in file **then**
3: $x_i^{(t)}, y_i^{(t)}, x_i^{(t+\Delta t)}, y_i^{(t+\Delta t)} \leftarrow$ GETPOSITION($t,\ t+\Delta t$)
4: $s_i^{(t)}, \leftarrow$ COMPUTESPEED($x_i^{(t)},\ y_i^{(t)},\ x_i^{(t+\Delta t)},\ y_i^{(t+\Delta t)},\ t,\ t+\Delta t$)
5: **if** $s_i^{(t)} >$ threshold **then**
6: $s_i^{(t)} \leftarrow$ INTERPOLATE(t,$t+2\Delta t$)
7: **if** valid $s_i^{(t)}$ **then**
8: vecspeed $\leftarrow$ DECOMPOSEABSSPEED($s_i^{(t)}$)
9: register $\leftarrow$ UPDATEREGISTER($s_i^{(t)}$,vecspeed)
10: WRITEOUT(register)
$\upsilon_i \leftarrow$ NEXT(υ_i)
11: **function** INTERPOLATE($t,\ t+2\Delta t$)
12: $x_i^{(t)}, y_i^{(t)}, x_i^{(t+2\Delta t)}, y_i^{(t+2\Delta t)} \leftarrow$ GETPOSITION($t,\ t+2\Delta t$)
13: $s_i^{(t)} \leftarrow$ COMPUTEABSSPEED($x_i^{(t)},\ y_i^{(t)},\ x_i^{(t+2\Delta t)},\ y_i^{(t+2\Delta t)},\ t,\ t+2\Delta t$)
14: **if** $s_i^{(t)} >$ threshold **then**
15: **return** IGNOREREGISTER( )
16: **else**
17: **return** $s_i^{(t)}$

4.3.2 Discovery of $(0,1)$-vicinity

In this step, we compute the $(0,1)$-vicinity for every pair of nodes in the network, considering they are all within $\mathcal{G}_{R_0}$. This means that nodes within a given radio range are considered to be in contact no matter the relative speed between them. Contacts in such situation are always at 1-hop distance. Hence, at a first moment, we do not take relative speeds into account and we use different fixed radio ranges to observe the effect of the medium attenuation. This step generates an output file that contains the time (t) at which the pair of nodes (υ_i, υ_j) exists in the trace simultaneously, the absolute (s_i, s_j) and the decomposed vector $(s_{ix}, s_{iy}, s_{jx}, s_{jy})$ of speeds of these nodes, the absolute relative speed between them $(r_{i,j})$, and a flag (n_{flag}) that indicates whether the nodes are 1-hop neighbors.

4.3.3 Computation of (κ,λ)-vicinity

Next, we obtain all (κ,λ)-vicinities for every node in the scenario, through the computation of the shortest paths (geodesics) between all nodes in the $(0,1)$-vicinity that exist within the same atomic interval. Each event is recorded in an output file containing the identification of the nodes (υ_i, υ_j), the number of hops (λ) between them, their relative speed $(r_{i,j})$, and the time (t) at which they are at λ-hop distance moving at $r_{i,j}$ relative speed. Note that at this point it is possible to obtain any

specific (κ, λ)-vicinity of any node v_i. To this end, it is only necessary to define the interval of relative speeds R_κ and the number of hops λ.

4.3.4 Generation of (κ, λ)-vicinity timeline

We then proceed to the generation of the (κ, λ)-vicinity timeline of a pair of nodes. We use as input the (κ, λ)-vicinity of all nodes and we filter the desired pair of nodes. The resultant (κ, λ)-vicinity provides information about the initial and final instants of time of each contact at λ-hop distance and r_{ij} relative speed, i.e., the state of the chosen pair of nodes. We store the state evolution over time to obtain the (κ, λ)-vicinity timeline of this pair of nodes.

Chapter 5

Vicinity Analysis: Results and Discussion

In this Chapter, we analyze nodes' vicinities using the method proposed in Section 4.3. We also discuss possible applications of our results and, based on our findings, we propose strategies that consider the relative speed of nodes during the forwarding process. We demonstrate through simulations that incorporating relative speed awareness into forwarding strategies does not affect the average packet delivery ratio and potentially reduce network resources consumption in multihop mobile networks.

5.1 Results

In this section, we investigate the following features:

1. Behavior of nodes' relative speed;
2. Influence of relative speeds on 1-hop contacts duration;
3. Behavior of number and duration of contacts per (κ, λ)-vicinity;
4. Average time spent by nodes in each State σ;
5. Number of useful contacts according to the (κ, λ)-vicinity.

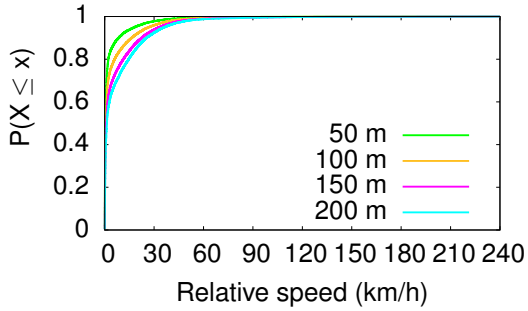
All results are obtained for different coverage ranges in $\mathcal{C} = \{50, 100, 150, 200\}$ meters.

5.1.1 Behavior of nodes' relative speed

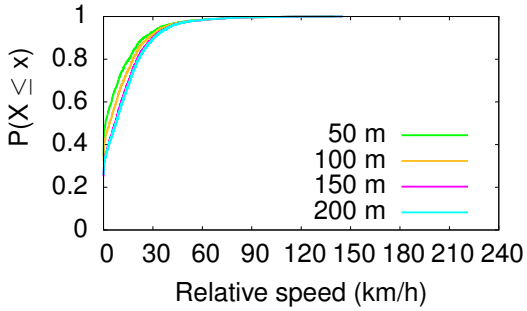
We first characterize the scenarios concerning the distribution of nodes' relative speed for all radio ranges $c \in \mathcal{C}$. The goal is to find at which relative speed contacts

are more common, considering that all nodes within c are in contact with each other at 1-hop distance. Figure 5.1 shows the resultant cumulative distribution function of relative speeds for each trace.

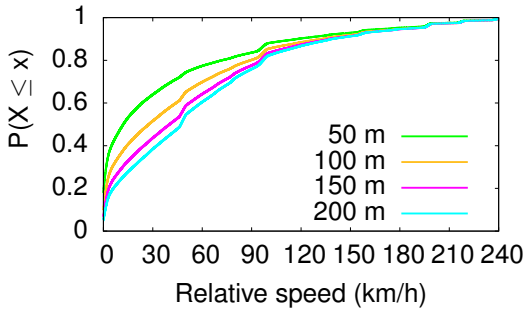
The Bus and Taxi scenarios present a similar distribution of relative speeds, as depicted in Figure 5.1. Most vehicles in contact move at low relative speeds, but this percentage decreases for longer ranges. For instance, 99% of nodes in contact move at relative speeds lower than 25 km/h for $c = 50$ m. Increasing c to 200 m, more nodes are included in the network graph and many direct contacts originated from the additional nodes happen at relative speeds higher than 25 km/h. Hence, the percentage of nodes moving at relative speeds lower than 25 km/h decreases to 90%. In the Bus scenario, we also have the majority of nodes in contact moving at speeds lower than 30 km/h for $c = 50$ m, as shown in Figure 5.1(b). For $c = 200$ m the percentage decreases from 92% to 90%. The Synthetic scenario shows a different behavior, with nodes in contact being able to move at very high relative speeds. For instance, in Figure 5.1(c), 90% of vehicles in contact move at relative speeds up to 130 km/h for $c = 200$ m. The presence of such high speeds in the Synthetic scenario can be a consequence of unlimited speed highways (autobahns) crossing the city.



(a) Taxi scenario.



(b) Bus scenario.



(c) Synthetic scenario.

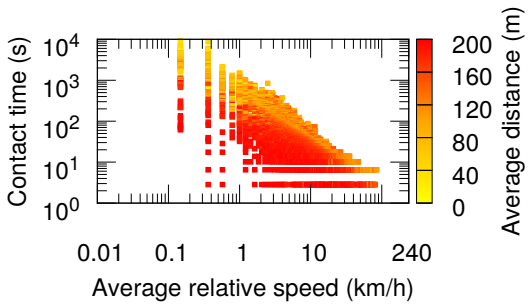
Figure 5.1: Cumulative distribution function of relative speeds for contacts at 1-hop distance. In the (a) Taxi and (b) Bus scenarios, the relative speeds of 1-hop contacts are usually low. In the (c) Synthetic scenario we observe a wider range of relative speeds.

Another remarkable characteristic is shared by all scenarios. In Figure 5.1, we observe that the distributions do not significantly change for $c = \{150, 200\}$ m, even though longer radio ranges may enclose more vehicles. Hence, we can conclude that, from a certain radio range on, the additional nodes enclosed are not able to change the distribution of relative speeds.

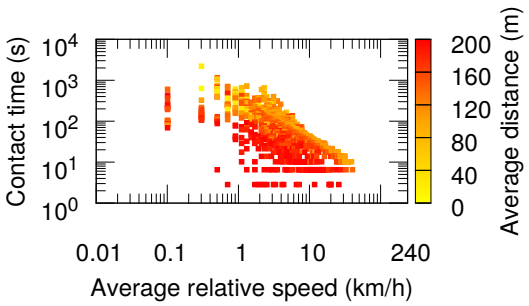
5.1.2 Influence of relative speeds on 1-hop contacts duration

We investigate the duration of contacts at 1-hop distance to evaluate if either long or short contacts are more common when nodes meet directly. We also investigate the relation with the physical distance between nodes. Note that when referring to distance in hops we always use "hop distance", whereas for physical distance we use only "distance". From our results so far, we observe that even though most vehicles in contact at 1-hop distance move at low relative speeds, we can also find some vehicles in contact at higher relative speeds. We expect that at higher relative speeds contacts at 1-hop distance are shorter, even if they happen at short physical distance. Obviously, the exact contact duration depends on the scenario. We investigate this assumption and we show the results for $c = 200$ m in Figure 5.2, for each scenario.

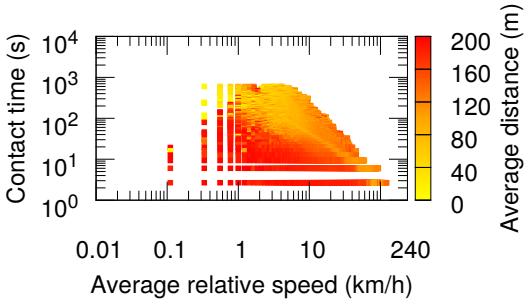
Each point in Figure 5.2 represents a contact between a pair of nodes. The du-



(a) Taxis, $c = 200$ m.



(b) Buses, $c = 200$ m.



(c) Synthetic, $c = 200$ m.

Figure 5.2: Contact duration for 1-hop contacts as a function of relative speed for each scenario.

ration of the contact is represented by the Y-axis. Both the relative speed and the distance between nodes during the contact can change, but we expect the variation to be small. Therefore, the X-axis represents the average relative speed and each point is colored according to the average distance during the contact. The distance is upper-limited by the radio range c used. The Synthetic scenario shows the greatest number of points, meaning that contacts occur more often than in the other scenarios, which can be due to the higher number of vehicles or the more intense mobility of nodes in this scenario. Note that the Y-axis in this scenario is upper-limited by the duration of the 10-minute dataset sample (600 s).

To evaluate Figure 5.2 better, we can divide it into distinct regions, according to the average relative speed, contact duration and average distance during the contact. If we choose an appropriate fixed value (threshold) for the average distance during contact, the regions are defined without much effort. Particularly, when the chosen threshold is equal to 120 m, we obtain 6 well defined regions where we can identify at which distance range more contacts happen. The boundaries of the regions are, then, defined by the thresholds for the following parameters: average relative speed (40 km/h), contact duration (20 and 200 s). We refer to distances $\geq$ 120 m as *longer* or *long range*, and $<$ 120 m as *shorter* or *short range*; average relative speeds $\geq$ 40 km/h as *high relative speeds*, and $<$ 40 km/h as *low relative speeds*; contacts duration $\geq$ 200 s as *long duration*, $<$ 20 s as *very short duration*, and $<$ 200 s but $\geq$ 20 s as *short duration*.

We notice in Figure 5.2 that a small percentage of 1-hop contacts can surpass a duration of 1,000 s. Contacts established at longer distances ($\geq$ 120 m) usually last for less than 200 s in all scenarios. Very few contacts at such distances have duration longer than 200 s. We also observe a clear correlation between the contact duration and the average distance during the contact. Most contacts at small distances ($<$ 120 m) tend to last longer than 200 s, whereas nodes farther away usually establish shorter contacts. Regarding the relation between the relative speed and the average distance during the contact, we observe that long distance contacts can happen at low or high relative speeds, as shown by the reddish points between $]0, 200]$. In addition, even though short distance contacts also happen at any relative speeds, most of them happen at lower relative speeds ($<$ 40 km/h), as shown by the yellowish points.

Focusing on the relation between the contact duration and the relative speed, it is clear that nodes are not able to establish long contacts at high relative speeds. For instance, contacts longer than 200 s happen only for relative speeds lower than 6 km/h in the Bus and Taxi scenarios, and for relative speeds lower than 11 km/h in the Synthetic scenario. Very short contacts, on the other hand, can happen at any relative speed, in all scenarios. This can be observed by the presence of several

Table 5.1: Relation between contact duration, relative speeds and physical distance between nodes. Non-existent contacts are represented as "Ø" and existing contacts as "+". We use "++" for each combination of relative speed and duration where contacts exist to indicate at which distance we observe more contacts.

		Distance	
Relative Speed	**Duration**	Short (< 120 m)	Long (≥ 120 m)
Low	Very short	Ø	+
Low	Short	++	+
Low	Long	++	+
High	Very short	+	+
High	Short	Ø	Ø
High	Long	Ø	Ø

points under 20 s within $]0, 200]$ range.

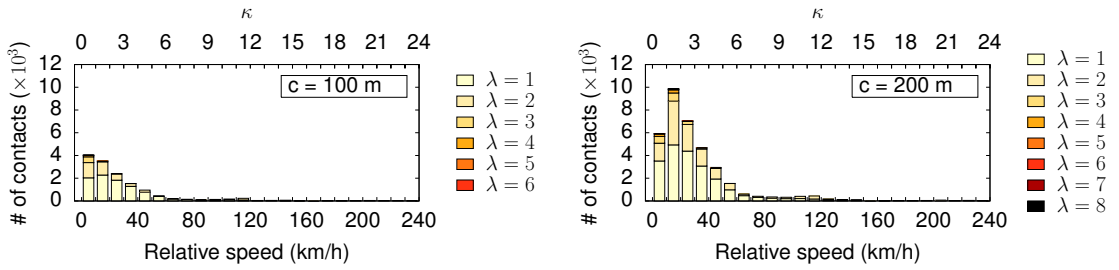
Correlating the aforementioned results to obtain the relation between contact duration, relative speed and distance between 1-hop contacts we obtain the results summarized in Table 5.1, where we consider low relative speeds as < 40 km/h, very short contacts as < 20 s and long contacts as ≥ 200 s. The results are clearer if we focus on the Synthetic scenario, because it has more data. In Table 5.1 we represent non-existent contacts as "Ø" and existing contacts as "+". For each combination of relative speed and duration where contacts exist, we increase the number of "+" to indicate at which distance we observe more contacts. The results are as follows: the majority of long contacts at low relative speeds are short distance, but we also observe very few at longer distances; short contacts at low relative speeds are usually short distance; very short contacts happen at any speed, but at low relative speeds they are always long distance, whereas at high relative speeds they can be either short or long distance; and neither short nor long contacts at high relative speeds exist.

The analyses carried out indicate that the best opportunities for exploiting the (κ, λ)-vicinity happen at low relative speeds and short distances, for which contacts last longer, as expected. Nevertheless, we can also exploit contacts at long physical distances, as long as they happen at low relative speed. At high relative speeds, contacts are very short and much likely are not useful for VANET entertainment applications, because contacts may not be long enough to successfully transfer messages. Note that considering the relative speed only is not enough to define the duration of the contact. Very short contacts can happen at low or high relative speeds, and at low relative speed they can be very short, short or long. Therefore, on the one hand, if we know that nodes are moving at high relative speeds, we can be sure that contacts are likely very short. On the other hand, if they are moving at low relative speeds, we cannot infer anything about contact duration.

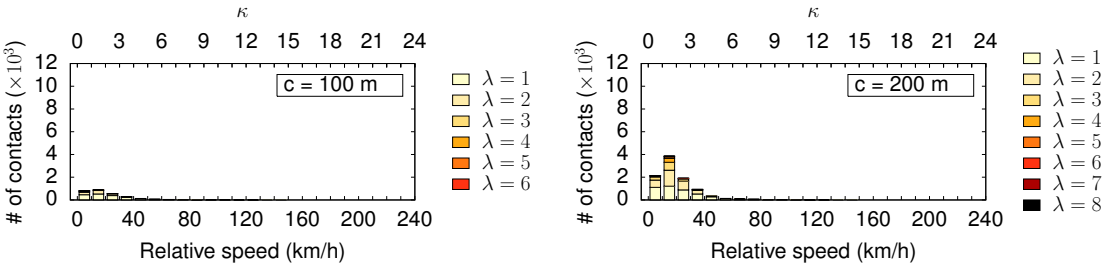
5.1.3 Behavior of number and duration of contacts per (κ, λ)-vicinity

We characterize the (κ, λ)-vicinity for each scenario regarding the total number and duration of contacts. We analyze the vicinity for $\lambda \leq 8$, because the contribution of longer hop distances to the number of contacts is insignificant compared to shorter hop distances, even using a radio range $c = 200$ m. The goal is to find out how far relative speeds influence multihop vicinity. This information can be used by routing protocols from VANETs to better adjust the maximum expected number of hops a message should be forwarded and prioritize sending messages to neighbors that most likely would provide longer contacts.

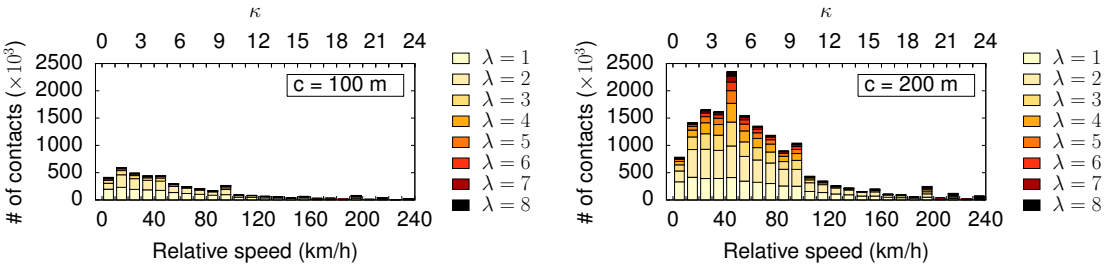
Figure 5.3 shows the results for the number of contacts, for $c = \{100, 200\}$ m. We omit the results for $c = 150$ m because they are similar to $c = 200$ m. Each bar



(a) Taxi scenario.



(b) Bus scenario.



(c) Synthetic scenario.

Figure 5.3: Total number of contacts as a function of the (κ, λ)-vicinity for each scenario, using $c = \{100, 200\}$ m. In (a) Taxi and (b) Bus scenarios there is a significant amount of contacts communicating at hop distances lower than 3. In (c) Synthetic scenario we can find several nodes communicating even at 6-hop distance.

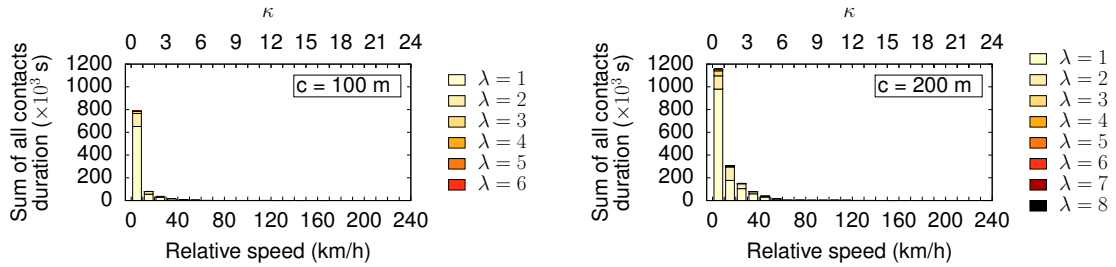
on the plot represents a different R_κ, as shown in the secondary X-axis, while the Y-axis shows the total number of contacts. Note that at each 10 km/h we have a new R_κ. In spite of the increasing number of contacts from one scenario to the other, they have some similar characteristics. For instance, we note that more contacts are established at lower relative speeds and, in all scenarios, the number of contacts increase for each R_κ when we increase the radio range. We further observe that the growth rate of the number of contacts for R_κ with $\kappa \geq 1$ tends to be greater than for $\kappa = 0$ when c increases. This is particularly true for R_1 in both Bus and Taxi scenarios. In the Synthetic scenario, the most affected relative speed range by the radio range is R_4. This is a consequence of highways in the scenario, which gives room to the higher relative speeds. In the Bus and Taxi scenarios, the contribution on the number of contacts for $\lambda \leq 3$ is more significant than for $\lambda \geq 4$, regardless of which R_κ we analyze. Concerning the Synthetic scenario, we can find significant number of contacts up to 6 hops. This occurs because the node density in this scenario is higher compared to the others.

Figure 5.4 presents the results for the sum of all contacts duration for $c = \{100, 200\}$ m. We quickly observe longer total duration in the Synthetic scenario, compared to the other scenarios. Again, Taxi and Bus scenarios present very similar behavior, with contacts lasting longer for R_0. The Synthetic scenario behaves differently, with the total duration of contacts being higher for R_4. We also observe that the most significant contribution to the total contact duration is obtained for $\lambda \leq 2$ in the Bus and Taxi scenarios, reaching up to 6 hops in the Synthetic scenario. If we compare Figures 5.3 and 5.4, we observe that although the taxis and buses do not have the majority of contacts in R_0, this is the range of relative speeds where we can find the longest total contact duration. Hence, we can infer that nodes within R_0, i.e., nodes almost stationary relatively to each other, tend to remain in contact for longer periods. Focusing on the Synthetic scenario, we observe that the highest sum of all contacts duration shifts from R_0 to R_4 when we increase the radio range from $c = 100$ m to $c = 200$ m, respectively. As the difference between the sum of all contacts duration is just slightly higher for R_4 compared to R_0, but the number of additional contacts is much higher for R_4 when we increase the radio range, we can infer that the additional contacts within R_0 tend to last longer than the new contacts within R_4. As a consequence, we can conclude that higher relative speeds contribute less to longer contacts. More than that, contacts at extremely low relative speeds ($0 \leq r_{i,j} \leq 10$ km/h) are the ones with the longest duration. Note that these nodes do not necessarily need to be moving at low absolute speeds. These results corroborate Figure 5.2.

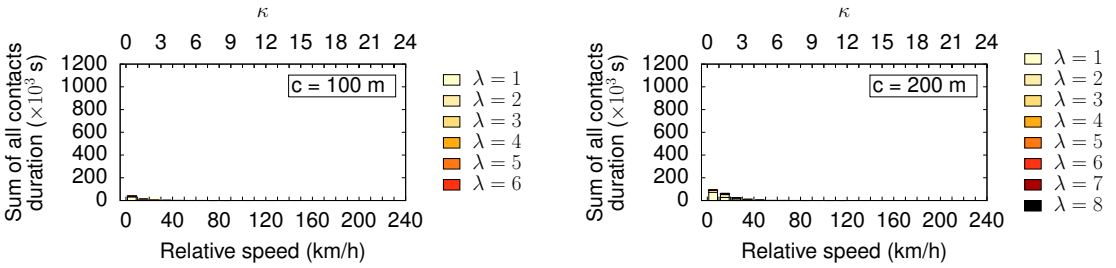
5.1.4 Average time spent by nodes in each State σ

We further investigate the average time spent by nodes in each State σ, regarding all relative speed intervals. Considering VANETs, this evaluation is important because it allows forwarding protocols to adjust parameters to exploit contact opportunities more efficiently.

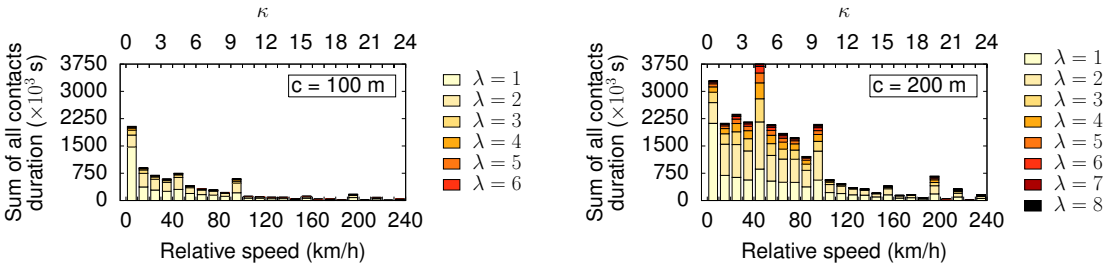
Table 5.2 shows the average time spent in each state, in seconds, for each scenario, using different radio ranges (c). We observe that the average time in contact is always longer for State 1 in all scenarios. In the Bus and Taxi scenarios, the contribution of 8-hop contacts is barely significant. In the Synthetic scenario, the average time in contact is quite similar for all states with multihop connectivity ($\lambda > 1$).



(a) Taxi scenario.



(b) Bus scenario.



(c) Synthetic scenario.

Figure 5.4: Sum of all contacts duration as a function of the (κ, λ)-vicinity for each scenario, using $c = \{100, 200\}$ m. In the (a) Taxi and (b) Bus scenarios, the most significant contribution to the total contact duration happen at very low relative speeds (< 10 km/h) and up to 2-hop distance. In the (c) Synthetic scenario the most significant contribution is obtained for nodes communicating up to 6-hop distance and at relative speeds between 40 and 50 km/h, followed by very low relative speeds (< 10 km/h).

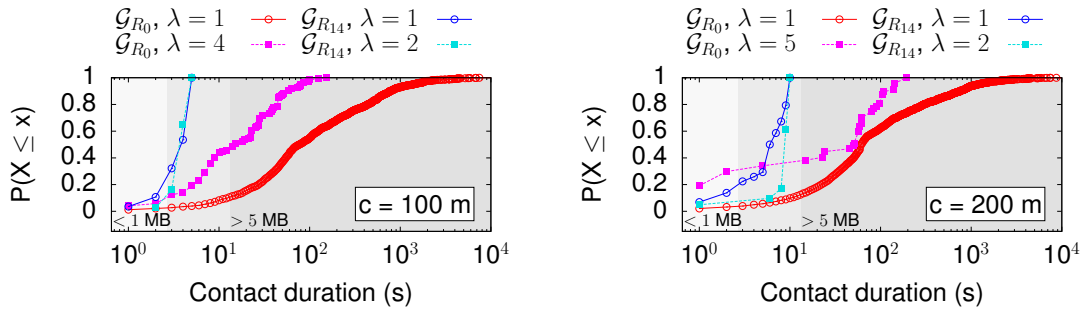
Table 5.2: Average time spent in each state, in seconds.

Traces	c **(m)**	**State** σ								
		∞	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Taxi	50	58.17	73.97	63.36	66.99	36.28	16.91	7.17	–	–
	100	58.00	79.41	38.53	33.25	24.24	27.32	36.60	–	–
	150	57.88	72.48	29.22	30.65	30.83	30.44	17.53	8.00	–
	200	57.77	65.77	27.06	31.87	27.80	24.96	11.92	6.33	5.00
Bus	50	29.62	23.31	10.56	4.60	7.00	–	–	–	–
	100	29.59	26.05	13.88	15.73	13.44	11.44	16	–	–
	150	29.56	27.80	18.68	14.28	8.20	4.55	2.36	1.00	–
	200	29.54	29.18	17.38	9.25	6.73	5.58	4.07	2.79	2.5
Synthetic	50	2.07	1.61	1.41	1.21	1.25	1.24	1.27	1.30	1.28
	100	2.08	2.17	1.73	1.39	1.33	1.32	1.34	1.36	1.36
	150	2.11	2.20	1.84	1.40	1.31	1.26	1.26	1.28	1.32
	200	2.08	2.06	1.82	1.37	1.26	1.19	1.17	1.13	1.13

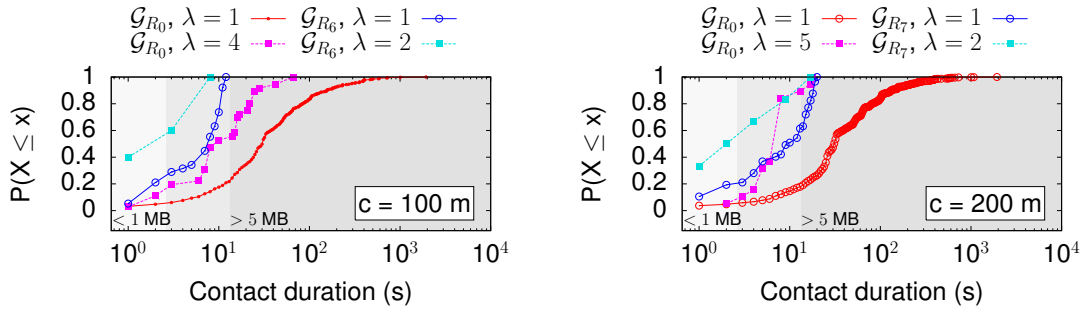
We observe in Table 5.2 that, contrarily to expected, sometimes the average time in State ∞ increases by a small quantity. For instance, in the Synthetic scenario, when we increase the radio range from 50 m to 100 m and from 100 m to 150 m, the average time out of contact increases by a small amount. Even though increasing c one would expect higher network connectivity, depending on the scenario, we can also have more transitions from vehicles that were never connected. In the Synthetic scenario, this phenomenon has as consequence a small increase on the average time spent in State ∞. This occurs because the number of vehicles transitioning, but still only connected for a short-time period, is higher than in the Taxi and Bus scenarios. The level of connectivity of these last scenarios does not change as much as the Synthetic one. Note that every time we change the radio range, a new graph appears with new transitions lasting for different amounts of time. As a consequence, non-linearities can happen on the average time spent on all states and not only on the State ∞. It is worth mentioning that no matter the radio range, the procedure proposed can be reproduced to calculate state transitions, even including the idea of different subgraphs that consider the range of relative speeds (G_{R_κ}).

5.1.5 Number of useful contacts according to the (κ, λ)-vicinity

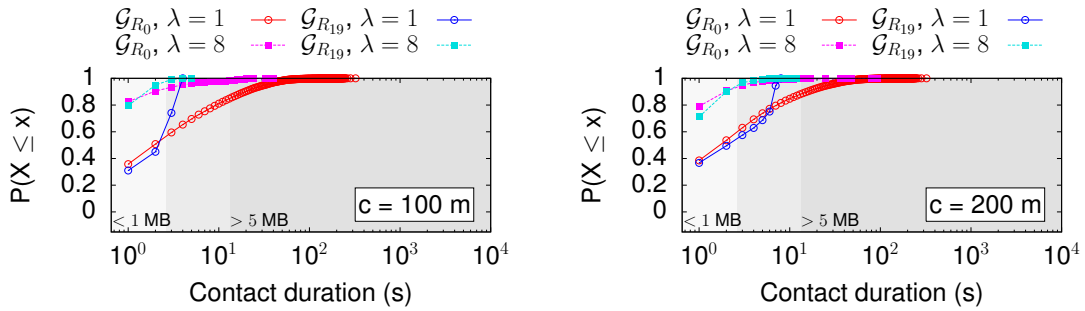
Finally, we investigate the number of useful contacts according to the (κ, λ)-vicinity, regarding all hop distances. We consider that a contact is useful if it can transfer a bundle using the lowest rate allowed by the IEEE 802.11p standard, 3 Mb/s. In this work, we use bundles of 1 MB and 5 MB, which need approximately 2.67 and 13.33 seconds, respectively, to be transmitted. We do not use single packets because the shortest contact in our datasets is equal to 1 second, which is enough to



(a) Taxi scenario.



(b) Bus scenario.



(c) Synthetic scenario.

Figure 5.5: Cumulative distribution function of the duration of contacts for the first vicinity and the last vicinity able to successfully communicate using packet bundles. In all scenarios both the hop distance and the relative speeds harshly influence the communication. In the (a) Taxi and (b) Bus scenarios there are more nodes able to fully transfer big packet bundles.

transfer up to 384 kB at the lowest rate of 3 Mb/s, 6 times more than the maximum supported IP packet size (64 kB) and at least 1966 times more than packets used in VANET safety applications [66].

We investigate all (κ, λ)-vicinities separately and we omit from our analysis all those not having enough number of contacts. We end up with a huge number of results, from which we only show the most significant: the first vicinity, represented by the subgraph $\mathcal{G}_{R_0}$, where relative speeds are within the range $[0, 10$ km/h$[$; and the vicinity with the highest possible relative speed and number of hops, where communications are still observed. Figure 5.5 plots the Cumulative Distribution

Function (CDF) of contact duration. The graphs in this figure are subdivided into 3 regions, marked by the different background colors. All contacts in the > 5 MB-region are able to transfer bundles greater than 5 MB. In the < 1 MB-region, it is possible to transfer only small bundles with less than 1 MB. Above this last region, nodes can transfer bundles greater than 1 MB. Note that, in this figure, curves with lower slopes are better. In all scenarios, we found that the greater the radio range, the more useful contacts happen for the transmission of both 1 and 5 MB bundle sizes, as expected. We clearly observe that the number of contacts that can transfer these bundles is a function of both the number of hops and the relative speed: the number of contacts that successfully transfers the bundles is more numerous for shorter hop distances at lower relative speeds and becomes less often for higher relative speeds or number of hops. An interesting finding is that relative speeds sometimes reduce more severely the number of useful contacts when compared to the hop distance. For instance, let us compare the $\mathcal{G}_{R_0}$ and $\mathcal{G}_{R_7}$ in the Bus scenario (Figure 5.5(b)), for $c = 200$ m. At 1-hop distance, approximately 97% of contacts in $\mathcal{G}_{R_0}$ are able to transfer a 1 MB bundle, against approximately 80% for contacts in $\mathcal{G}_{R_7}$. For $\lambda = 5$, we observe a reduction of 7% of contacts in $\mathcal{G}_{R_0}$ able to transfer such bundles, whereas in $\mathcal{G}_{R_7}$ the reduction is approximately 38%, when $\lambda = 2$. We further observe that the influence of relative speeds and hop distance on the number of useful contacts is more harsh for the 5 MB bundle. This behavior is repeated in all evaluated scenarios, although the specific number of useful contacts changes from one scenario to the other. Despite only confirming the notion that lower relative speeds provide better contact opportunities, the results show that increasing the relative speed is worse to wireless communications than increasing the hop distance. Besides that, we also found that many useful contacts between nodes at very high relative speeds communicating at multihop distances also exist, even though less often. For instance, 4% of the nodes can successfully communicate at 8-hop distance within $[190, 200$ km/h$[$ at 200 m radio range, in the Synthetic scenario. Considering the same radio range, we observe that, much more frequently, nodes can communicate at 2-hop distance within $[60, 70$ km/h$[$ in the Bus scenario, and within $[140, 150$ km/h$[$, in the Taxi scenario.

5.2 Discussion

The main goal of Part I is to demonstrate the impact of nodes' relative speed on multihop communications. Moreover, we aim at deriving recommendations taking into account the correlation between hop distance and relative speed so as to improve wireless communications. Although absolute speeds are more easily obtained in vehicular networks, we consider that relative speeds are more important since

they determine the contact duration. The relative speed between nodes within a predetermined radio range can be estimated even in real time, as shown by Wang et al. [65]. So far, our results showed that:

1. nodes are normally out of contact in dynamic networks, but when they are in contact they tend to remain in the same state at low relative speeds. At high relative speeds, contacts usually happen at few-hop distances;

2. direct contacts between nodes within mutual radio range last longer when their relative speed is low and their distance is less than 120 m. If the distance is greater than 120 m, most contact get shorter, independent of the relative speed. If, however, the relative speed is high, the contact is always very short;

3. contacts are more frequent at low relative speeds in sparser scenarios (Bus and Taxi), and they tend to happen at most at 3-hop distance. In high density scenarios (Synthetic), in turn, contacts at high relative speeds are found and they can happen even at longer hop distances. A significant number of such contacts, however, cannot be used to transfer data due to their short duration;

4. lastly, the number of useful contacts decreases when the relative speed or the number of hops increases.

We observed in each evaluation that the Synthetic scenario shows a different behavior compared to the other two GPS-based scenarios. Surprisingly, despite being the most dense scenario, with the highest number of contacts, even achieving longer hop distances (6 hops), we found that it has the least number of useful contacts. We attribute this to the singularities of the scenario, such as the heterogeneity of roads, including unlimited speed highways, and we do not generalize these findings to all dense scenarios. Following, we describe some possible applications of our findings and a proof of concept to highlight the importance of our observations.

5.2.1 Applications

Let us consider the purest and simplest idea of routing in challenging networks, such as VANETs. We do not have information, a priori, about contacts and there is no infrastructure to provide connectivity. Routing protocols in such networks need to use node mobility information to maximize the exploitation of contact opportunities. Shelly et al. [61], for instance, analyze the mobility history of nodes to predict the residual lifetime of a direct contact. The accuracy of their protocol could be improved, for instance, if authors included in their model the relation between the hop distance to the destination and the relative speed (Figure 5.3). As such, if the destination is too far, the current node could give up sending the message.

If we take into account some kind of infrastructure, we can use a communication model based on clusters, where a special node, the cluster head, is responsible for disseminating messages to nodes in the cluster [67]. The cluster head must be carefully chosen in order to increase message delivery while minimizing the overhead. To this end, the relative speed could be a suitable parameter to determine the nodes in the cluster, including the cluster head. For instance, the cluster should be created considering only nodes for which relative speeds are very low, so that the cluster remains unchangeable for longer periods. He et al. [68] propose a minimum delay routing algorithm considering that messages are disseminated using clusters. In their scenario, the destination point is fixed and vehicles in the same road travel with the same speed, with zero relative speed as a consequence. The cluster formation could benefit from our findings if authors considered that relative speeds can assume values different from zero. According to our results, protocols based on clusters or communities could use up to 3-hop communications at low relative speeds in sparser scenarios, i.e., a cluster could be composed of the (κ, λ)-vicinities of the cluster head, with $0 \leq \kappa \leq 4$ and $1 \leq \rho \leq 3$. The algorithm proposed by He et al. neither takes into account the relative speed nor the contact duration between clusters crossing each other paths. These quantities could be used to determine if the message can be fully forwarded to the next cluster. For instance, considering a sparse scenario and a source node that needs to send a 5 MB file to the destination, if the relative speed between the cluster and the next cluster is higher than 70 km/h, there is a high probability that the link between the two clusters will break even before the transfer completes (results from Figure 5.5(b)). One must question if nodes at high relative speeds are always considered useless contacts. We know that if two vehicles are moving at high relative speed, at least one of them must be moving at high absolute speed if they are moving in the same direction, which means that it will likely encounter a significant number of nodes. This fast node could be used as a data mule that could disseminate small messages among clusters encountered along its trip.

The outcome of this work could be used to improve prediction models, and consequently, reduce the amount of resources wasted to forward packets that are not likely to arrive at the destination. It could be useful also to improve packet delivery ratio. Bazzi et al. [60] propose two routing algorithms focusing on cellular networks offloading in the VANET context. The hop-count-based algorithm considers that if a path to the nearest Roadside Unit (RSU) does not exist, the vehicle must send the packet using the cellular network. Bazzi et al. do not focus on packet delivery ratio improvement, or on saving vehicular network resources. They could improve their protocol by considering that packets sent to a given RSU will not arrive for sure, even when a path exists, due to vehicle mobility. Considering a dense scenario,

like the Synthetic one, if the RSU is 8 hops away from the current node that carries the message and the relative speed with the next hop is higher than 10 km/h, a 1 MB message will probably not arrive at the RSU (results from Figure 5.5(c)). Therefore, the routing algorithm could determine that this packet should be sent to the cellular network, consequently increasing the packet delivery ratio (at the cost of increasing the cellular network usage). Hence, the information gathered in this work can be used by VANET routing protocols to better adjust the expected lifetime of a message in the network, or to choose the next hop, improving prediction algorithms. As a consequence, we can reduce the waste of resources due to useless packets, or even increase the packet delivery ratio if we rely on cellular networks.

We also studied the behavior of state transitions and we found singular patterns for both low and high relative speeds (Figure 4.2). The model can be used, for instance, to artificially extend a dataset to predict node mobility beyond the duration of the provided dataset. The resulting mobility pattern can be used in simulations without being limited by the dataset duration.

5.2.2 Relative-speed-aware packet forwarding

Our discussions in Subsection 5.2.1 raised the possibility of taking better decisions concerning packet forwarding, if relative speeds between nodes are considered. In this section, we evaluate this notion by showing that a local forwarding decision made by the source node can have significant impact on the network performance. More precisely, we can potentially save network resources by avoiding useless transmissions, without reducing packet delivery ratio. Using the Network Simulator 3 (NS-3), we simulate three different forwarding strategies that take into account information regarding relative speed. These three strategies are proposed herein and individually applied to the Optimized Link State Routing protocol (`OLSR`) [44]. We evaluate their efficiency, and furthermore our analysis, using packet forwarding decisions based on relative speed awareness as a use case. We chose `OLSR` because in the current state of our work, we need a routing protocol capable of providing global information about the network structure. Nevertheless, this protocol is not the most suitable for VANETs and our results confirm this.

In our simulations, the `OLSR` is used without any modification to play the role of baseline routing protocol. We then modify several classes in NS-3, linked to the operation of `OLSR`, so that we can implement some restrictions based on our results. The source files that contain the modified classes are shown in Table 5.3. We know that each `OLSR` node needs to know the global topology of the network in order to compute the shortest paths (geodesics) to other nodes. If we use the number of hops as cost metric, after computing the routing table, we know the hop distance to the

destination and the next hop towards it. We need information about node speed, but OLSR does not provide any. Thus we need to find a way to disseminate such information. To this end, we choose the Hello packets, which are sent periodically to discover neighbor nodes. We modify the header of these packets to include the vector absolute speed of the sender node, ⟨AbsSpeedX, AbsSpeedY⟩, as shown in Figure 5.6. As this information is included in a field that was previously reserved and has no use in current implementations, we do not add any overhead to the communication.

Each time a node broadcasts a Hello, the neighbor receiving the packet will have information about the decomposed vector absolute speed of the Hello sender ⟨AbsSpeedX, AbsSpeedY⟩. As such, the receiver will be able to compute the relative speed to the Hello sender. All the proposed forwarding strategies rely on this modification. The forwarding decision is made hop-by-hop, but, to make the protocol as simple as possible, and to reduce the processing overhead, the sender node is the only one that can give up forwarding a packet to the next hop because of high relative speed. More detailed discussion about the sources containing the modified classes follows.

- `olsr-header`: modified the Hello header to include nodes vector absolute speeds and methods to convert this information from and to a floating-point representation. We also included `getters` and `setters` to manipulate the additional speed field in the header.

- `olsr-repositories`: we included the vector of absolute speeds in the neighbor tuple that is used by the `olsr-routing-protocol` in the neighborhood discovery process.

- `olsr-routing-protocol`: we included the possibility of defining a relative speed threshold to be applied in the neighborhood discovery process. Forwarding decisions based on nodes' relative speed are not processed in this class. Here we construct the routing table only, based on the neighborhood limited by the chosen relative speed threshold. We added to the nodes the capability of computing the relative speed using the Hello messages. We also included the relative speed information to the routing table entries.

Table 5.3: Source files in NS-3 that implement the modified classes related to OLSR's operation.

Source file	**Comment**
`olsr-routing-protocol`	Defines the operation of OLSR
`olsr-header`	Defines the headers of OLSR packets
`olsr-repositories`	Defines the structures needed by an OLSR node
`ipv4-route`	Defines the IP (version 4) route cache entry
`ipv4-l3-routing-protocol`	Defines the operation of IP (version 4)

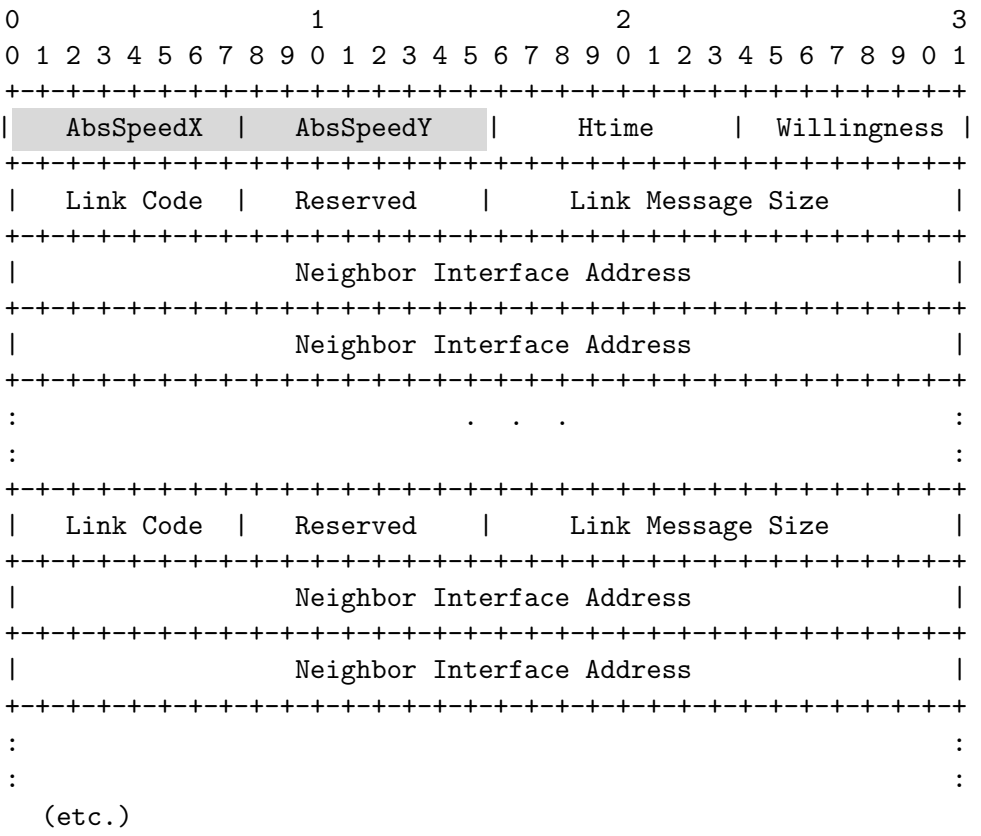


Figure 5.6: Modified format of the `OLSR` Hello header. The `AbsSpeedX` and `AbsSpeedY` (marked in gray) fields replace a previously 16-bit `Reserved` field.

- `ipv4-l3-protocol`: we included the possibility of defining relative speed thresholds to be used for forwarding decisions. Before sending the packet, a node must check the routing table and based on the relative speed it chooses whether it will send the packet to the interface. We also add a new reason for packet drops (`DROP_GIVE_UP`), which happens when the node gives up sending the packet due to an over-speeding neighbor.

- `ipv4-route`: we included `getters` and `setters` to allow the methods in `olsr-routing-protocol` to manipulate the cost metric and the relative speed between nodes.

The forwarding strategies proposed in this work are discussed hereafter. All of them use the `OLSR` as base protocol.

1. *Relative Speed Restricted forwarding* – `RelSpeedR`: after computing the routing table, the sender node, i.e., the origin of the packet, checks the hop distance, the next hop and the relative speed to the next hop, before forwarding a packet. The node will only forward the packet to the next hop if the neighbor is within a relative speed range that still allows reaching the destination at that hop distance. Hence, we impose a relative speed restriction according to

Table 5.4: Relative speed thresholds according to the hop distance to the destination.

Hop Distance	1	2	3	4	5	6	7	8
Relative Speed (km/h)	140	120	100	100	100	80	80	70

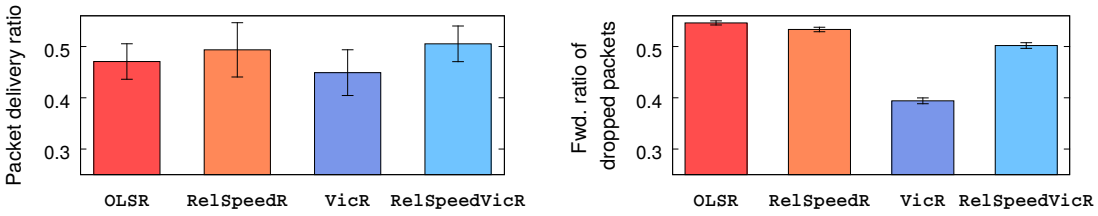
the hop distance to the destination right before the packet forwarding at the first hop.

2. *Vicinity Restricted forwarding* – `VicR`: we use the concept of (κ, λ)-vicinity to modify how the `OLSR` sees its vicinity. We consider that two nodes are neighbors only if they are within a relative speed that allows 1-hop communications. Hence, during the neighborhood discovery process, we prune some links based on the relative speed between the adjacent nodes. The forwarding happens normally, each node checks its routing table and forward the packet to the next hop based only on the cost metric. Hence, the relative speed restriction is imposed during the neighborhood discovery.

3. *Relative Speed and Vicinity Restricted forwarding* – `RelSpeedVicR`: we first prune links during the neighborhood discovery, as in `VicR`. Then, at the first hop, at the moment the packet is being forwarded, we check the relative speed to the next hop and the hop distance to the destination, as in `RelSpeedR`. Thus, the relative speed restriction is imposed in different moments.

The relative speed thresholds used in all the strategies are based on the results of Figure 5.3(c). We chose values that would reduce the number of contacts by only 20%. This quantity could be parameterized according to the application. The thresholds are summarized in Table 5.4. If nodes are moving at a relative speed higher than 140 km/h or if they are more than 8-hops away from the destination, the source node automatically discards the message.

We used a subset of the 10-minute TAPASCologne dataset as the mobility model. This subset consists of an area with approximately 4.8 km^2 containing 688 nodes. `OLSR` was used as the baseline routing protocol and we randomly installed 100 UDP client-server applications on the available nodes. A single node can host several applications simultaneously. The only restriction is that a client node cannot be its own server, and vice-versa. Each application sends packets of 1,500 bytes at a rate of 1 packet per second. The effect of the propagation medium is modeled by the combination of two propagation models. We assume that vehicles move in an urban scenario, in which losses are well modeled by the 3-Log-Distance propagation model. This model allows different attenuation factors for each distance range between the transmitter and the receiver, dividing the path loss into three separate processes. The first process is the free-space loss and the second happens due to the different



(a) Average packet delivery ratio

(b) Forwarding ratio of dropped packets

Figure 5.7: Comparison of the performance metrics using `OLSR` and each relative-speed-aware forwarding strategy based on the `OLSR`. The average packet delivery ratio for all strategies is statistically equal, but `VicR` stands out because it significantly reduces resource consumption. `RelSpeedVicR` is also able to reduce resource consumption, even though less than `VicR`.

kind of materials the signal has to travel through. The third is a random loss process that happens due to the urban environment itself, which causes multipath interferences and shadowing. In addition to this model, we use the Nakagami-m propagation model that represents the losses due to fast fading only. We consider that the signal can reach up to 230 m [69]. Even though we use this radio range, after 90 m, the fast fading is intense [70] and we model it to significant degrade the signal-to-noise-ratio. Both the 3-Log-Distance and the Nakagami-m models are implemented in NS-3.

We analyze the packet delivery ratio at the application level and the forwarding ratio of dropped packets provided by each forwarding strategy. The delivery ratio is defined as the total number of packets correctly delivered at the destination divided by the total number of packets sent, i.e., `#delivered_packets/#sent_packets`. The second metric, the forwarding ratio of dropped packets, is used to investigate the number of times that packets not delivered are forwarded before being dropped. With this metric, we aim to have an idea of the amount of resources wasted with useless transmissions. The forwarding ratio of dropped packets is then defined as the number of hops used before the packet is dropped divided by the number of necessary hops to reach the destination. Hence, the highest the forwarding ratio of dropped packets, the greater the amount of resources wasted with useless transmissions. We emphasize that we do not include delivered packets in this metric. Hence, we have that `#used_hops/#necessary_hops`≤ 1, and the unitary upper bound happens when the last hop before the destination transmits the packet but the destination does not receive it.

Figure 5.7(a) shows the packet delivery ratio for each scenario, using a 95% confidence interval. We observe that the average packet delivery ratio is statistically equal if we compare all strategies and the `OLSR`. The packet delivery ratio is low, independent of the used strategy, confirming that `OLSR` based protocols are not the

most suitable for VANETs in terms of successfully delivered packets.

Figure 5.7(b) shows the results for the forwarding ratio of dropped packets, also using a 95% confidence interval. Comparing the forwarding strategies, `VicR` wastes less network resources with useless transmissions than the others, reaching a forwarding ratio of dropped packets equal to 39.4%. The `RelSpeedVicR` is the second best forwarding strategy in terms of wasting less resources with packets. This strategy uses, in average, 50.2% of the necessary number of hops before dropping the packet, whereas `RelSpeedR` and `OLSR` achieve 53.3% and 54.6%, respectively. Therefore, we can conclude that, in the TAPASCologne scenario, submitted to the simulation parameters we described earlier, there is a clear tradeoff between delivery ratio and resource utilization. Considering the average forwarding ratio of dropped packets, the best strategy is `RelSpeedVicR`, which combines number of hops and relative speed knowledge. Although it does not improve the average delivery ratio, it does not deteriorate it. The `RelSpeedVicR` reduces the amount of useless transmissions compared with `OLSR`, but `VicR` is the most resource-efficient strategy. Depending on the application, e.g., if saving network resource is a major concern, one can decide to use `VicR` instead of `RelSpeedVicR`.

Part II

Assessment of Node Importance Using Quasi-Shortest Paths

Chapter 6

The ρ-Geodesic Betweenness Centrality

In this chapter, we focus on the importance of nodes in general multihop networks, considering their topological position and the multiple paths between them. To this end, we first present the existing centrality metrics, focusing on the betweenness centrality. Then, we propose a new centrality metric, the ρ-Geodesic Betweenness Centrality, which uses other paths, besides the shortest ones, to assess node importance.

6.1 Background

Each node in a system plays a different role on the transfer of flows across the network. The quantification of the importance of each role is not trivial. Knowledge about node importance is essential to better distribute functionalities across the network, or to better understand the relations between nodes. Frequently, node importance is assessed using centrality metrics, which are based on nodes topological positions. Examples of centrality metrics are degree, closeness, and betweenness [15–17, 42]. The degree relates to the popularity of a node, the closeness to how quickly it can access or spread resources (e.g., information), and the betweenness relates to the control that a node can exert over the flow between other nodes [15]. This work focuses on the betweenness centrality, which was formally introduced by Freeman in 1977 [43], based on the intuitions revealed in several previous works, using paths to determine the importance of a node. Many types of network can benefit from this information. For instance, in social networks, nodes with high betweenness are known as gatekeepers or brokers: those that control what information enters or leaves a group. In computer networks, we may find appealing to place data flow analyzers on nodes with high betweenness, because they have higher probability

of lying between other nodes. In the remaining of this section we first review the traditional definition of betweenness centrality and then we present its main variants used in the literature.

6.1.1 Betweenness centrality

The traditional idea behind the betweenness centrality is that nodes that fall on many shortest paths (geodesics) between other nodes are more central, because they can potentially control the flow of information in the network [71–74]. Note that this notion assumes that the flows always follow shortest paths. Hence, considering a pair of nodes v_i, v_j, the number of shortest paths between them ($n^*_{i,j}$) passing through v_k ($n^*_{i,j}(v_k)$) increases the control exerted by v_k over the flows between v_i, v_j. Freeman [43] mathematically formalizes this idea, assuming that a message passes through one of the existing shortest paths between v_i, v_j with probability equal to $1/n^*_{i,j}$. The chance that one of such paths passing through v_k is randomly picked is given by [43]:

$$b_{i,j}(v_k) = \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}, \tag{6.1}$$

which can be computed considering all pairs in the network, defining the overall *betweenness centrality* of v_k as:

$$B_{trad}(v_k) = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}, \tag{6.2}$$

where $i \neq k \neq j$ and $j \neq i$[1]. The betweenness can be further normalized by the maximum value that could be assigned to a node in a network; this is obtained for the central node in a star graph with the same number of nodes as the network in analysis. The betweenness for this central node is equal to the number of paths it falls on: $0.5 \cdot \big(|\mathcal{V}| - 1\big) \cdot \big(|\mathcal{V}| - 2\big)$ [43]. Note that Freeman initially considered only undirected graphs, but it is possible to derive, analogously, the normalization factor for directed graphs, which is equal to $\big(|\mathcal{V}| - 1\big) \cdot \big(|\mathcal{V}| - 2\big)$.

Freeman suggests that his metric is suitable for networks where knowledge about node betweenness can affect the network operation somehow. For instance, in communication networks, if we compute his metric we can find nodes that have the potential to control the communication between other nodes [43]. As such, we could instruct these nodes to act as flow selectors, allowing specific flows and blocking others. Although very important, Freeman's betweenness is limited to simple graphs, leaving aside the strength or cost of the relationship between adjacent nodes, i.e.,

[1] We always consider these constraints. Therefore, they will be omitted in the remainder of the work.

the weight of the link between a pair of nodes. In the remainder of this work, we refer to Freeman's betweenness as the *traditional betweenness*.

Freeman also assumes that the flow of information is always governed by the shortest-path rule, which may not be true in some cases. For instance, rumors and diseases are good examples of information that spread randomly. Rumors can be, in addition, intentionally channeled through specific intermediaries [75]. Policies [76] and the placement of virtual machines, on the other hand, are neither necessarily ruled by randomness nor shortest paths. Instead, they usually follow previously defined requirements, e.g., to meet energy constraints or performance goals [77].

Many works already questioned the shortest-path rule, proposing new metrics to quantify the importance of a node [38–42, 78]. Some of them also tried to tackle this issue in weighted networks [38, 42]. The simplest proposals were made by Borgatti and Everett [41] and Geisberger et al. [78]. All of them still focus on the use of shortest paths. In the following sections we discuss the main variants of the traditional betweenness.

6.1.2 Accounting path length: bounded-distance, distance-scaled and linearly-scaled betweenness

Borgatti and Everett [41] argue that the length of the path should influence the betweenness because longer paths are less valuable to be controlled or may not be realistic for some networks, such as friendship. Hence, they propose the *bounded-distance* and the *distance-scaled betweenness centralities*, which reduce the importance of longer paths when computing node betweenness. These metrics are formalized in Equations 6.3 and 6.4, respectively.

$$B_{bounded}(v_k) = \sum_{\substack{i \in \mathcal{V} \\ \lambda^*_{i,j} \leq \kappa}} \sum_{j \in \mathcal{V}} \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}. \tag{6.3}$$

The parameter κ in Equation 6.3 is used as a threshold, so that paths longer than κ hops are discarded.

$$B_{dist}(v_k) = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} \frac{1}{\lambda^*_{i,j}} \cdot \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}. \tag{6.4}$$

In the distance-scaled betweenness, instead of discarding longer paths, the authors use all shortest paths and weight their contribution to the importance of a node according to their length, $\lambda^*_{i,j}$, as shown in Equation 6.4.

Geisberger et al. [78] propose a complementary variation of the distance-scaled betweenness, the *linearly-scaled betweenness*, in which they also account for the dis-

tance between the source and the intermediary node. They argue that intermediary nodes closer to the destination should have more control over the communication. Their metric is formalized in Equation 6.5:

$$B_{lin}(v_k) = \sum_{i \in |\mathcal{V}|} \sum_{j \in |\mathcal{V}|} \frac{\lambda^*_{i,k}}{\lambda^*_{i,j}} \times \frac{n^*_{i,j}(v_k)}{n^*_{i,j}}. \tag{6.5}$$

6.1.3 Analyzing weighted networks: flow and Opsahl's betweenness

None of the aforementioned works is able to quantify the importance of a node in weighted networks. Such networks are essential to represent the strength or the cost of a relationship between two nodes. To assess node importance in these networks, Freeman et al. [38] and Opsahl et al. [42] propose the *flow betweenness* and a weighted variant of the traditional betweenness, to which we refer as *Opsahl's betweenness*. Freeman et al. interpret the weight of each edge in the graph as the capacity of a channel. This channel can be modeled as an elastic pipe connecting two nodes and it becomes narrower as the distance between the nodes increases. As a consequence, the more distant the nodes, the smaller the channel capacity. At some point, the pipe can no longer be stretched and it breaks, interrupting the flow between the nodes [38]. The authors determine the maximum flow, $m_{i,j}$, between a pair of nodes and the maximum flow between these nodes that passes through v_k, $m_{i,j}(v_k)$. To this end, they use the concept of i-j cut sets, which are subsets of edges that disconnect nodes v_i, v_j when removed. The overall dependency on v_k to the maximum flow defines the *flow betweenness*, which can be understood as a measure of the flow amount supported by a node when the maximum flow is pumped in the network. The metric is formalized as in Equation 6.6.

$$B_{flow}(v_k) = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} m_{i,j}(v_k). \tag{6.6}$$

This value can be normalized by the total flow between all pairs of nodes, given by $\sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} m_{i,j}$.

The major drawback of the flow betweenness centrality is the need to know all the independent sets between each pair of nodes in the network, which increases time complexity. Opsahl et al. proposed a simpler workaround to handle weighted networks [42]. They extend the traditional concept [43] to take into account both the number of nodes in-between other nodes, known as intermediary nodes, and the strength of the relationship between the nodes. The authors use a slightly modified implementation of Dijkstra's algorithm to find the shortest path using inverted weights $\omega_{i,j}$ tuned by a parameter $\alpha \in \mathbb{R}_+$, which determines the relative

importance of the number of links compared to the link weights. Depending on the value of α, the metric accounts for: only the number of shortest paths ($\alpha = 0$), only the inverse of the weights ($\alpha = 1$), favors the length of the path over the cost ($0 < \alpha < 1$), or favors the cost of the path over the length ($\alpha > 1$). The weighted adjacency matrix ω and the tuning parameter α are used during the computation of shortest paths. The metric is formalized as in Equation 6.7.

$$B^{(\omega\alpha)}(v_k) = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} \frac{n_{i,j}^{(\omega\alpha)}(v_k)}{n_{i,j}^{(\omega\alpha)}}, \tag{6.7}$$

where $n_{i,j}^{(\omega\alpha)}$ and $n_{i,j}^{(\omega\alpha)}(v_k)$ represent, respectively, the number of shortest paths and how many of them pass through v_k. Note that ω and α are used in Equation 6.7 to indicate that the network is weighted and a tuning parameter can be used.

6.1.4 Including longer paths: current flow and random walk betweenness

The aforementioned works consider that the information in the network will always follow some kind of optimal path. Newman claims that a realistic betweenness measure should include paths that are not necessarily the shortest [40]. Thus, he tackles the problem by completely relaxing the idea of following optimal paths. Newman suggests that the information in the network can wander around essentially at random until it finds its destination and, thus, we should include contributions from many paths that are not optimal in any sense [40]. Hence, he proposes the *random walk betweenness*, which measures the expected number of times a random walk starting at a source and ending at a destination passes through a node v_k along the way, computed considering all pairs of nodes. Such a metric is computed using matrix methods, and it is proven [40] to be equivalent to the *current flow betweenness* [39, 40]. Mathematically, the random walk betweenness is formalized as in Equation 6.8.

$$B_{rnd}(v_k) = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} I_k^{(ij)} \tag{6.8}$$

where $I_k^{(ij)}$ is computed by Equation 6.9:

$$I_k^{(ij)} = \frac{1}{2} \sum_{l \in \mathcal{V}} A_{k,l} |T_{k,j} - T_{k,i} - T_{l,i} + T_{l,j}| \tag{6.9}$$

and $I_i^{(ij)}$ and $I_j^{(ij)}$ can be 0, when edge nodes v_i, v_j are not considered, or 1, otherwise. These equations can be computed using the following steps [40]:

1. Construct the diagonal matrix of vertex degrees, D;
2. Construct the adjacency matrix, A;
3. Compute the matrix $D - A$;
4. Remove the x^{th} row and the corresponding x^{th} column;
5. Invert the resulting matrix;
6. Add back the x^{th} row and x^{th} column, but now filled with 0's. The resulting matrix is named T;
7. Calculate the betweenness as in Equation 6.8.

Note that, in the traditional betweenness, the flow knows exactly where it is going to and which path is the best to arrive there; whereas in the random walk betweenness, it has no prior idea of where the destination is, wandering around at random until the destination is found. Hence, we can consider these metrics as the two extremes of a betweenness centrality spectrum with the other metrics based on shortest paths lying between them.

Summarizing, the aforementioned works use the frequency of participation on paths to determine the importance of a node. Each one defines the type of path they use: all of them or only the shortest ones, and specifies whether the length or the cost of the path influences on the contribution to the importance of the node. This is why there are several betweenness metrics. We show the possible combinations between type and length of paths in Table 6.1.

In this work, we argue that shortest paths alone are not enough to determine the betweenness centrality of a node, because flows can also follow non-ideal paths, depending on the network we analyze and the application that we need. Hence, our idea is similar to the flow and random walk betweenness, according to this aspect, but it differs from the traditional, bounded-distance, distance-scaled, linearly-scaled and Opsahl's betweenness. In addition, we agree that the contribution of the paths

Table 6.1: Summary of betweenness centrality metrics discussed in this work.

Metric Name	**Symbol**	**Influence of Path Length**	**Type of Path**	**Weighted Networks**
Traditional Betweenness [43]	B_{trad}	No	Shortest	No
Bounded-Distance Betweenness [41]	$B_{bounded}$	Yes	Shortest	No
Distance-Scaled Betweenness [41]	B_{dist}	Yes	Shortest	No
Linearly-Scaled Betweenness [78]	B_{lin}	Yes	Shortest	No
Flow Betweenness [38]	B_{flow}	Yes	All	Yes
Opsahl's Betweenness [42]	$B^{(\omega\alpha)}$	Yes	Shortest	Yes
Random Walk Betweenness [40]	B_{rnd}	Yes	All	Yes

should be weighted by their length (or cost), because flows tend to concentrate on the shortest paths. Therefore, considering this claim, our proposal is similar to all metrics discussed in this section, except the traditional betweenness, when length and cost are interchangeable. Nevertheless, even though we consider the influence of path lengths, we weight the contribution of each path relatively to the cost of the shortest path. Remember that, in this work, the *length* and the *cost* of a path are used interchangeably. If we account both the influence of path length (cost) and the type of path considered by each metric, our proposal is different from all metrics, except the flow and random walk betweenness. In fact, conceptually, our idea is similar to the one used by these two metrics, but we claim that it is not necessary to use every path. Instead, it is sufficient to consider only the shortest paths and *slightly* longer ones. This is because paths much longer than the shortest one do not concentrate significant amount of flows and, thus, they cannot have much influence on the importance of a node [40]. Using such paths on the computation of centrality metrics unnecessarily increases the complexity of the metric. We argue, then, that paths longer than a certain threshold could be discarded without significantly affecting the node importance. Hence, our proposal uses shortest and slightly longer paths, and consider the cost of each path. Thus it can be applied to both unweighted and weighted networks This is better discussed in the next section.

6.2 The ρ-geodesic betweenness

In this section we present the weighted betweenness centrality metric proposed in this work, discussing its conceptual idea, formalization, properties, and implementation.

6.2.1 Metric overview

We saw in Chapter 1 and in Section 6.1 that the traditional betweenness is frequently used to estimate node importance. This metric, however, only accounts for

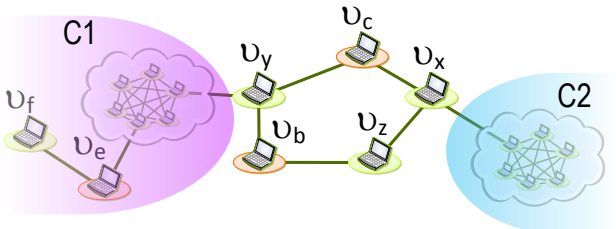


Figure 6.1: Example of network topology where betweenness centrality metrics can fail to capture the importance of a node on a quasi-shortest path. The clouds represent any type of network topology, as long as it is connected.

shortest paths, ignoring nodes that participate on slightly longer paths. In multihop networks, we may have multiple paths between pairs of nodes. In this case, we argue that shortest paths only are not enough to determine node importance and we must also consider other paths. The use of multiple paths, even if they are not as short as the shortest one, can provide network reliability and increase the total throughput.

In this thesis, we aim to capture the potential of intermediary nodes neglected by shortest-path-based betweenness centralities in multipath networks. To this end, we propose the *ρ-Geodesic Betweenness Centrality*. In a nutshell, the ρ-geodesic betweenness of a node v_k is computed using the proportion of shortest and *quasi*-shortest paths that v_k falls on between all possible pairs of nodes in the network. This proportion is weighted by the ratio between the cost of the shortest path connecting a pair of nodes and the cost of the *quasi*-shortest path between the same pair of nodes passing through v_k. The search for *quasi*-shortest paths is limited by the spreadness factor ρ, which defines the maximum extra cost that the proposed ρ-geodesic betweenness can take into account (Chapter 2). We will see that a small ρ is enough to capture well the idea of *quasi*-shortest paths while keeping the computational load low.

Some nodes neglected by the shortest-path-based betweenness centralities can be crucial to the network operation, but they are not accounted by typical betweenness metrics just because they do not fall on a sufficiently large number of shortest paths. This problem is illustrated in Figure 6.1, in which both v_c and v_b are important to maintain the network components `C1` and `C2` connected, and v_e connects an edge node to the rest of the network. Let us consider only the paths between `C1` and `C2`, and their contribution to the importance of nodes v_b, v_c and v_e. The traditional betweenness of v_c will be the highest among these three nodes followed by v_e. The node v_b would receive null importance. This happens because v_c and v_e participate on many shortest paths, while v_b does not fall on any. Nevertheless, v_b must be given some importance, because it is part of a backup path that can maintain both network components connected at a small additional cost if v_c fails. Even if one of the paths is slightly longer, it could also be used to distribute the load between both network components in two paths, to avoid congestion, or as part of a congestion recovery process. If we consider other paths besides the shortest one to compute the centrality of nodes, v_b would be assigned some importance.

Table 6.2 compares the betweenness computed for v_e, v_c, and v_b using the traditional (B_{trad}), distance-scaled (B_{dist}), and random walk (B_{rnd}) betweenness, considering two mesh networks composed of six nodes, one connected between v_e and v_y, and the other connected to v_x, as shown in the clouds of each component in Figure 6.1. We observe in Table 6.2 that the random walk betweenness is the only metric able to assign a significant importance to v_b. Hence, we say that this metric

Table 6.2: Comparison of traditional (B_{trad}), distance-scaled (B_{dist}) and random walk (B_{rnd}) betweenness metrics for nodes in Figure 6.1.

Node	B_{trad}	B_{dist}	B_{rnd}
v_c	63.0	12.1	47.6
v_b	9.0	2.5	35.6
v_e	17.0	3.9	17.0
v_x	73.0	16.1	80.4
v_y	81.0	17.7	86.8
v_z	7.0	2.1	34.8
v_f	0.0	0.0	0.0

proportionally captures the importance of v_b to the network. This happens because the random walk betweenness also accounts for paths longer than the shortest one.

6.2.2 Metric formalization

Our approach differs from the works discussed in Section 6.1 by considering in a single metric the number of shortest and *quasi*-shortest paths between all pairs of nodes, as well as the cost of each path. These costs are introduced as a ratio between the cost of the shortest path, $\delta^*_{i,j}$, and the cost of the *quasi*-shortest path through v_k, given by $\delta_{i,j} = \delta_{i,k} + \delta_{k,j}$. Hence, the ρ-geodesic betweenness weights the paths proportionally to their costs, assigning higher importance to nodes on shorter paths. The *maximum* cost of the *quasi*-shortest path depends on the spreadness factor ρ. For instance, if $\rho = C$, the maximum cost for the *quasi*-shortest path is $\texttt{max}(\delta_{i,j}) = \delta^*_{i,j} + C$, thus, paths that cost $\delta^*_{i,j} + (C + \varphi)$, where $\varphi \in \mathbb{R}^*_+$, are ignored in the computation. Note that we account for all the paths with cost $\delta_{i,j} \leq \delta^*_{i,j} + C$, including the shortest one ($C = 0$).

The concept of the ρ-geodesic betweenness (B_ρ) is quite similar to the one of the traditional betweenness, which can be understood as the frequency with which v_k falls on shortest paths between all pairs of nodes in the network. Analogously, the proposed metric measures the frequency with which v_k falls on paths that cost less than or equal to $\delta^*_{i,j} + \rho$. The idea behind the limitation imposed by ρ is based on the fact that the throughput of information traveling through paths for which $\delta_{i,j} \gg \delta^*_{i,j}$ is expected to be low. The proposed metric is formalized in Equation 6.10.

$$B_\rho(v_k) = \sum_{\substack{i \in \mathcal{V} \\ \delta_{i,k}+\delta_{k,j}-\delta^*_{i,j} \leq \rho}} \sum_{j \in \mathcal{V}} \frac{n^*_{i,j}(v_k) + n_{i,j}(v_k)}{n^*_{i,j} + n_{i,j}} \times \frac{\delta^*_{i,j}}{\delta_{i,k} + \delta_{k,j}}. \tag{6.10}$$

If $\rho = 0$, $\delta_{i,j} = \delta^*_{i,j}$, and only the shortest paths are considered. In addition, the metric is computed for source-destination nodes that lie in the same component, such that each partial value is equal to zero if these nodes are in different components.

6.2.3 Properties

The ρ-geodesic betweenness centrality has the following properties:

- It considers the number of multiple paths between nodes, both shortest and *quasi*-shortest paths;
- It increases with the participation of v_k in both shortest and *quasi*-shortest paths;
- It prioritizes low cost paths by decreasing the contribution of costly paths through a cost ratio;
- It grows with the centrality of the node.

Note that, in this work, the node is considered more central if it participates on multiple paths, either shortest or *quasi*-shortest. The reason behind this consideration is that nodes that participate in several *quasi*-shortest paths should not be discarded just because they are not on the shortest path, as they could be important in many situations. For instance, such nodes that are so close to the shortest path could serve as backup nodes during a network failure. In Figure 6.1, for example, v_b is part of a possible backup path between both sides of the network. The ρ-geodesic betweenness of nodes v_c, v_b, v_e, v_x, v_y, v_z, v_f for $\rho = 3$ are equal to 117.1, 39.6, 35.6, 156.9, 173.1, 36.2 and 0.0 respectively, considering unitary cost for each link.

The upper and lower limits of each partial term of the metric depend on the proportion of shortest and *quasi*-shortest paths that v_k participates. In addition, these limits depend on the ratio between the costs of such paths. The value of ρ can modify the proportion of node participation on shortest and *quasi*-shortest paths. As a consequence, it can influence the limits of each partial term, being able to decrease the lower limit down to 0 if $\delta_{i,k} + \delta_{k,j} - \delta^*_{i,j} > \rho$.

Using larger values for ρ, we can find more *quasi*-shortest paths, and if ρ is sufficiently large to account for at least one of these paths, the lower limit will tend to 0 if the cost of the *quasi*-shortest paths is much greater than the cost of the shortest path between the same pair of nodes, i.e., $\delta_{i,k} + \delta_{k,j} \gg \delta^*_{i,j}$. Note that if the cost is ∞, the nodes are considered as unreachable, meaning that the contribution to the ρ-geodesic betweenness is null. The lower limit will also tend to 0 if the value of ρ provides too many *quasi*-shortest paths, such that the number of existing paths between v_i, v_j is much greater than the number of such paths that v_k falls on, i.e., $n^*_{i,j} + n_{i,j} \gg n^*_{i,j}(v_k) + n_{i,j}(v_k)$. In the best case scenario, the upper limit of each term is equal to 1, when v_k only falls on shortest paths and participates in all shortest paths connecting v_i, v_j, meaning that $n^*_{i,j}(v_k) + n_{i,j}(v_k) = n^*_{i,j} + n_{i,j}$ and $\delta^*_{i,j} = \delta_{i,k} + \delta_{k,j}$.

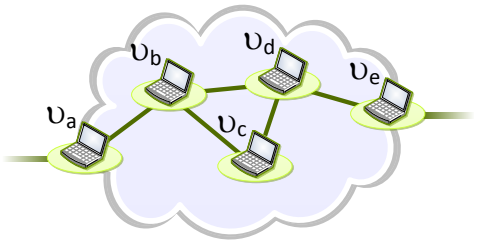


Figure 6.2: The nodes in the example can be divided in 3 sets according to their ability to intermediate flows: $\{v_a, v_e\}$ will never intermediate them, $\{v_b, v_d\}$ intermediate the great majority, and $\{v_c\}$ intermediates if necessary. Nevertheless, shortest-path-based centralities, such as the traditional and the distance-scaled betweenness, classify v_c in the same set of v_a and v_e.

The ρ-geodesic betweenness presents an intrinsic higher variance, compared to other shortest-path-based centrality metrics, such as the traditional and distance-scaled betweenness. This happens due to the inclusion of the *quasi*-shortest paths on the computation of node importance. Therefore, we can have a broader spectrum to classify nodes according to their importance and, thus, we achieve a more fine-grained node ranking. This is especially true for larger ρ. Furthermore, the ρ-geodesic betweenness is able to assign importance to nodes even if their ego network density is unitary, whereas the aforementioned metrics cannot, as we observe in Figure 6.2. If we consider only the set of nodes $\{v_a, v_b, v_c, v_d, v_e\}$, it is clear that using only shortest paths will lead us to a condensed ranking, with only two positions, occupied by two groups of nodes: $\{v_b, v_d\}$ and $\{v_a, v_c, v_e\}$. Nevertheless, v_c is clearly different from v_a and v_e in the sense that it can obviously intermediate communications, if necessary, while the former cannot because they are endpoints. Therefore, we argue that the second group should not be composed by $\{v_a, v_c, v_e\}$. Instead, v_c should be reranked as more important than the other two nodes in this group, broadening the ranking. This reranking is achieved by both the random walk and the ρ-geodesic betweenness, as we observe in the results of Table 6.3. Note, however, that the ρ-geodesic betweenness metric ($B_{\rho=1}$) with $\rho = 1$ assigns higher importance

Table 6.3: Comparison of the betweenness of nodes in Figure 6.2. Both the random walk (B_{rnd}) and ρ-geodesic betweenness (B_ρ) are able to broaden the node ranking. The distance-scaled (B_{dist}) and traditional (B_{trad}) betweenness cannot capture the importance of node v_c.

Node	B_{trad}	B_{dist}	B_{rnd}	$B_{\rho=1}$
v_a	0.0	0.0	0.0	0.0
v_b	3.0	1.3	3.7	2.7
v_c	0.0	0.0	1.3	1.3
v_d	3.0	1.3	3.7	2.7
v_e	0.0	0.0	0.0	0.0

to v_c proportionally to v_b and v_d (1.3 *vs.* 2.7) than the random walk betweenness (1.3 *vs.* 3.7).

6.2.4 Algorithm for computing the ρ-geodesic centrality

The steps used to compute our metric for one node v_k are as follows.

1. Look for the shortest paths from every v_i to every other node v_j in the network;
2. For each pair v_i, v_j, store the cost of the shortest path, the cost threshold and the number of shortest paths between these nodes;
3. Count and store the number of shortest paths between each pair v_i, v_j that passes through v_k;
4. Look for the *quasi*-shortest paths from every v_i to every other node v_j in the network, using the stored cost threshold to limit the search;
5. For each pair v_i, v_j, store the cost of the *quasi*-shortest paths and the number of such paths between these nodes;
6. Count and store the number of *quasi*-shortest paths between each pair v_i, v_j that passes through v_k;
7. Calculate the betweenness of v_k as in Equation 6.10.

At the end of these steps we have the ρ-geodesic betweenness of v_k. This can be easily implemented if we want to compute the betweenness of only one node. Without any modification, however, if we followed these steps to compute the ρ-geodesic betweenness of all nodes in the network, the algorithm would not be efficient because steps 1 and 4 are computationally intensive and would need to be repeated for each pair v_i, v_j. Hence, to compute our metric for all nodes in the network, we modify Brandes' algorithm for the traditional betweenness [79] to also look for *quasi*-shortest paths and account their contribution to the importance of nodes. The algorithm used to compute the ρ-geodesic betweenness for all nodes in the network is described in Algorithm 2.

In this work, we use the number of hops as cost metric and, thus, $\rho \in \mathbb{N}$. We use $\Delta_{max} = \rho + 1$ vectors $D^{\Delta}_{\texttt{src}}$ and $N^{\Delta}_{\texttt{src}}$ for each `src` to account for the paths that cost from 0- to ρ-hops more than the shortest one. Vectors $D^{\Delta}_{\texttt{src}}$ and $N^{\Delta}_{\texttt{src}}$ are composed of `numNodes` elements each and D^{Δ}_{src} represents the path cost between `src` and all other nodes, while N^{Δ}_{src} has the number of paths between these nodes. Hence, for a given ρ, we have $D^{\Delta}_{\texttt{src}} = [\delta_{\texttt{src},1}, \ldots, \delta_{\texttt{src},\texttt{numNodes}}]$ and $N^{\Delta}_{\texttt{src}} = [n_{\texttt{src},1}, \ldots, n_{\texttt{src},\texttt{numNodes}}]$. Note that $0 \leq \Delta \leq \rho$, where $\Delta = 0$ refers to arrays concerning the shortest paths, and

Algorithm 2 BASIC ρ-GEODESIC BETWEENNESS

Input: ρ, G
Output: ρ-GB

1: **for** $\texttt{src} \leftarrow 1, \texttt{numNodes}$ **do**
2: $\quad D^{\Delta}_{\texttt{src}}, N^{\Delta}_{\texttt{src}}, N\_k^{\Delta}_{\texttt{src}}, \texttt{T}_{\texttt{src}} \leftarrow$ INITIALIZE(G, ρ)
3: $\quad D^{0}_{\texttt{src}}, N^{0}_{\texttt{src}}, \texttt{T}_{\texttt{src}} \leftarrow$ FIND_SP$(\texttt{src}, \rho)$
4: $\quad D^{\forall\Delta>0}_{\texttt{src}}, N\_k^{\forall\Delta>0}_{\texttt{src}}, N^{\forall\Delta>0}_{\texttt{src}} \leftarrow$ FIND_QSP$(src, \texttt{T}_{\texttt{src}})$
5: $\quad \rho$-GB$\leftarrow$ACCUMULATE$(\rho, \texttt{src}, D^{\Delta}_{\texttt{src}}, N^{\Delta}_{\texttt{src}}, N_k^{\Delta}_{\texttt{src}}, \rho$-GB$)$
6: **function** ACCUMULATE$(\rho, \texttt{src}, D^{\Delta}_{\texttt{src}}, N^{\Delta}_{\texttt{src}}, N_k^{\Delta}_{\texttt{src}}, \rho$-GB$)$
7: $\quad$ **for** $\texttt{dest} \leftarrow 1, \texttt{numNodes}$ **do**
8: $\quad\quad$ **for** $k \leftarrow 1, \texttt{numNodes}$ **do**
9: $\quad\quad\quad$ **if** $v_k \neq v_{src}$ & $v_k \neq v_{\texttt{dest}}$ & $v_{\texttt{src}} \neq v_{\texttt{dest}}$ **then**
10: $\quad\quad\quad\quad$ **for** $\Delta \leftarrow 0, \rho$ **do**
11: $\quad\quad\quad\quad\quad$ **if** $\exists$ SP through $v_k \parallel \exists$ QSP through v_k **then**
12: $\quad\quad\quad\quad\quad\quad \rho\text{-GB}_k \leftarrow \rho\text{-GB}_k + \frac{N\_k^{0}_{\texttt{src,dest}} + N\_k^{\Delta}_{\texttt{src,dest}}}{N^{0}_{\texttt{src,dest}} + N^{\Delta}_{\texttt{src,dest}}} \cdot \frac{D^{0}_{\texttt{src,dest}}}{D^{\Delta}_{\texttt{src,dest}}}$

$\Delta > 0$, to the ones regarding the *quasi*-shortest paths of cost $\delta^*_{i,j} + 1 \leq \delta_{i,j} \leq \delta^*_{i,j} + \rho$. Yet, $N\_k^{\Delta}_{\texttt{src}}$ represents $\Delta_{max} = \rho + 1$ matrices with size $\texttt{numNodes} \times \texttt{numNodes}$, where each matrix represents one source node. In these matrices, each element is the number of paths between $\texttt{src}$ and all other nodes that v_k falls on. Hence, $N\_k^{\Delta}_{\texttt{src}}$ for a given ρ is represented by:

$$N\_k^{\Delta}_{\texttt{src}} = \begin{bmatrix} n_{\texttt{1,1}}(v_k) & \dots & n_{\texttt{1,numNodes}}(v_k) \\ \dots & \dots & \dots \\ n_{\texttt{numNodes,1}}(v_k) & \dots & n_{\texttt{numNodes,numNodes}}(v_k) \end{bmatrix}.$$

Vector $\texttt{T}_{\texttt{src}}$ is composed of $\texttt{numNodes}$ elements and it contains the maximum allowed cost for the *quasi*-shortest path between $\texttt{src}$ and all other nodes, i.e., $\texttt{T}_{\texttt{src}} = [\delta_{\texttt{src,1}} + \rho, \dots, \delta_{\texttt{src,numNodes}} + \rho]$. Finally, ρ-GB is a vector composed of $\texttt{numNodes}$ elements, where each element is the ρ-geodesic betweenness of an intermediary node v_k.

Algorithm 2 uses as input the matrix representation of the network, G, and the spreadness factor, ρ. It returns the vector ρ-GB, which contains the ρ-geodesic betweenness of every node. The function INITIALIZE is responsible to create the arrays and initialize them with the proper values. As we use the number of hops as the cost of the path, the function FIND_SP implements the Breadth-First Search (BFS) algorithm to find the shortest paths, while the function FIND_QSP implements the Depth-First Search (DFS) algorithm, constrained in depth by vector T, to find the *quasi*-shortest paths. Note that FIND_SP uses ρ as input only to compute the maximum allowed length of the *quasi*-shortest paths. If a real cost was used, for instance, these functions must be changed. A possible candidate is to modify the Dijkstra algorithm to compute paths that costs more than the shortest one.

The `ACCUMULATE` function described in line 6 of Algorithm 2 is responsible for summing up the contribution of each pair of nodes v_i, v_j to the betweenness of the intermediary node v_k. Each time this function is called it updates the ρ-geodesic betweenness of the intermediary nodes that participate in the paths from `src` to all the other nodes. Consequently, in the end of the `for` loop in line 1 of Algorithm 2, the vector ρ-`GB` will contain the ρ-geodesic betweenness of every node.

The analysis of Algorithm 2 shows that the functions `INITIALIZE`, `FIND_SP` and `FIND_QSP`, and `ACCUMULATE` are, respectively, $O(n^2)$, $O(m+n)$, and $O(\rho n^2)$, where ρ is a constant. Hence, the complexity of the ρ-geodesic betweenness metric can be computed in $O(n^2)$ or $O(n^3)$, depending if the network is sparse or dense, respectively.

Chapter 7

ρ-Geodesic Betweenness: Characterization and Discussion

We explore the ρ-geodesic betweenness to verify the impact of the *quasi*-shortest paths on the importance of nodes in both dynamic and static networks. To this end, we characterize the metric according to several aspects. Each one is investigated in this Chapter in a separate section. We then discuss some possible applications of our metric.

7.1 Analysis guidelines

We analyze the impact and relevance of our metric on four datasets using $\rho \leq 5$. As we use the number of hops as cost metric, $\rho \in \mathbb{N}$, and we account for all *quasi*-shortest paths for which $\delta_{i,j} \leq \delta^*_{i,j} + \{1, 2, 3, 4, 5\}$. Our analysis captures the importance of nodes according to their topological distribution in the network, using multiple paths. We use the traditional betweenness as the baseline centrality metric to assess the characteristics of the ρ-geodesic betweenness. We use the following guideline:

1. *Metrics correlation*: we analyze the correlation between the random walk, distance-scaled and ρ-geodesic betweenness with the traditional betweenness. The goal is to discover how closely related the metrics are to the traditional betweenness. This is important to verify if they are using similar characteristics to determine the importance of a node. In addition, we want to verify whether they can pinpoint nodes to be reranked, even if they are strongly correlated to the traditional betweenness.

2. *Metrics reranking ability*: we investigate the behavior of the ranking obtained for each metric, studying the level of agreement between the metrics and the

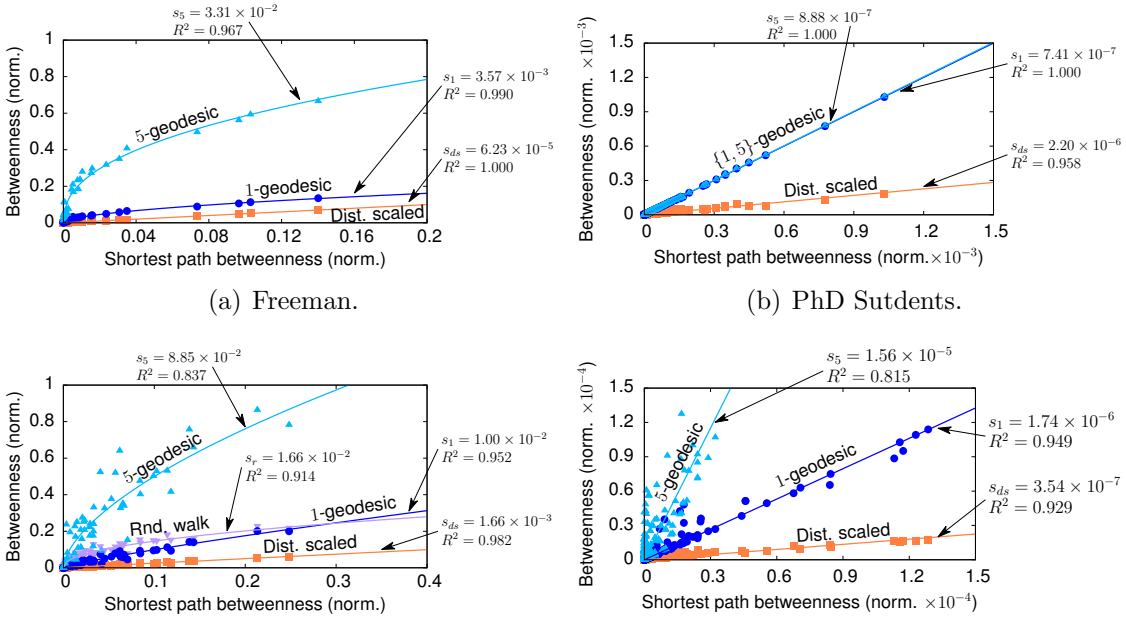
reranking of nodes. Note that the rank position of a node depends on the value of betweenness assigned to it, such that the first node (most important) has the highest betweenness. Nodes can be reranked in higher or lower positions, according to its new value of betweenness.

3. *Intermediation ability*: we examine the distance-scaled and ρ-geodesic betweenness to determine whether they can pinpoint nodes that are potentially a better choice to intermediate flows than the ones preferred by the traditional betweenness. We also investigate for how long nodes can keep the same position. A node can be a better choice to intermediate flows if it remains in the same rank position for longer or, at least, loses its intermediation ability less frequently. We do not use the random walk betweenness for this analysis because the dynamic dataset is not a connected network and, thus, the random walk betweenness cannot be computed using Newman's algorithm (Subsection 6.1.4), considering the entire dataset as a single network.

4. *Fault tolerance analysis*: in this analysis we aim to investigate the performance of a dynamic network in the presence of single node failures. We assume that flows travel through shortest paths (geodesics). We investigate the ranking of the articulation points for each metric and the impact on the network throughput (bandwidth consumption) when a single fail happens on the nodes claimed as the most central by each metric.

7.2 Metrics correlation

It is important to know how the metrics relate to the traditional betweenness to find how the different considerations of each metric influence the similarity between them. We need the metrics to be strongly correlated, so that we are sure that they measure similar characteristics. Nevertheless, we also need that each metric pinpoints nodes that should be reranked according to its own definition of node importance. The results are shown in Figure 7.1, where the X-axis is the normalized traditional betweenness and each curve represents one of the other three metrics, also normalized. The normalizing factor is given by $0.5 \cdot (|\mathcal{V}| - 1) \cdot (|\mathcal{V}| - 2)$ for the undirected graphs and by $(|\mathcal{V}| - 1) \cdot (|\mathcal{V}| - 2)$ for the directed ones, as explained in Chapter 6. The axes in Figures 7.1(b) and 7.1(d) are scaled for better visualization. The random walk betweenness is computed only for the Dolphins dataset due to restrictions of Newman's algorithm related to the network structure, as we discussed previously in this Chapter. We only show the curves for the 1- and 5-geodesic betweenness ($\rho = \{1, 5\}$, respectively), for the sake of clearness. The curves for the other values of ρ lie between these two.



(a) Freeman. (b) PhD Sutdents.

(c) Dolphins. (d) Cologne #1.

Figure 7.1: The correlation with the traditional betweenness is clearly strong for all metrics, being stronger for the distance-scaled betweenness. The random walk and ρ-geodesic betweenness show more capability to identify nodes that could receive a different value for the betweenness. Note that the axes are normalized by $\big(|\mathcal{V}|-1\big)\cdot\big(|\mathcal{V}|-2\big)$ if the graph is directed, and by $0.5\cdot\big(|\mathcal{V}|-1\big)\cdot\big(|\mathcal{V}|-2\big)$ otherwise.

Figure 7.1 shows that all metrics are strongly correlated to the traditional betweenness, as their coefficient of determination (R^2) is high. The strongest correlation is found for the distance-scaled and 1-geodesic betweenness. As ρ increases the correlation decreases, since the participation of nodes in *quasi*-shortest paths increases, pinpointing more nodes to be reranked. Nonetheless, $\rho = 1$ is already enough to pinpoint some nodes. This is shown by the dispersion of points around the curve or, mathematically, by the standard deviation of each fitting, combined with the R^2 value. By comparing these parameters for each curve, we find that the higher the standard deviation and the lower the R^2, the more nodes are identified as candidates to be reranked. Note, that R^2 cannot be very small (< 0.35), as it would give us only a moderate correlation, meaning that the metrics are almost completely different, which is not of our interest. If the metrics are completely different from the traditional betweenness, we cannot assure that they keep the idea of flow control.

Figure 7.1(b) shows a singular behavior. The ρ-geodesic betweenness is quite identical to the traditional betweenness in the PhD Students network, independently of ρ. This happens because the relationships between nodes in this network have strong socialization tendencies, which turns out to produce few multiple paths. The correlation between the random walk and traditional betweenness can be observed in Figure 7.1(c). In this scenario, we note that this metric is similar to the 1-geodesic

betweenness ($\rho = 1$), and both can almost equally identify that some nodes could be reranked.

7.3 Metrics reranking ability

Knowing that the ρ-geodesic betweenness can identify nodes that could be reranked, we further investigate the level of agreement between the metrics regarding the reranking. We also analyze the influence of ρ on the ranking, which is established using the betweenness of the nodes, such that the most important node has the highest betweenness and is the first in the rank, while the node with the lowest betweenness is the less important and, thus, the last in the rank.

We use the node ranking for the Dolphins network to analyze the level of agreement between the metrics. To this end, we compute the Kendall's W coefficient for each pair combination of the metrics. Hence, using the terminology of the coefficient, the metrics are the judges and every node in the dataset is judged by them. Kendall's W defines if the judges agree with the classification assigned to the set of nodes, i.e., if each metric agrees with the ranking provided by the other metrics. In Figure 7.2, the closer to the border the blue octagon is, the higher the level of agreement between the metric and the others. The ranking provided by the distance-scaled betweenness, for instance, is almost in perfect agreement with the one for the traditional betweenness. The disagreement between the random walk and the traditional betweenness is higher than the one between the 1-geodesic betweenness and the traditional betweenness. As ρ increases, in turn, the disagreement with the traditional betweenness also increases, because the *quasi*-shortest paths accounted become significant. This also happens if we compare the concordance between the random walk betweenness and our metric. This discussion does not reflect, however, the rate with which nodes are reranked. Although the concordance between the metrics is high, the reranking rate is also high. For instance, we found that compared to the traditional betweenness, several nodes are reranked independently of the metric we use. We have a reranking rate of 66.1% using the distance-scaled betweenness, 75.8% for the random walk betweenness, and for the ρ-geodesic betweenness we have 74.2%, 77.4%, 79.0%, 75.8%, and 77.4% for $\rho \in \{1, 2, 3, 4, 5\}$, respectively. This happens because, contrary to Kendall's W coefficient, the reranking rate does not take into account whether the change in the rank position is significant.

In order to investigate the intensity of the reranking, we analyze in Figure 7.3 how the rank varies according to the metrics we use. The X-axis represents the transition between the metrics, while the Y-axis shows the number of positions that a node gained or lost when we change from one metric to the other. The color grid shows how frequently the nodes gain or lose y positions. Figure 7.3(a) illustrates the

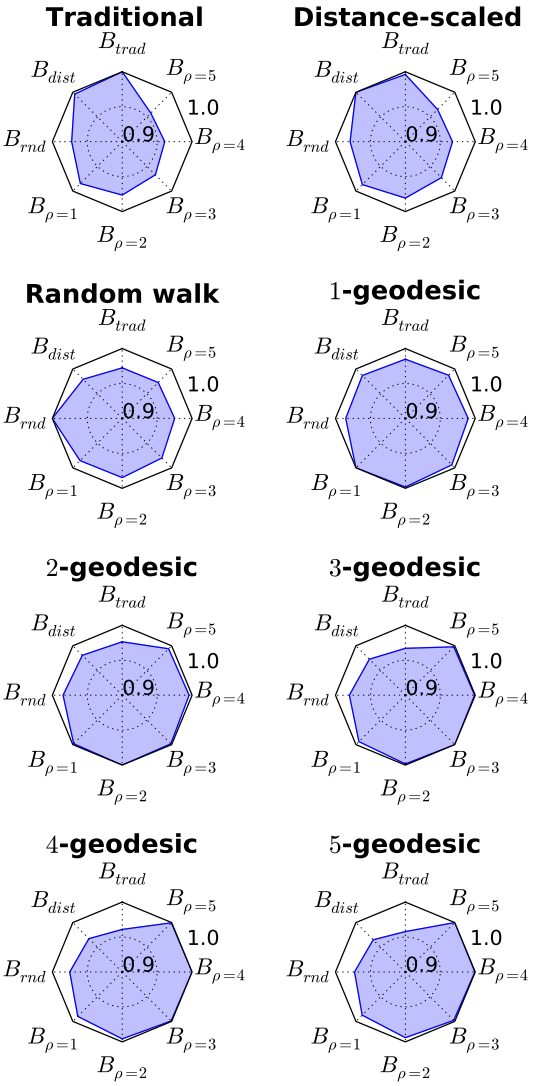
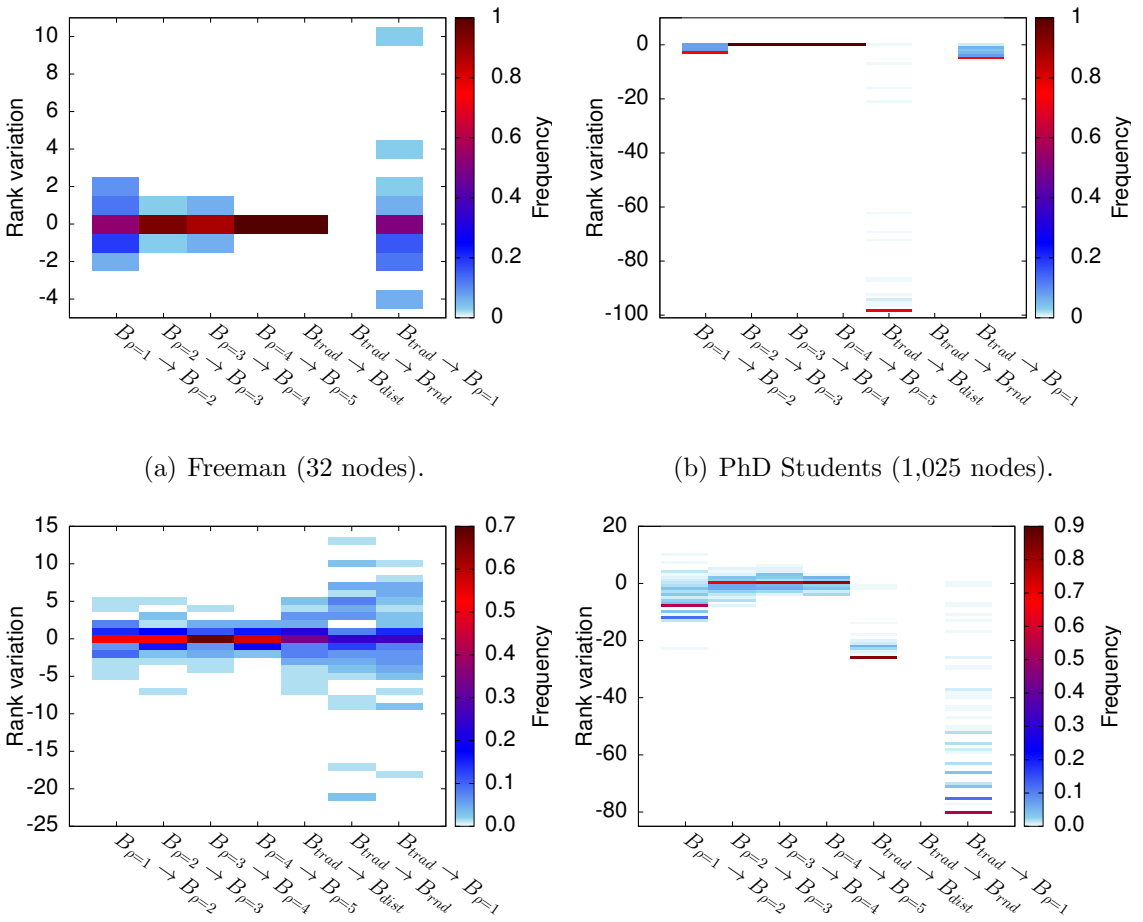


Figure 7.2: The Kendall's W coefficients plotted to each pairwise combination of betweenness centrality metrics show a high level of agreement between them. The lowest concordance happens between the traditional and the ρ-geodesic betweenness for $\rho = 5$. This is due to the potentially numerous *quasi*-shortest paths considered on the metric computation, which vary more significantly the importance of nodes and, consequently, their rank positions.

results for the Freeman dataset. We observe that at least half of the nodes keep the same position when we change from the traditional to the 1-geodesic betweenness ($\rho = 1$), as shown by the purplish color for $y = 0$. Note that we can find nodes that gain up to 10 positions if we use the proposed metric. In turn, if we use the distance-scaled betweenness, 100% of nodes stay in the same position. We also observe that increasing ρ affects the ranking with nodes losing or gaining up to 2 positions. The variation stops at $\rho = 4$, as for $\rho = 5$ the influence of *quasi*-shortest paths ends. We highlight that, in all scenarios, for $\rho > 2$ most nodes keep their positions unchanged,



(a) Freeman (32 nodes).

(b) PhD Students (1,025 nodes).

(c) Dolphins (62 nodes).

(d) Cologne #1 (1,579 nodes).

Figure 7.3: The distance-scaled, random walk and ρ-geodesic betweenness are able to redistribute the node ranking to different extents, compared to the traditional betweenness.

as shown by the reddish rectangles for $y = 0$.

Figure 7.3(b) shows the result for the PhD Students dataset. Some nodes change their position when we use the ρ-geodesic betweenness, but the distance-scaled betweenness has the most significant influence on the ranking for this scenario, as it spreads the classification. This corroborates the correlation results found for this dataset. We observe in Figures 7.3(c) and 7.3(d) that all metrics change significantly the node ranking on the Dolphins and Cologne #1 networks, respectively. For the Dolphins dataset the random walk and ρ-geodesic betweenness redistribute several nodes in the rank. The variation on the positions of the rank is representative, with a range that lies, approximately, between $[-21, 13]$ for the random walk betweenness and $[-18, 10]$ for our metric. In the Cologne network, the ρ-geodesic betweenness has more power of modification compared with the other metrics, ranging from -80 to 28 positions. Note that the number of nodes in each dataset is very different and changing a certain number of positions in each one has a different impact. For

instance, if we disregard ties, i.e., each position can be occupied by one node only, a node in the Freeman dataset promoted by 10 positions improves its importance by 31.25%. In contrast, a node in the sample 1 of the TAPASCologne dataset, in the same condition, would improve its importance by only 0.64%. We do not imply that the results observed in Figure 7.3 elects one or other metric as the best centrality metric. We only intend to present the power of changing nodes ranking for each one of them. Such reranking reflects what the metric defines as an important node. To be the best metric depends on the analyzed application.

We further investigate how the centrality metrics evaluated behave in the presence of ties. The idea is to find how the metrics assess the importance of nodes once tied in the same rank position by the traditional betweenness. If the metrics are able to differentiate more nodes, the node ranking will be broadened. Consequently, considering that one of these nodes could be chosen to execute a functionality, it would be possible to choose from a finer-grained list of options. The result is shown in Figure 7.4 for all datasets, considering $1 \leq \rho \leq 5$. We define the rate of broken ties as $1 - (\#\ \texttt{tied nodes other metrics} / \#\ \texttt{tied nodes trad metric})$. As the values are consistent for all ρ in this range, we label the blue bar in the graph as ρ-geodesic. We observe that the tiebreaking rate for the ρ-geodesic betweenness is usually equal or greater than the other metrics, even though we use a small spreadness factor ($\rho = 1$). The only exception is the PhD Students scenario, singular in its construction, which does not provide many multiple paths that could benefit the ρ-geodesic betweenness. Note that, in this figure, a negative tiebreaking rate means that the number of tied nodes increased.

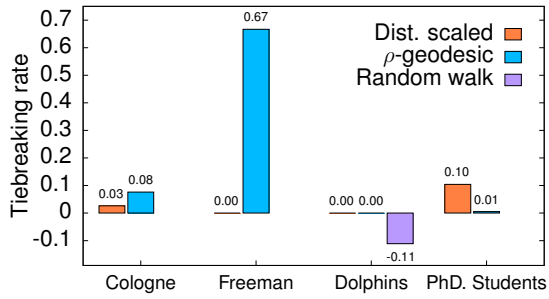


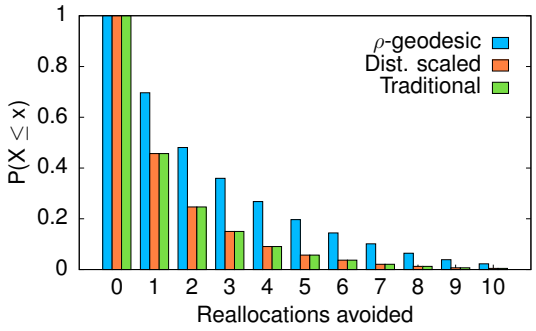
Figure 7.4: Compared to the traditional betweenness, the distance-scaled and ρ-geodesic betweenness are able to spread the classification rank, giving room to more positions. Hence, we find less nodes tied in the same position. The random walk betweenness surprisingly increases the number of tied nodes in the Dolphins dataset.

7.4 Intermediation ability

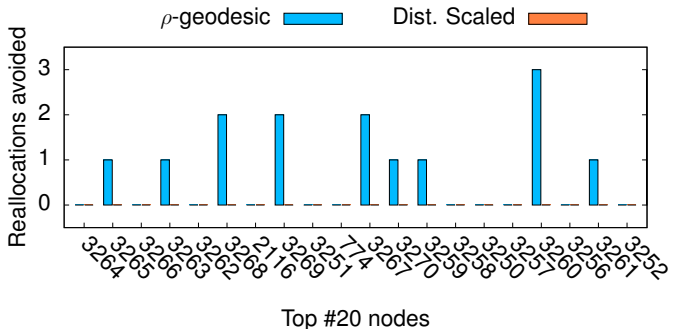
The ρ-geodesic betweenness is strongly correlated to the traditional betweenness but it moves away from the latter as ρ increases. As such, we can identify nodes that could be given more (or less) importance. Even for low values of ρ, such as $\rho = 1$, for instance, we can find nodes that can be reranked, but to a lesser proportion. In the following analysis, we suppose that nodes are chosen to play a specific role on the network based on their betweenness. This role could be, for instance, to act as a packet flow analyzer in a network where flows follow shortest paths. The placement of this functionality is determined by node centrality, so that nodes with high betweenness are better candidates because they much likely fall on the majority of shortest paths between other nodes in the network. We investigate, then, how the betweenness metrics capture node centrality overtime, which can also be interpreted as the ability to intermediate flows. To this end, we took 10 samples from the Cologne network, each 10-seconds.

Nodes that do not fall on any shortest path cannot intermediate flows and it is expected that the number of nodes that participate in shortest paths influences how many nodes can play specific roles. In dynamic networks, or in the presence of node failure, having less options of nodes to play specific roles can lead to service disruption or even break the network into several connected components. Note that services in this context can point to different functionalities, such as data storage, packet forwarding and flow analysis.

Let us continue our analysis supposing that packet analyzers were allocated on the most central nodes, defined by the traditional betweenness. If in the next instant some of these nodes come across an event that reduces their ability to intermediate flows, we will need to reallocate the packet analyzers. This can happen, for instance, because they became a leaf node or the number of paths through them was severely reduced. In this scenario, we examine the *intermediation ability* through the analysis of the *number of reallocations avoided* over time. We use both expressions interchangeably. Figure 7.5(a) is a cumulative distribution that shows how often we can avoid to reallocate the packet analyzers. The X-axis is the number of times that we could avoid reallocations, i.e., prevented nodes to lose their intermediation ability. The Y-axis represents how often at least x reallocations were potentially avoided during the period in analysis ($P(X \leq x)$). While $x = 0$ means that the reallocations were never prevented, $x = 10$ means that nodes never lost their intermediation ability. Clearly, the ρ-geodesic betweenness is always capable of avoiding more reallocations than the other metrics. Of course, this result is of poor use if the nodes that keep the intermediation ability are the ones that are never used because their betweenness is very small, ergo the last nodes in the rank. Therefore, we an-



(a) Prevention of loss of intermediation ability.



(b) Number of times the loss of intermediation ability was prevented for the top 20 nodes.

Figure 7.5: The proposed ρ-geodesic betweenness is able to reduce the number of times that nodes lose their ability to intermediate flows in the network compared to the other metrics, even for the most important nodes. The ρ-geodesic betweenness can prevent up to 3 reallocations more than the traditional and distance-scaled betweenness.

alyze the top initial 20 nodes in the network to further verify if they also benefit from this behavior.

Figure 7.5(b) shows how many times more we can keep nodes intermediation ability when we change from the traditional to the distance-scaled and ρ-geodesic betweenness, for the initially top-ranked nodes in the network. The X-axis is the node label in ascending order of importance. The Y-axis indicates the difference between the number of times the reallocation was avoided by the distance-scaled and ρ-geodesic betweenness compared to the traditional betweenness, in absolute values. We observe that for the top 20 nodes, the distance-scaled betweenness never avoids more losses than the traditional betweenness. On the other hand, the ρ-geodesic betweenness is able to avoid more than 1 reallocation to almost half of these nodes, reaching up to 3 reallocations avoided.

Lastly, we analyze for how long nodes can remain in the same rank position using each metric. Figure 7.6 shows how frequently we can find nodes that can keep the same position during the 90 seconds time interval analyzed for the Cologne dataset. The majority of nodes in this interval frequently jumps between rank positions and

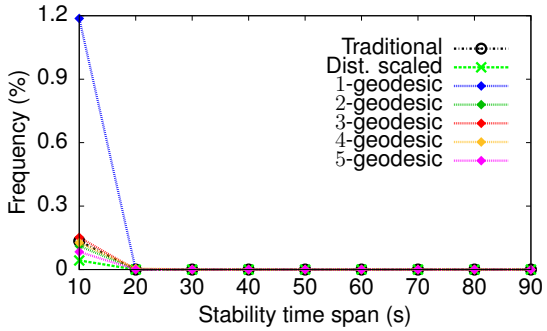


Figure 7.6: In the highly dynamic scenario of the Cologne vehicular network, the ρ-geodesic betweenness is the metric that changes less frequently the ranking of nodes. Of all nodes in the network, 1.2% remain in the same rank position for up to 10 seconds.

none of them is able to maintain the same position for more than 20 seconds. Hence, in Figure 7.6, we only show the results for $10 \leq x \leq 30$. We observe that few nodes remain in the same rank position and they do so for approximately 10 seconds maximum. Although this is valid for the minority of nodes in this network, we can quickly note that the ρ-geodesic betweenness is the metric that achieves the highest number of nodes that can keep the same rank position for longer, reaching 1.2% of nodes for $\rho = 1$, which is 6 times higher than the traditional and $\{3, 4\}$-geodesic betweenness. Therefore, we claim that the ρ-geodesic betweenness is the metric that can better keep the ranking unchanged for the highly dynamic scenario provided by the Cologne network.

7.5 Fault tolerance analysis

Suppose the scenario of Section 7.4, where we have allocated packet analyzers on the nodes that participate the most on shortest paths. The idea of this allocation choice is because flows tend to follow shorter paths. Nodes with high betweenness are responsible for intermediating a great number of flows between pairs of nodes in the network. The problem with that approach is that the proportion of participation on shortest paths is related to the criticality of the node to the network connectivity [80]. Hence, depending on the topology of the network, a failure on a single node can affect a great number of flows. As the identification of critical nodes is not trivial [81, 82], one can use the traditional betweenness to identify some of them [80], because they usually have higher betweenness than ordinary nodes. We use, then, the Cologne network to investigate the relation between the articulation points and the traditional, distance-scaled and ρ-geodesic betweenness centralities.

We analyze only the first 6 snapshots (10 seconds each) from the Cologne network. Figure 7.7 shows the result considering the top #5 nodes of the metric in

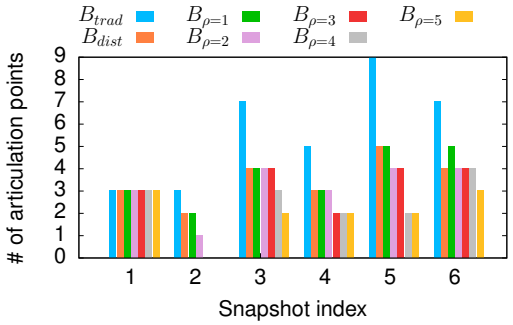


Figure 7.7: The number of articulation points existing on each snapshot of the dynamic Cologne network considering the top #5 positions. In general, this number is lower for the distance scaled (B_{dist}) and ρ-geodesic (B_{ρ}) betweenness, when compared to the traditional betweenness (B_{trad}).

each snapshot. The X-axis is the snapshot index and the Y-axis is the number of articulation points existing among the top #5 nodes in each snapshot. We observe that the number of articulation points is different according to the snapshot and the metric used. It is clear that both the distance scaled and the ρ-geodesic betweenness have less articulation points in the top positions. Table 7.1 shows the number of components existing before and after a failure on a single articulation point in the top #5 positions. In each snapshot, at least one more connected component is created. In snapshot #4, up to two components can be created, depending on the node classification.

A single node failure can have severe impact on the communications, depending on the network topology. If this node is also an articulation point, the failure can be fatal to the network operation. We investigate, then, the severity of single failures on nodes selected as the most central by each betweenness metric. To this end, we use the Network Simulator (NS-3) aiming to analyze the behavior of the average system throughput before and after a single node failure on each of the top #5 positions. The average system throughput is the rate of successful message delivered over the communication medium, measured in bits per second. We evaluate the results for each metric by comparing them to the traditional betweenness. We consider that each node in the network intends to communicate with all the other nodes, even if they are not in the same connected component. Nevertheless, if two nodes in

Table 7.1: Number of components in each snapshot of the Cologne network before and after failure on a single articulation point.

	Snapshot index					
	#1	**#2**	**#3**	**#4**	**#5**	**#6**
No failure	550	549	546	576	596	578
Single failure	551	550	547	577 / 578	597	579

Table 7.2: The probability that a failed node is also an articulation point is higher in the majority of cases for the traditional betweenness, compared to the other metrics.

Snapshot	B_{trad}	B_{dist}	$B_{\rho=1}$	$B_{\rho=2}$	$B_{\rho=3}$	$B_{\rho=4}$	$B_{\rho=5}$
#1	0.60	0.60	0.60	0.60	0.60	0.60	0.60
#2	0.43	0.33	0.33	0.17	0.0	0.0	0.0
#3	0.88	0.80	0.80	0.80	0.80	0.38	0.25
#4	0.72	0.60	0.60	0.60	0.40	0.40	0.40
#5	1.0	1.0	1.0	0.67	0.80	0.33	0.33
#6	1.0	0.80	1.0	0.80	0.80	0.80	0.43

different components try to communicate they will not succeed because there is no path between them. Hence, we install one UDP server-client application for each pair of nodes. All messages exchanged between them have the same size. We consider that the topology changes each 10 seconds, to match our snapshots. Lastly, the node to be failed is chosen randomly among each rank position. This means that if we want to investigate the failure of a node in position n, and there are m nodes in this position, we randomly chose one of the m nodes to fail. The results depend on the ranking of each metric. In sparser networks, which have less alternative paths between pairs of nodes, we should have a behavior similar to the traditional betweenness. On the other hand, if the offer of alternative paths is higher, the difference between the metrics will be more accentuated.

We investigate the average network throughput after a failure on a single node in the top #5 positions. Note that the failed node can be an articulation point or not. The probability that the failed node is also an articulation point differs for each metric, as shown in Table 7.2. Comparing the traditional betweenness with the distance scaled and the ρ-geodesic betweenness, this probability is often higher for the traditional betweenness. Focusing on the distance scaled and the ρ-geodesic betweenness, in the studied scenario, the probability that a failed node is also an articulation point is usually higher for the distance scaled betweenness. In addition, regarding the variation within the ρ-geodesic betweenness itself, this probability is reduced when ρ increases. Therefore, we expect the impact of failure on the throughput of the network to be less harsh for the ρ-geodesic betweenness.

In Figure 7.8 we highlight the maximum average throughput of the network, achieved when there is no node failure, equal to 21.17 Mb/s. We use it as the reference value. We observe that a single failure can be disastrous on the average network throughput, independently of the node rank. The variation among the node ranks is very small and the throughput loss is lower for less important nodes. For instance, the worst failure in the first rank causes the drop of the average throughput to 28.18% of its maximum original value, while for the last rank it is reduced to 28.40% in the worst case. Frequently, nodes classified as more important are the

same for all metrics, and, thus, the variation of the average throughput within the same rank is small. Figure 7.9 shows how often the top #5 nodes change during the 60 seconds of simulation. During 83% of the time, the nodes classified on the first position is the same for both the traditional and distance scaled betweenness. Compared to the ρ-geodesic betweenness, it drops to 67% of the time. This value is not enough, however, to influence the average network throughput. This throughput is influenced only when the coincidence is smaller than 50% of the time.

The major point of this section was to highlight that the ρ-geodesic betweenness is able to reduce the number of articulation points in the most central positions. Additionally, although the throughput results are similar to all metrics, in the evaluated scenario, failure on nodes with high ρ-geodesic betweenness usually reduces slightly less the average throughput of the network. It is important to design a network where no articulation points exist, but if they do, it is also important not to place critical functionalities on these points, as they represent major vulnerabilities. Therefore, one should not use the traditional betweenness for such placement. The ρ-geodesic betweenness, in turn, could be used, as it elects less articulation points as central nodes and the cost to deviate the flow from the shortest path to a *quasi*-shortest path is potentially small, depending on the spreadness factor ρ.

7.6 Discussion

The main goal of Part II is to propose a new betweenness metric to determine node importance, considering, in addition to shortest paths, longer paths between nodes in multipath networks. This is important because, in many networks, flows do not follow only shortest paths. In the literature, we already have a well-known centrality metric, the random walk betweenness, presented in Chapter 6, that accounts for longer paths, besides the shortest ones. One may ask, then, why bother to create another metric to account longer paths if the random walk betweenness proposed

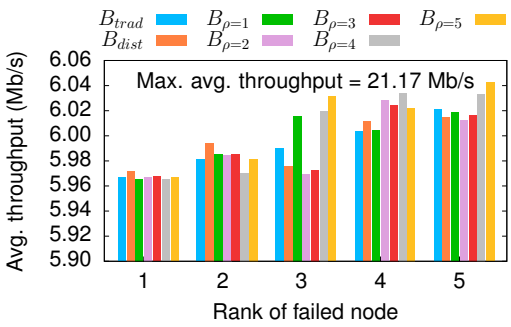


Figure 7.8: Compared to the traditional (B_{trad}) and distance scaled (B_{dist}) betweenness, a single failure on one of the top #5 nodes of the ρ-geodesic betweenness (B_{ρ}) is generally less harsh to the average network throughput.

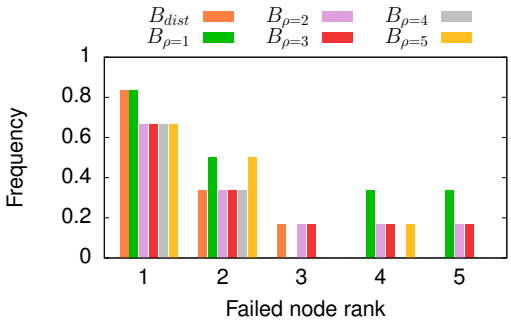


Figure 7.9: Nodes classified in the first position according to the distance scaled and ρ-geodesic betweenness are frequently the same as the traditional betweenness, considering all samples of the Cologne dataset. For lower ranks, the frequency is significantly reduced.

by Newman [40] does the same job gracefully. The main reason for that is the basis conjecture behind the random walk betweenness itself, which cannot be applied in some scenarios. Newman states in his work [40] that his metric suits well situations where information may follow random paths until it finds its destination, and he considers that information may not know where it is going to. This may be true if no global knowledge about the network structure exists, or if we are trying to model the natural spread of diseases, for instance. As a counter example, in a computer network where end-to-end paths do exist, and the source of information knows exactly who is the target of the message, it will always *try* to use the most efficient path. Similarly, in transport networks, a driver or a delivery vehicle will always be more interested in using the shortest path. In both networks, a packet delivery model fits better the flow process in the network. Nevertheless, in these networks, and others, it is not always that the shortest path is the best option. For instance, if every packet delivery happens using the shortest paths, it is possible that they become congested, being no longer a good option. This is better discussed in Subsection 7.6.2. Hence, it is interesting to consider longer paths, in addition to the shortest ones.

The random walk betweenness considers *all* existing paths between any pair of nodes in the network, no matter the path length. The contribution of each path to the importance of a node is proportional to the probability of using the path. This probability, in turn, varies simultaneously with the cost of number of hops in the path and the degree of the nodes on it. If successive nodes have high degree or if the path is long, the contribution will be lower. Unlike random walk betweenness, the ρ-geodesic betweenness weights the contribution only as a function of the path cost. In addition, we reduce the number of paths according to the spreadness factor. Hence, considering that number of hops is used as cost metric, the contribution of longer paths tends to decrease more quickly for the random walk betweenness, as

long as the nodes on the path have degree greater than 2.

To incorporate the several paths considered by the random walk betweenness we can use mainly two methods. The first one simulates several random walks between pairs of nodes. This method allows for the computation of approximated values of the random walk betweenness in a distributed fashion [83, 84]. Either sequential or distributed algorithms, however, require extra attention to not allow the random walks to loop over the same sequence of nodes, which would erroneously increase the importance of nodes that are traversed many times. Moreover, we need to be able to stop the simulation at a step where the values computed for the random walk betweenness approximate the exact value (given by Newman's algorithm [40]). Note that the convergence time can be unfeasible for some applications when using this approach [83]. The second method computes the metric using Newman's algorithm, which applies a matrix approach in a very elegant fashion. This approach, however, is neither appropriate for disconnected nor directed graphs, due to the generation of null determinants that prevents further computation. The complexity of this algorithm is $O((m+n)n^2)$, which is roughly $O(n^3)$ in sparse and $O(n^4)$ in dense networks. The ρ-geodesic betweenness, in turn, does not have any restriction regarding the structure of the network. Additionally, since it only considers paths *up to* a length and not all the paths as the random walk betweenness, it is less time consuming. Taking a look at Algorithm 2 (Chapter 6), we observe that the functions `INITIALIZE`, `FIND_SP` and `FIND_QSP`, and `ACCUMULATE` are, respectively, $O(n^2)$, $O(m+n)$, and $O(\rho n^2)$, where ρ is a constant. Hence, the complexity of the ρ-geodesic betweenness metric is reduced compared with Newman's algorithm and it can be computed in $O(n^2)$ or $O(n^3)$, depending if the network is sparse or dense, respectively. The complexity of our metric can be further reduced if the algorithm is parallelized, which is a matter of parallelizing the single-source shortest paths (SSSP) and the accumulation functions in Brandes' algorithm [79], considering unweighted networks. This is feasible [85–88] and the graph traversal performed in the SSSP needs to be run $\rho + 1$ times to find all the paths that we need to compute the ρ-geodesic betweenness. In addition, if only local knowledge is available, it is possible to modify a distributed algorithm as the one proposed by Lehman and Kaufman [89] to compute our metric.

7.6.1 Random walk and ρ-geodesic betweenness centralities in scale-free networks

As we previously discussed, the random walk and the ρ-geodesic betweenness are quite different, even though their purpose is to account non-ideal paths. The main differences between them is the number of paths considered in the computation and

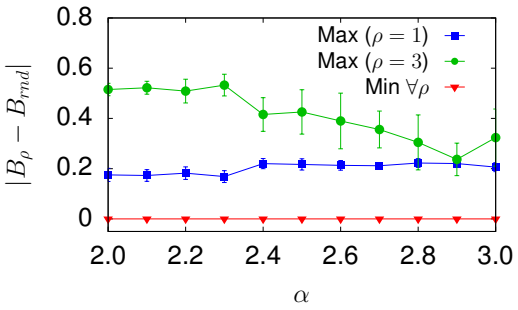


Figure 7.10: Averaged results for the maximum and minimum absolute differences between the random walk and the ρ-geodesic betweenness, for $\rho = \{1, 3\}$. The minimum difference is independent of ρ, while the maximum difference becomes more significant for higher ρ.

the weight assigned to each one of them. We investigate the impact of this difference on the importance of nodes in synthetic random networks with power law degree distribution ($P \propto degree(v_i)^{-\alpha}$), generated by the Havel-Hakimi[56, 57] algorithm, as we explained in Chapter 3. Such networks are also known as scale-free networks and they are the most common networks in the real world.

We compute the absolute maximum and minimum differences between the values assigned by the random walk and the ρ-geodesic betweenness, considering all nodes in each random graph. Figure 7.10 shows the averaged results for each α. We observe that the minimum difference is always close to zero, while the maximum difference depends on the value of ρ. Considering the 1-geodesic betweenness ($\rho = 1$) we note that the maximum difference between the metrics remains almost constant for all α. For the 3-geodesic betweenness ($\rho = 3$), the maximum difference becomes more significant for $2 \leq \alpha \leq 3$. We believe that within this interval the number of longer paths considered by the ρ-geodesic betweenness becomes much greater than the ones for the random walk betweenness. As such paths can be used as backup or offloading paths, nodes that participate on them should be valuable.

7.6.2 Applications

Given that the random walk and the ρ-geodesic betweenness are different, although both use longer paths to assess node importance, each metric should be applied in different situations. We argue that, in some applications, it is reasonable to exclude all paths longer than a threshold to compute the importance of a node, as these paths are much likely neglected. Hence, it is not necessary to compute all the paths between every pair of nodes as the random walk betweenness does. The use of *quasi*-shortest paths, or at least considering them as reasonable alternatives, is a choice that can be driven by (i) reactive or (ii) proactive situations. In the first case,

entities try to escape from the common-sense, a.k.a., the shortest path, to avoid unwanted consequences that are already expected to happen. For instance, a packet may be sent through a *quasi*-shortest path if the shortest path between two nodes in a computer network is congested, or a node, or link, in this path is expected to fail. Also, the driver in a transport network can choose slightly longer paths during rush hours to avoid jammed shortest paths. The idea is that it is better to take a little extra time to arrive, when compared to the normal shortest path, than to either risk being blocked on the normal shortest path or being forced to take alternative paths on-the-fly. As for the proactive situations, the idea is to avoid, beforehand, to damage the shortest path in the near future. This situation can happen whenever multiple alternatives exist and any one of them could be picked according to a given criteria. For instance, each packet flow can be sent through different paths so as to prevent congestion in computer networks. In the same sense, the audience of a soccer game may also follow different trajectories using different gates to enter a stadium. Even in social networks we can observe the situation where information may occasionally follow a path that is neither shortest nor random, e.g., the act of a friend telling a secret of a third person to a common friend.

In both proactive and reactive situations, the entity arbitrarily chooses a slightly longer path when there is a high chance that the shortest path is damaged or will be damaged in the near future. When they choose to follow other paths, they start to increase the centrality of other nodes. Thinking about city streets, if each intersection is a node, when cars begin to deviate from the shortest path, other streets will be used, increasing the centrality of the intersections on this new path. We use the spreadness factor to denote the additional cost the entity is willing to pay to arrive at the destination the fastest possible, considering that the shortest path can be damaged in some sense. This can be better understood using an analogy. Suppose that we have a set of pipes, with different diameters, ending in a container. The shorter pipes are also the largest ones, whereas thin pipes are long. We want to fill the container the fastest we can with some kind of solid particle. We cannot push all the particles through the shorter pipe because at some point it will be clogged. Hence, the fastest way to fill the container is to push the particles through all the pipes. Nevertheless, we cannot use some thinner pipes because the solid particles do not fit into them. In this case, the spreadness factor could model the diameter of the pipe, so that only the ones into which the particles fit can be used.

We envision some applications for the ρ-geodesic betweenness that benefit from the use of *quasi*-shortest paths. We briefly discuss two of them hereafter. The first one uses the ρ-geodesic definition as it is. The second excludes the shortest paths from the computation of the metric, using only the *quasi*-shortest paths to assess node importance.

Delivery of goods

Suppose that a truck wants to go from a factory (υ_A) to a store (υ_B) the fastest possible to deliver some goods. The graph through which the truck moves is modeled such that each street is an edge and each intersection or traffic light is a node. The topology of the delivery problem is depicted in Figure 7.11. The truck can use three paths to arrive at the destination, $p_1 = \langle \upsilon_A, \upsilon_B, \upsilon_C, \upsilon_D, \upsilon_E \rangle$, $p_2 = \langle \upsilon_A, \upsilon_B, \upsilon_F, \upsilon_G, \upsilon_D, \upsilon_E \rangle$, $p_3 = \langle \upsilon_A, \upsilon_B, \upsilon_H, \upsilon_I, \upsilon_J, \upsilon_D, \upsilon_E \rangle$. In this scenario we have a black truck that can differentiate into three types of drivers: type `I` (red truck) is ordinary and always follow the path with the highest betweenness nodes, in this case p_1; type `II` (blue truck) is impatient and does not want to risk wasting time on a shortest path that will probably be congested; and type `III` (green truck) is altruist and chooses to not follow the suggested path, trying to avoid congesting it. Note that driver type `II` chooses path p_2 only if p_1 is congested, or will likely be congested. Hence we also represent driver type `II` on path p_1 using dashed lines. Driver type `III` always chooses p_2 because he believes that it is better to take a longer path than to risk contribute to the traffic jam or to be trapped in it. We could use the partial values of the ρ-geodesic betweenness to decide which path should be taken. The values of the 1-geodesic betweenness for these nodes due to the pair υ_A, υ_E, only, are:

- $B_\rho(\upsilon_A) = B_\rho(\upsilon_E) = B_\rho(\upsilon_H) = B_\rho(\upsilon_I) = B_\rho(\upsilon_J) = 0$
- $B_\rho(\upsilon_B) = B_\rho(\upsilon_D) = 0.8$
- $B_\rho(\upsilon_C) = 0.5$
- $B_\rho(\upsilon_F) = B_\rho(\upsilon_G) = 0.4$

Note that the centralities for $\upsilon_H, \upsilon_I, \upsilon_J$ are null because the truck is willing to pay only 1 hop more to arrive at the destination. At each step, the truck could choose the nodes with the highest centrality. Hence, at first, it would follow the path p_1. Nevertheless, when it arrives at υ_B it discovers that there is too much traffic on the link $\varepsilon_{B,C}$. Both type `II` and `III` drivers are willing to pay only 1 hop more for an alternative path. Hence, when they arrive at υ_B, they choose the node with the second highest centrality, υ_F. In the end, they use a *quasi*-shortest path with different intentions, driver type `II` hopes to take less time to deliver the goods, while driver type `III` tries to not contribute to the congestion.

Data concentrators

Suppose now that we have a mobile wireless sensor network deployed in a city, using, for instance, personal vehicles to collect data. We exclude buses because

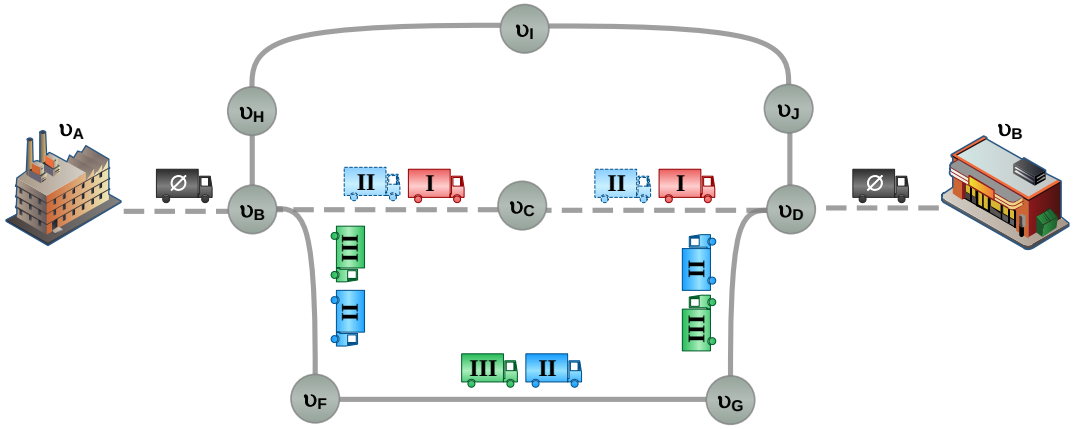


Figure 7.11: Example topology to model the delivery of goods. Three types of truck drivers want to travel from the factory (υ_a) to the store (υ_b) to deliver some products. Driver type `I` (ordinary) always follow paths with the most central nodes; driver type `II` (impatient) does not want to risk wasting time on a possibly congested shortest path; and driver type `III` (altruist) does not want to congest the shortest paths and, thus, follow slightly longer paths.

these vehicles have fixed routes. The mobile sensors need to send the data to be processed in the cloud or the fog. We need to know, then, where to place the data concentrators. Such data concentrators could be seen as cloudlets. We can also consider that we have both Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications. Hence, nodes that are part of the infrastructure could play the role of both data concentrators and forwarders. It is reasonable to think that these functionalities should be placed on the most central nodes, because most vehicles will pass on these nodes, which could lead to the overload of the node. Therefore, we need to separate these functionalities.

We could use the traditional betweenness to determine where the data forwarders should be placed, thus using shortest paths only. After this decision, we could exclude the shortest paths from the computation of the ρ-geodesic betweenness and use the result to decide where to place the data concentrators. We could have, then, the placement of forwarders based on shortest paths only and the placement of concentrators based on *quasi*-shortest paths only. This is reasonable if we notice that data forwarding needs faster routes, while concentrators could have slightly longer delays and, hence, could be placed on slightly longer paths. As cars will not always follow shortest paths between two locations, we can expect that placing concentrators on central nodes that participate in several *quasi*-shortest paths would still allow a huge number of sensors to offload their data while avoiding burdening the forwarders with the additional function of a concentrator. In this scenario, we

could additionally use the results of Part I of this work to better forward packets, using V2V communications, from a mobile sensor to the data concentrator.

Chapter 8

Conclusions and Future Work

This thesis studied on the influence of the network structure on system properties over time. More specifically, we investigated how (i) data forwarding and the (ii) assessment of nodes relative importance are affected by the structure of multihop networks. We also (iii) introduced a study of the network performance under failure of central nodes. In the first analysis, we investigated the impact of node mobility on the establishment of multihop communications, based on the behavior of node vicinity. To this end, we proposed the (κ, λ)-vicinity, which extends the concept of node vicinity to take into account nodes at multihop distance from the ego node and the relative speed between them. In the second analysis, we studied how node centrality is affected by the use of longer paths to assess their importance. To this end, we proposed the ρ-geodesic betweenness centrality, which takes into account all paths between two nodes that are shorter than a threshold. This threshold is defined as the cost of the shortest path (geodesic) between the pair of nodes in analysis added by the spreadness factor ρ. Any cost metric in the real domain, excluding zero, can be used. Consequently, ρ will be limited to the same domain, but we need to include zero because we also use shortest paths, which are found when $\rho = 0$. In the third analysis, given that each centrality metric provides a different node ranking, we investigated the number of critical nodes elected as the most central and how the throughput of the network is affected by a single failure on one of the most central nodes according to each metric.

We analyzed both dynamic and static networks, using publicly available datasets and randomly generated scale-free networks. Dynamic networks are characterized for the frequent topology changes that can happen due to node mobility, causing intermittent connectivity. We showed that, based on the results of our analyses, we can potentially reduce resource consumption, without decreasing the average packet delivery ratio in computer networks, using simple forwarding strategies. Moreover, if we use slightly longer paths to compute node importance, we are able to reduce the number of critical nodes among the most central nodes in the network. Hence,

we are moving in the right direction on the search for better network resilience. The investigation performed in this work is fundamental to provide better insights on the behavior of the network structure so that we can maximize the exploitation of contact opportunities and have more information about nodes roles to make better decisions.

The proposed extended vicinity uses the relative speed of nodes to determine whether a link between two nodes should exist. This definition allows identifying conditions for multihop communications, discarding contacts that are not useful because they do not last long enough to successfully finish the transmission. We analyzed the behavior of the proposed (κ, λ)-vicinity using three vehicular network datasets with distinct node distribution across the covered area. Nodes in each dataset represent a different type of vehicle: taxis (Taxi scenario), buses (Bus scenario) and personal cars (Synthetic scenario). The Synthetic scenario is, in fact, a traffic model based on real observations. The majority of relative speeds in the Bus and Taxi scenarios are lower than 30 km/h, but in the Synthetic scenario we observed several relative speeds reaching over 100 km/h. In fact, in this scenario, the majority of relative speeds are lower than 140 km/h.

We quantitatively confirmed the intuition behind "contacts happen more often at few hops and low relative speeds" for three scenarios. The relation between number of hops, contact duration and relative speeds of vehicles has never been quantified, as far as we know. Our results showed that most contacts often happen at few hops and low relative speeds, as expected. Nevertheless, we also demonstrated that useful contacts can happen between nodes at higher relative speeds, separated by longer hop distances, even though less often. In such conditions, some nodes are still capable of transferring MB-sized messages. We also observed that nodes spend a lot of time out of contact, in average, and the longest average contact duration is found for 1-hop contacts. In the Bus scenario, whenever nodes are in contact, in whichever state, there is a high probability that they remain in this state. When the hop distance increases, the probability of moving to State ∞ also increases, e.g., at relative speeds lower than 40 km/h and 1-hop distance, the probability of going to State ∞ is approximately 1.1%; at 7 hops, it increases to 50%. At higher relative speeds, the probability increases to 11.6% at 1-hop distance, and to 66.7% at 3 hops, which is the maximum hop distance in this case. We also observed that the duration of direct contacts at low relative speeds in all scenarios can be very short, short or long, when the physical distance between nodes is longer than 120 m; and they are either short or long when the distance is shorter. In turn, contacts between nodes at high relative speeds are always very short, independently of the distance between the nodes.

In the Bus and Taxi scenarios, the majority of contacts happen for hop distances

shorter than 3 hops and at relative speeds lower than 40 km/h (Bus) and 60 km/h (Taxis). In the Synthetic scenario the hop distance reaches up to 6 hops and we can find a very significant number of multihop contacts at very high relative speeds. We attribute this to the higher density of the scenario compared to the other two. The total duration of contacts is also higher in the Synthetic scenario. Nevertheless, even with higher density and longer total duration of all contacts, the number of individual useful contacts in this scenario is small. The longest average time spent in contact in this scenario is 2.20 s at 1-hop distance considering a radio range of 150 m. Even in sparser scenarios, such as the Bus scenario, the average time in each state is higher. The number of useful contacts in all scenarios is highly influenced by both hop distance and relative speed. For security applications, which use very small packets, all contacts in our scenarios can be used, because they last at least 1 s. If we use MB-sized bundles, the number of useful contacts drops significantly. In addition, increasing the relative speed sometimes reduces more severely the number of useful contacts when compared to the hop distance.

The aforementioned results showed the importance of also considering the relative speeds for path establishment when developing an application for vehicular networks to avoid unnecessary transmissions, especially when transmitting bigger messages. The same idea could also be applied to adjust the number of hops a message could be forwarded. This would be a function of the relative speed between nodes in contact. To demonstrate the impact of our results on mobile communications, we proposed three forwarding strategies based on the outcomes of this work, `RelSpeedR`, `VicR`, and `RelSpeedVicR` and we ran simulations using each one. Although the `OLSR` is not the most suitable routing protocol for VANETs, all proposed forwarding strategies are based on it due to our need to have information about the global network structure, in the current state of the work. To provide information about nodes' relative speed, we used a reserved field of the `OLSR` Hello packet, which is not currently used by the implementations of this protocol. Hence, we do not add any overhead to the communication.

We increased the restrictions imposed to each forwarding strategy, starting with `RelSpeedR`, where nodes can only forward a packet to the next hop if the neighbor is within a relative speed range that still allows reaching the destination at that hop distance. In `VicR`, we restrict the neighborhood of a node, according to the relative speed between the node and its adjacent neighbors. `RelSpeedVicR` uses both restrictions. We showed that both `RelSpeedVicR` and `VicR` can potentially reduce the waste of network resources without decreasing the average packet delivery ratio, with `VicR` presenting the best performance considering this aspect. Our results showed it is important to consider the relative speed for path establishment when developing an application for vehicular networks. Nodes moving at high rela-

tive speed must not waste time sending unnecessary control messages because each contact opportunity is frequently short. Hence, the outcomes of this work could be used to improve existing forwarding schemes.

In this work, we also proposed a different strategy to assess node importance. We claim that, in some cases, using only the shortest paths to compute node importance is not sufficient. For instance, the spread of rumors in a social network does not follow only shortest paths. In a computer network, we could be interested in splitting the load on some important node on the shortest path, even if we need to use slightly longer paths. As such we could prevent failure of the node on the shortest path due to overload. In the context of smart cities, we could avoid node overload by placing different functions in different locations, even if they perform better when placed, separately, on the same location. For instance, we could place routing functions on important nodes of the shortest paths, while data concentrators could be placed on nodes around the shortest paths, i.e., nodes in *quasi*-shortest paths. To this end, we propose the ρ-geodesic betweenness centrality, which is a variant of the traditional betweenness that uses both shortest and *quasi*-shortest paths to assess node importance. The idea is to increase the importance of nodes that do not necessarily fall on shortest paths, but can still be considered critical to the network operation. We characterized the ρ-geodesic betweenness using several randomly generated networks that follow a power-law degree distribution, and four datasets with distinct characteristics, for which we also computed the traditional and distance-scaled betweenness. We additionally computed the random walk betweenness using Newman's algorithm, when possible. The random walk betweenness follows the same idea of using more paths in addition to the shortest ones. It considers, however, that information travels at random using all existing paths. This is not the case in some situations, such as in the majority of computer and transport networks, and even in some social networks. In addition, the complexity of this metric is higher than the one of our metric. Moreover, although similar in concept, the ρ-geodesic betweenness is quite different from the random walk betweenness in practice, especially for networks that follow a power-law degree distribution with $2 \leq \alpha \leq 3$.

Our characterization showed that the ρ-geodesic betweenness is able to rerank nodes, promoting those that participate in many paths. Even though, it remains strongly correlated to the traditional betweenness, meaning that they are still measuring the same characteristics of the node. In addition, the level of agreement between the rankings obtained by each metric is also high, usually over 95% if we compare the metrics in pairs. Although the concordance between the metrics is very high, the rate with which nodes are reranked is also high, i.e., several nodes change positions compared to the traditional betweenness. This happens because, contrary

to the coefficient of concordance (Kendall's W), the reranking rate does not account whether the change in the rank position is significant. The ρ-geodesic betweenness spreads the classification rank, giving room to break ties between nodes, as the number of *quasi*-shortest paths they fall on can be very different. The random walk betweenness also presents these characteristics, but depending on the dataset, it can increase the number of nodes tied in the same position. We claim that having several nodes tied in the same position may waste their potential to contribute to the operation of the network.

Node reranking according to the ρ-geodesic betweenness also presents the effect of reducing the number of articulation points among the most central nodes. Such nodes are critical to the network operation because if they fail they disconnect areas of the network, i.e., they have the potential to split the network into several connected components. We found that the probability that a failed nodes is also an articulation point, considering the top #5 nodes according to each metric, is never higher in the ρ-geodesic betweenness when compared to the distance-scaled betweenness and it is always lower compared to the traditional betweenness. This happens because nodes in the first position are frequently the same, but in the following positions they are usually different. For instance, during 83% of the time, nodes classified on the first position are the same for both the traditional and distance scaled betweenness, but comparing the former with the proposed metric this frequency drops to 67%. We ran separate simulations where we failed a single node from the top #5 positions from each rank of each metric, to investigate the impact on the average throughput of the network, i.e., the impact on the rate of successful message delivered over the communication medium. We observed that a single failure is disastrous to the average network throughput. Failures on important nodes according to the proposed metric are usually equal to the other metrics and, sometimes, slightly less impacting. We believe that this happens due to the frequency of coincidence between the nodes elected as the most important by each metric. When the coincidence is smaller than 50% of the time, the impact on the throughput is slightly lower.

We also observed that the ρ-geodesic betweenness has the potential to reduce the number of resources reallocations, in networks that use shortest path based rules to distribute resources. Such resources can range from information flow to real or virtual machines, for example. Some of these networks can quickly change their topology and in the vehicular network scenario that we analyzed, we found that the ρ-geodesic betweenness is able to provide longest rank stability to a larger number of nodes compared to the other metrics. Hence, if we consider a scenario where central nodes are in charge of a network function, we could decrease the operational costs due to the reduction on the number of migrations because the central node

will continue to be central for longer periods.

The vicinity analysis performed in this work is important to understand the dynamics of the network. The centrality analysis, in turn, is important to understand the roles played by nodes within the network. With both analyses, we can have more knowledge to make better decisions, improving the performance of the network. As future work we plan to extend our analyses to other real datasets using data mining tools. The idea is to use these tools due to the huge amount of data produced, which is difficult to evaluate with ordinary analysis tools. We have already stepped towards a new routing protocol when we proposed the three forwarding strategies based on the outcomes of the vicinity analysis. We intend to continue on this path and fully design such protocol. To this end, we need to consider other routing protocols, besides the `OLSR`, that are more suitable for VANETs, but that can provide information about the topology of the region of interest. In addition, we need to extend the proposed strategies to also include the relative speeds of the complete path between the communicating nodes. The performance of the proposed strategies has to be further evaluated considering that nodes can store messages for a short period before dropping it, which could increase the rate of successfully delivered messages.

Lastly, we will investigate the performance of the network running under rules based on the ρ-geodesic betweenness, studying its relevance in different use cases, including weighted networks. To this end, we first need to optimize the algorithm to compute the metric, so that we can apply it to larger networks. Then we will choose different scenarios where the flow behavior can benefit from the use of the proposed metric to develop specific applications, such as load balancing or increase the communication throughput. The proposed metric can also be adapted to improve the scenario model. For instance, the spreadness ρ can be dynamic, changing according to the current status of the network, or it could include not only the cost of the paths, but also the bandwidth of each path, being in closer agreement with the communication capacity. We also intend to propose a mechanism to find the optimum value for ρ according to the characteristics of the scenario.

Bibliography

[1] BARABÁSI, A.-L. "The Network Takeover", *Nature Physics*, v. 8, n. 1, pp. 14–16, 2012.

[2] BOCCALETTI, S., LATORA, V., MORENO, Y., et al. "Complex Networks: Structure and Dynamics", *Physics Reports*, v. 424, n. 4, pp. 175–308, 2006.

[3] BELBLIDIA, N., SAMMARCO, M., COSTA, L. H. M. K., et al. "EPICS: Fair Opportunistic Multi-Content Dissemination", *Trans. on Mobile Computing*, v. 14, n. 9, pp. 1847–1860, 2015.

[4] CAMPISTA, M. E. M., ESPOSITO, P. M., MORAES, I. M., et al. "Routing Metrics and Protocols for Wireless Mesh Networks", *IEEE Network*, v. 22, n. 1, pp. 6–12, 2008.

[5] REZENDE, C. G., PAZZI, R. W., BOUKERCHE, A. "An Efficient Neighborhood Prediction Protocol to Estimate Link Availability in VANETs". In: *MobiWAC '09*, pp. 83–90, 2009.

[6] TALEB, T., SAKHAEE, E., JAMALIPOUR, A., et al. "A Stable Routing Protocol to Support ITS Services in VANET Networks", *Trans. on Vehicular Technology*, v. 56, n. 6, pp. 3337–3347, 2007.

[7] KISS, C., BICHLER, M. "Identification of Influencers - Measuring Influence in Customer Networks", *Decision Support Systems*, v. 46, n. 1, pp. 233–253, 2008.

[8] THILAKARATHNA, K., VIANA, A. C., SENEVIRATNE, A., et al. "Mobile Social Networking Through Friend-to-friend Opportunistic Content Dissemination". In: *ACM MobiHoc*, pp. 263–266, 2013.

[9] BOUET, M., LEGUAY, J., COMBE, T., et al. "Cost-Based Placement of vDPI Functions in NFV Infrastructures", *International Journal of Network Management*, v. 25, n. 6, pp. 490–506, 2015.

[10] DALY, E. M., HAAHR, M. "Social Network Analysis for Routing in Disconnected Delay-tolerant MANETs". In: *ACM MobiHoc*, pp. 32–40, 2007.

[11] ARULSELVAN, A., COMMANDER, C. W., ELEFTERIADOU, L., et al. "Detecting Critical Nodes in Sparse Graphs", *Computers & Operations Research*, v. 36, n. 7, pp. 2193–2200, 2009.

[12] BORGATTI, S. P., MEHRA, A., BRASS, D. J., et al. "Network Analysis in the Social Sciences", *Science*, v. 323, n. 5916, pp. 892–895, 2009.

[13] PANTAZOPOULOS, P., KARALIOPOULOS, M., STAVRAKAKIS, I. "On the Local Approximations of Node Centrality in Internet Router-Level Topologies". In: *IWSOS*, pp. 115–126, 2013.

[14] GHAHRAMANI, S. A. A. G., HEMMATYAR, A. M. A., KAVOUSI, K. "A Network Model for Vehicular Ad Hoc Networks: An Introduction to Obligatory Attachment Rule", *Trans. on Network Science and Engineering*, v. 3, n. 2, pp. 82–94, 2016.

[15] FREEMAN, L. C. "Centrality in Social Networks: Conceptual Clarification", *Social Networks*, v. 1, n. 3, pp. 215–239, 1978.

[16] HUI, P., CROWCROFT, J., YONEKI, E. "Bubble Rap: Social-based Forwarding in Delay Tolerant Networks". In: *ACM Mobihoc*, pp. 241–250, 2008.

[17] WEHMUTH, K., ZIVIANI, A. "DACCER: Distributed Assessment of the Closeness Centrality Ranking in Complex Networks", *Computer Networks*, v. 57, n. 13, pp. 2536–2548, 2013.

[18] SOMMER, C., DRESSLER, F. "Progressing Toward Realistic Mobility Models in VANET Simulations", *Communications Magazine*, v. 46, n. 11, pp. 132–137, 2008.

[19] HARRI, J., FILALI, F., BONNET, C. "Mobility Models for Vehicular Ad Hoc Networks: A Survey and Taxonomy", *Communications Surveys & Tutorials*, v. 11, n. 4, pp. 19–41, 2009.

[20] GAIKWAD, D. S., ZAVERI, M. "VANET Routing Protocols and Mobility Models: A Survey". In: *Trends in Network and Communications*, pp. 334–342, 2011.

[21] SPAHO, E., BAROLLI, L., MINO, G., et al. "VANET Simulators: A Survey on Mobility and Routing Protocols". In: *BWCCA '11*, pp. 1–10, 2011.

[22] ZEADALLY, S., HUNT, R., CHEN, Y.-S., et al. "Vehicular Ad Hoc Networks (VANETS): Status, Results, and Challenges", *Telecomunication Systems*, v. 50, n. 4, pp. 217–241, 2012.

[23] MADI, S., AL-QAMZI, H. "A Survey on Realistic Mobility Models for Vehicular Ad Hoc Networks (VANETs)". In: *ICNSC*, pp. 333–339, 2013.

[24] GERHARZ, M., DE WAAL, C., FRANK, M., et al. "Link stability in mobile wireless ad hoc networks". In: *LCN*, pp. 30–39, 2002.

[25] MENOUAR, H., LENARDI, M., FILALI, F. "Movement Prediction-Based Routing (MOPR) Concept for Position-Based Routing in Vehicular Networks". In: *VTC '07*, pp. 2101–2105, 2007.

[26] BARGHI, S., BENSLIMANE, A., ASSI, C. "A Lifetime-Based Routing Protocol for Connecting VANETs to the Internet". In: *WoWMoM*, pp. 1–9, 2009.

[27] CONAN, V., LEGUAY, J., FRIEDMAN, T. "Characterizing Pairwise Inter-contact Patterns in Delay Tolerant Networks". In: *Autonomics*, pp. 19:1–19:9, 2007.

[28] GONZALEZ, M. C., HIDALGO, C. A., BARABASI, A.-L. "Understanding Individual Human Mobility Patterns", *Nature*, v. 453, n. 7196, pp. 779–782, 2008.

[29] PASSARELLA, A., CONTI, M. "Characterising Aggregate Inter-contact Times in Heterogeneous Opportunistic Networks". In: *NETWORKING*, pp. 301–313, 2011.

[30] MENEGUETTE, R. I., FILHO, G. P. R., BITTENCOURT, L. F., et al. "Enhancing Intelligence in Inter-vehicle Communications to Detect and Reduce Congestion in Urban Centers". In: *ISCC*, pp. 662–667, 2015.

[31] PHE-NEAU, T., DIAS DE AMORIM, M., CONAN, V. "Vicinity-based DTN Characterization". In: *MobiOpp*, pp. 37–44, 2012.

[32] PHE-NEAU, T., DIAS DE AMORIM, M., CAMPISTA, M. E. M., et al. "Examining Vicinity Dynamics in Opportunistic Networks". In: *PM2HW2N*, pp. 153–160, 2013.

[33] DOLEV, S., ELOVICI, Y., PUZIS, R. "Routing Betweenness Centrality", *Journal of the ACM*, v. 57, n. 4, pp. 25:1–25:27, 2010.

[34] GILES, A. P., GEORGIOU, O., DETTMANN, C. P. "Betweenness Centrality in Dense Random Geometric Networks". In: *ICC*, pp. 6450–6455, 2015.

[35] YIM, J., AHN, H., KO, Y.-B. "The Betweenness Centrality Based Geographic Routing Protocol for Unmanned Ground Systems". In: *IMCOM*, pp. 74:1–74:4, 2016.

[36] MAGAIA, N., FRANCISCO, A. P., PEREIRA, P., et al. "Betweenness Centrality in Delay Tolerant Networks: A survey", *Ad Hoc Networks*, v. 33, pp. 284–305, 2015.

[37] JAIN, A. "Betweenness Centrality Based Connectivity Aware Routing Algorithm for Prolonging Network Lifetime in Wireless Sensor Networks", *Wireless Networks*, v. 22, n. 5, pp. 1605–1624, 2016.

[38] FREEMAN, L. C., BORGATTI, S. P., WHITE, D. R. "Centrality in Valued Graphs: A Measure of Betweenness Based on Network Flow", *Social Networks*, v. 13, n. 2, pp. 141–154, 1991.

[39] BRANDES, U., FLEISCHER, D. "Centrality Measures Based on Current Flow". In: *STACS*, pp. 533–544, 2005.

[40] NEWMAN, M. J. "A Measure of Betweenness Centrality Based on Random Walks", *Social Networks*, v. 27, n. 1, pp. 39–54, 2005.

[41] BORGATTI, S. P., EVERETT, M. G. "A Graph-Theoretic Perspective on Centrality", *Social Networks*, v. 28, n. 4, pp. 466–484, 2006.

[42] OPSAHL, T., AGNEESSENS, F., SKVORETZ, J. "Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths", *Social Networks*, v. 32, n. 3, pp. 245–251, 2010.

[43] FREEMAN, L. C. "A Set of Measures of Centrality Based on Betweenness", *Sociometry*, v. 40, n. 1, pp. 35–41, 1977.

[44] CLAUSEN, T., JACQUET, P. *Optimized Link State Routing Protocol (OLSR)*. RFC 3626, 2003.

[45] MEDEIROS, D. S. V., HERNANDEZ, D. A. B., CAMPISTA, M. E. M., et al. *Impact of Relative Speed on Node Vicinity Dynamics in VANETs*. Technical Report 18, Universidade Federal do Rio de Janeiro, 2017.

[46] TRIVEDI, K. S. *Probability and Statistics with Reliability, Queuing and Computer Science Applications*. John Wiley and Sons Ltd., 2002. ISBN: 0-471-33341-7.

[47] PIORKOWSKI, M., SARAFIJANOVIC-DJUKIC, N., GROSSGLAUSER, M. "CRAWDAD data set epfl/mobility (v. 2009-02-24)". [Online]. Available ago/2017 in http://crawdad.org/epfl/mobility/, fev. 2009.

[48] JETCHEVA, J. G., HU, Y.-C., PALCHAUDHURI, S., et al. "CRAWDAD data set rice/ad_hoc_city (v. 2003-09-11)". [Online]. Available ago/2017 in http://crawdad.org/rice/ad_hoc_city/, set. 2003.

[49] UPPOOR, S., FIORE, M. "Large-scale urban vehicular mobility for networking research". In: *VNC*, pp. 62–69, 2011.

[50] FREEMAN, S., FREEMAN, L. *The Networkers Network: A Study of the Impact of a New Communications Medium on Sociometric Structure*. Social Sciences Research Reports. School of Social Sciences, University of California, 1979.

[51] LUSSEAU, D., SCHNEIDER, K., BOISSEAU, O. J., et al. "The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations", *Behavioral Ecology and Sociobiology*, v. 54, n. 4, pp. 396–405, 2003.

[52] JOHNSON, D. S. "The Genealogy of Theoretical Computer Science", *SIGACT News*, v. 16, n. 2, pp. 36–44, 1984.

[53] BARABÁSI, A.-L., BONABEAU, E. "Scale-Free Networks", *Scientific American*, v. 288, n. 5, pp. 50–59, 2003.

[54] BARABÁSI, A.-L., ALBERT, R. "Emergence of Scaling in Random Networks", *Science*, v. 286, n. 5439, pp. 509–512, 1999.

[55] HOLME, P., KIM, B. J. "Growing Scale-Free Networks with Tunable Clustering", *Physical Review E*, v. 65, n. 2, pp. 026107–1–026107–4, 2002.

[56] HAVEL, V. "A Remark on the Existence of Finite Graphs", *Časopis pro pěstování matematiky*, v. 80, n. 4, pp. 477–480, 1955. (*in Czech*).

[57] HAKIMI, S. L. "On Realizability of a Set of Integers as Degrees of the Vertices of a Linear Graph. I", *Journal of the Society for Industrial and Applied Mathematics*, v. 10, n. 3, pp. 496–506, 1962.

[58] DOROGOVTSEV, S. N., MENDES, J. F. F. "Evolution of Networks", *Advances in Physics*, v. 51, n. 4, pp. 1079–1187, 2002.

[59] OLARIU, S., YAN, G., SALLEH, S. "A Probabilistic Routing Protocol in VANET", *International Journal of Mobile Computing and Multimedia Communications*, v. 2, n. 4, pp. 21–37, 2010.

[60] BAZZI, A., MASINI, B. M., ZANELLA, A., et al. "IEEE 802.11p for Cellular Offloading in Vehicular Sensor Networks", *Computer Communications*, v. 60, pp. 97–108, 2015.

[61] SHELLY, S., BABU, A. V. "Link Residual Lifetime-Based Next Hop Selection Scheme for Vehicular Ad Hoc Networks", *EURASIP Journal on Wireless Communications and Networking*, v. 2017, n. 1, pp. 23:1–23:13, 2017.

[62] SPYROPOULOS, T., RAIS, R. N. B., TURLETTI, T., et al. "Routing for Disruption Tolerant Networks: Taxonomy and Design", *Wireless Networks*, v. 16, n. 8, pp. 2349–2370, 2010.

[63] HOQUE, M. A., HONG, X., DIXON, B. "Efficient Multi-Hop Connectivity Analysis in Urban Vehicular Networks", *Vehicular Communications*, v. 1, n. 2, pp. 78–90, 2014.

[64] HERNANDEZ, D. A. B., MEDEIROS, D. S. V., CAMPISTA, M. E. M., et al. "Uma Avaliação da Influência da Velocidade dos Nós no Estabelecimento de Caminhos em Redes Ad Hoc Veiculares". In: *SBRC*, pp. 431–444, 2015.

[65] WANG, X., WANG, C., CUI, G., et al. "Practical Link Duration Prediction Model in Vehicular Ad Hoc Networks", *International Journal of Distributed Sensor Networks*, v. 11, n. 3, pp. 216934:1–216934:14, 2015.

[66] HARTENSTEIN, H., LABERTEAUX, K., EBRARY, I. *VANET: Vehicular Applications and Inter-Networking Technologies*. Wiley Online Library, 2010.

[67] FATHIAN, M., JAFARIAN-MOGHADDAM, A. R. "New Clustering Algorithms for Vehicular Ad-Hoc Network in a Highway Communication Environment", *Wireless Networks*, v. 21, n. 8, pp. 2765–2780, 2015.

[68] HE, J., CAI, L., PAN, J., et al. "Delay Analysis and Routing for Two-Dimensional VANETs Using Carry-and-Forward Mechanism", *Trans. on Mobile Computing*, v. 16, n. 7, pp. 1830–1841, 2017.

[69] CHENG, L., HENTY, B. E., STANCIL, D. D., et al. "Mobile Vehicle-to-Vehicle Narrow-Band Channel Measurement and Characterization of the 5.9 GHz

Dedicated Short Range Communication (DSRC) Frequency Band", *Journal on Selected Areas in Communications*, v. 25, n. 8, pp. 1501–1516, 2007.

[70] YIN, J., HOLLAND, G., ELBATT, T., et al. "DSRC Channel Fading Analysis from Empirical Measurement". In: *ChinaCom '06*, pp. 1–5, 2006.

[71] BAVELAS, A. "A Mathematical Model for Group Structures", *Human Organization*, v. 7, n. 3, pp. 16–30, 1948.

[72] SHIMBEL, A. "Structural Parameters of Communication Networks", *Bulletin of Mathematical Biophysics*, v. 15, n. 4, pp. 501–507, 1953.

[73] SHAW, M. E. "Group Structure and the Behavior of Individuals in Small Groups", *The Journal of Psychology*, v. 38, n. 1, pp. 139–149, 1954.

[74] COHN, B. S., MARRIOTT, M. "Networks and Centres of Integration in Indian Civilization", *Journal of Social Research*, v. 1, pp. 1–9, 1958.

[75] STEPHENSON, K., ZELEN, M. "Rethinking Centrality: Methods and Examples", *Social Networks*, v. 11, n. 1, pp. 1–37, 1989.

[76] SAVLA, K., COMO, G., DAHLEH, M. A. "Robust Network Routing under Cascading Failures", *Trans. on Network Science and Engineering*, v. 1, n. 1, pp. 53–66, 2014.

[77] DUPONT, C., SCHULZE, T., GIULIANI, G., et al. "An energy aware framework for virtual machine placement in cloud federated data centres". In: *e-Energy*, pp. 1–10, 2012.

[78] GEISBERGER, R., SANDERS, P., SCHULTES, D. "Better Approximation of Betweenness Centrality". In: *ALENEX*, pp. 90–100, 2008.

[79] BRANDES, U. "On Variants of Shortest-Path Betweenness Centrality and Their Generic Computation", *Social Networks*, v. 30, n. 2, pp. 136–145, 2008.

[80] AUSIELLO, G., FIRMANI, D., LAURA, L. "The (Betweenness) Centrality of Critical Nodes and Network Cores". In: *IWCMC*, pp. 90–95, 2013.

[81] ZHANG, D., STERBENZ, J. P. G. "Modelling Critical Node Attacks in MANETs", *LNCS Series*, v. 8221, pp. 127–138, 2014.

[82] COUTO, R. D. S., SECCI, S., CAMPISTA, M. E. M., et al. "Reliability and Survivability Analysis of Data Center Network Topologies", *Journal of Network and Systems Management*, v. 24, n. 2, pp. 346–392, 2016.

[83] KERMARREC, A.-M., LE MERRER, E., SERICOLA, B., et al. "Second Order Centrality: Distributed Assessment of Nodes Criticity in Complex Networks", *Computer Communications*, v. 34, n. 5, pp. 619–628, 2011.

[84] LULLI, A., RICCI, L., CARLINI, E., et al. "Distributed Current Flow Betweenness Centrality". In: *SASO*, pp. 71–80, 2015.

[85] BADER, D. A., MADDURI, K. "Parallel Algorithms for Evaluating Centrality Indices in Real-world Networks". In: *ICPP*, pp. 539–550, 2006.

[86] BADER, D. A., MADDURI, K. "Designing Multithreaded Algorithms for Breadth-First Search and St-connectivity on the Cray MTA-2". In: *ICPP*, pp. 523–530, 2006.

[87] SARIYÜCE, A. E., KAYA, K., SAULE, E., et al. "Betweenness Centrality on GPUs and Heterogeneous Architectures". In: *GPGPU*, pp. 76–85, 2013.

[88] BERNASCHI, M., CARBONE, G., VELLA, F. "Scalable Betweenness Centrality on multi-GPU Systems". In: *CF*, pp. 29–36, 2016.

[89] LEHMANN, K., KAUFMANN, M. *Decentralized Algorithms for Evaluating Centrality in Complex Networks*. Technical Report 10, Wilhelm Schickard Institut, 2003.