Università di Pisa

Department of Computer Science

Master Degree - Data Sciences & Business Informatics

Route diversity in urban environments

Candidate

Charlotte Brimont

Supervisors

Luca Pappalardo

Giuliano Cornacchia

A.Y. 2022-2023

ABSTRACT

Facing the climate urgency and the impact of global warming, limiting our emissions has become one of the main challenges of our societies. The impact of transportation, especially of motorised vehicles, on carbon emissions and air quality is not to prove anymore. The decreasing of our impact on emissions can only be achieve with a better understanding of human mobility in general, especially individual behaviors on road networks and recurring patterns of our societies and by deepening our comprehension of the routing algorithms and their impact on road usage [29]. The alternative routing strategies aim to diversify the routes taken by drivers, and, consequently, to reduce the potential congestion and the associated emissions. However, to build more efficient alternative routing algorithms, studying the route diversity in cities is necessary to understand how to adapt them to each road network. In our work, we develop a procedure to assess the existing diversification of paths in urban environments and the factors impacting it. In particular, we studied the impact of distance on the set of paths between an origin and a destination points. We also evaluated how the topology of a city changes the potential paths diversification in an urban environment.

Table of Contents

Chapter 1

Introduction

In 1988, the United Nations Environment Programme (UNEP) and the World Meteorological Organisation (WMO) established a panel regrouping scientists from all around the world to study the impact of humans on climate change. The Intergovernmental Panel on Climate Change (IPCC) aims to gather, summarise and provide a critical analysis of scientific literature on this field of study. At the end of each evaluation cycle of 6-7 years, the major discoveries are compiled in the form of a report addressed to governments.

The latest report regroups the knowledge acquired between 2015 and 2021. It evaluates the global warming to 1.5°C by 2030, and this, regardless the emissions produced from now on. Yet, the limitation of this warming to 1.5°C-2°C could only be achieved with a drastic decrease of greenhouse gases and carbon dioxide emissions. According to the International Energy Agency (IEA), the transportation sector represents around 20% of the CO_2 emissions[1]. In particular, almost 50% of these emissions are caused by passenger road vehicles[2], that is to say, cars, motorbikes and buses. Therefore, reducing the emissions of passenger vehicles is a *sine qua non* condition in the limitation of global temperature rising.

In the past few years, many urban limitations were established in the major European cities[3] among them we can note restricted traffic area, that is to say zones in which only residents are allowed to drive. In France, the government established another regulation policy a few years ago. It consists of the classification of all motorised vehicles according to its motor, the fuel it uses and the emissions. This classification is created alongside the progressive prohibition of the most polluting ones in major french cities, especially in Paris. The final target of this policy is to reach 100% of electric vehicles in restricted areas by 2030. This regulation's purpose is not only to reduce the carbon emissions [4] but it also is a national health matter, especially, through air quality improvement [31].

Motorised vehicles are a major source of pollution. By rejecting carbon dioxide and particulate matter, they participate in the global temperature rising and create major health issues for people exposed to these emissions. Therefore, controlling car's

[1]Global CO_2 emissions by sector, 2019-2022

[2]Transport sector CO_2 emissions by mode in the Sustainable Development Scenario, 2000-2030

[3]`https://urbanaccessregulations.eu/`

emissions, especially in urban environments, is essential to limit the global warming and protect human health [5, 6]. Through years, many scientists studied the relation between traffic and emissions in urban environments. In particular, many research were led on the impact of congestion on the emissions of carbon dioxide, greenhouse gases and particulate matter [3, 17, 43]. The generalisation of the use of navigation apps in our everyday trips also increased the emissions, especially in the city centres [11, 30]. Thus, mitigating the congestion and studying the potential diversification of paths is crucial for ecological and for public health concerns.

The first causes of congestion are physical obstacles or modification of the road network, for instance, a car accident causing lanes' closure on the road or the presence of work zones at some points of the itinerary. However, congestion can also be caused by human factors. Indeed, road users tend to be very selfish when it comes to routing choices. If an individual needs to take a car from an origin to a destination, they would most likely take the shortest or fastest path without considering the rest of the users who shares the road. The selfish behavior of individuals might be considered as a gain of time for drivers. Yet, it only decreases the global performances of road networks causing unnecessary congestion [14, 37]. Indeed, the routing choices of road users may cause congestion in urban areas [9, 10, 12].

However, during the peak hours or for long trajectories, users tend to diversify their routes more than they usually do, by taking alternative routes to decrease their total travel time [40], thereby, paradoxically reducing their own impact on congestion. Thus, it is reasonable to assume that using alternative routing (AR) alleviate congestion by better distributing vehicles on the road network. Alternative routing consists of diversifying the shortest or fastest paths by generating other routes. The main idea behind this concept is to reduce the load of drivers on the most used road segments by offering different possible paths to different drivers. Finally, to understand to what extend it is possible to reduce traffic load on these congested road segments, it is crucial to study how much a city can allow paths diversification.

The main objective of this thesis is to asses the diversity of routes in urban environments. We develop a paths enumeration algorithm, based on alternative routing, capable of generating sets of thousands different paths in a reasonable time. The sets are used to compare the potential diversification existing between Milan, Rome and Florence and to highlight structural parameters influencing, positively or negatively, the number of paths. With this procedure, we analyse the impact of the distance between the origin and the destination, the length of the paths comparing to the shortest one and the influence of the core-periphery structure in this diversification. The purpose of our research is to understand how the road network's layout impact the number of possible paths between two locations.

This thesis is organised in different chapters. Chapter 2 contains a presentation of essential notions from graph theory and human mobility and a state of the art regarding paths from both perspective and the impact of graph's topology in our every day trips. Chapter 3 details our procedure to assess the diversity in a city, the road networks, the chosen sampling methods and, more importantly, the paths enumeration algorithm developed to analyse potential diversification. Chapter 4 compares the results obtained for each one of the studied cities: Rome, Milan and Florence, and explains the different

factors impacting the diversity in these cities. Finally, in Chapter 5, we discuss the limitations of our analysis and presents possible improvements and future related works.

1.1 Open Science

To guarantee the reproducibility of our work and the results obtained, the code is available on a dedicated github project: `https://github.com/charlottebrimont/route_diversity_in_urban_environements`.

Chapter 2

State of the art and background

This chapter presents some major concepts of our fields of study and reviews the existing scientific literature on these topics. In Section 2.1, we will introduce some essential notions of human mobility and graph theory. Then, Section 2.2 summarises some scientific works about paths problems (Subsection 2.2.1) and about the topology of road graph and how it is influencing human mobility (Subsection 2.2.2).

2.1 Background

In this Section, we will present the essential notion to correctly apprehend the rest of our work. The Subsection 2.1.1 defines a fundamental concept of human mobility to represent and store human flows, the origin-destination (OD) matrix. Then, Subsection 2.1.2 introduces the detour index, a major topological measure of human mobility. Subsection 2.1.3 presents the K_{road}, an assessment of the attractiveness of road segments. In Subsection 2.1.4, we will introduce an important algorithm of sampling used to study the topology of cities. Finally, in Subsection 2.1.5, we will present the Dijkstra algorithm and some alternative routing strategies based on it.

2.1.1 OD matrix

The human mobility field covers a large spectrum of problems from everyday trips to migratory flows. Most of this questions require large sets of data to analyse human behavior at every level. Among the different type of data available to study the movement of humans, one of the most important is the *OD* matrix.

An *OD* matrix, namely origin-destination matrix, represents human flows. Each cell m_{ij} of the matrix represents the flow from a location i to another j one and the cells on the diagonal, m_{ii} represents people staying within the same area. Usually, these matrices are created based on geographical divisions, such as countries to study migratory flows or administrative divisions for everyday trips. However, *OD* matrices can also be created based on a tessellation, that is to say an artificial territorial cutting such as the Voronoi's tessellation or an hexagonal one.

The advantage of an *OD* matrix is that it gathers data representing human flows

at different levels. For instance, an *OD* matrix can represent either migratory flows between every countries, but also every day commuting between the neighborhoods of a metropolitan area.

2.1.2 Detour Index

When studying the topology of graphs, it can be interesting to use path-based measures. Amongst other, we can find the detour index ω [2, 26, 44] and the Wiener index W [18]. Both of these measures are topological indices used, especially in chemical graph theory, to describe and compare layouts of graphs, such as molecules. The Wiener index has been formalised by Harry Wiener in 1947 [39]. First named *path number*, it is computed by adding up the lengths of the shortest paths between each pair of nodes of a molecule. This measure was then generalised for all networks and renamed Wiener index. Similarly, years later, the detour index ω was introduced as the sum of the lengths of the longest paths between each nodes.

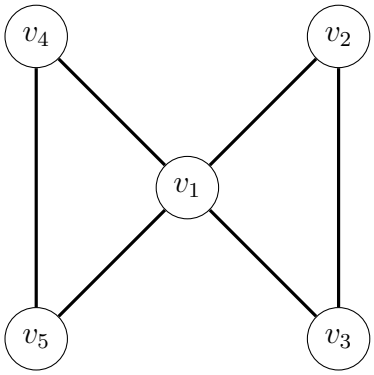


Figure 2.1: Undirected and connected graph $G = (N, E)$.

To compute both, the Wiener and the detour indices of a graph, it is mandatory to build the distance matrix d and the detour matrix Δ. The distance matrix d represents the length of all the shortest paths between two nodes, whereas the detour matrix Δ gathers the length of the longest paths between each pair of vertices. For instance, for the graph G described by Figure 2.1, the distance matrix is given by the Matrix 2.1 and the detour matrix is represented by the Matrix 2.2.

$$D = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 & 2 \\ 1 & 1 & 0 & 2 & 2 \\ 1 & 2 & 2 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 \end{bmatrix} \quad (2.1) \qquad \Delta = \begin{bmatrix} 0 & 2 & 2 & 4 & 4 \\ 2 & 0 & 2 & 4 & 4 \\ 2 & 2 & 0 & 2 & 2 \\ 4 & 4 & 2 & 0 & 2 \\ 4 & 4 & 2 & 2 & 0 \end{bmatrix} \quad (2.2)$$

If G is connected, meaning that the graph is composed by a single component, the Wiener index W and the detour index ω are the half-sums of the corresponding matrices:

$$W = \frac{1}{2}\sum_{i,j} D_{ij} \qquad (2.3) \qquad\qquad \omega = \frac{1}{2}\sum_{i,j} \Delta_{ij} \qquad (2.4)$$

Thus, for the graph G, it is possible to compute its indices, $W_G = 14$ and $\omega_G = 28$, whereas for a 5-complete graph the results are: $W = 10$ and $\omega = 10$. These results are mainly used in chemistry. For instance, if they are computed only between carbon atoms these measures can help to determine boiling points of molecules.

Yet, those measures are hard to compute for bigger graphs such as road networks. Moreover, the definition of detour index used in chemical graph theory does not make a lot of sense for road networks since they are related to human mobility. Indeed, comparing length of shortest and longest paths in road networks is not the meaningful since people are never using longest routes to go from one point to another.

Therefore, in human mobility, the notion of detour index (DI) represents by how much the shortest path differ from the distance in straight line between two points. For instance, this index has been explored by Lee et al. [21]. In their work, the detour index is a measure representing the length of the extension of the shortest paths comparing to the euclidean distance of an origin-destination pair *OD*. Thus, to generalise this notion, they extended it to compute the *DI* of a city by averaging the detour indices of every *OD* samples of this specific city.

The detour index is a good topological indication of the efficiency of the network of the city. Also known as circuity [23], the detour index of an *OD* pair is computed as:

$$\mathcal{D}_{OD} = \frac{d_{OD}}{s_{OD}} \qquad (2.5)$$

where d_{OD} is the length of the shortest path from the origin to the destination and s_{OD} is the euclidean distance. The closer it is to 1, the less the shortest path deviate from the straight line between the departure and the arrival locations. Therefore, when the average *DI* is low, the network is, theoretically, more efficient than a city with a detour index of 2. However, this measure is only a topological efficiency measure and does not reflect the behavior of road users and the congestion issues that they could cause.

2.1.3 $\mathbf{K_{road}}$

Unlike the detour index which is a topological, the K_{road} is a usage measure. The K_{road} assesses the attractiveness of a road segment by counting the number of locations, either city neighborhoods or tiles of a tessellation, responsible of the majority of the traffic flow on this particular road segment [38].

To compute the value of the K_{road}, it is compulsory to build the *network of road usage*. This graph is a bipartite graph gathering, on one side, the edges of the road network and, on the other side, the major driver sources (MDS). The MDS of a road segment are the top ranked areas that produces 80% of the traffic flow of this edge. Once the bipartite graph defined, the K_{road} of each edge is computed by taking the degree of

the corresponding node, that is to say to count of the number of neighborhoods nodes in its direct vicinity. The Figure 2.2 provides an example a the bipartite graph of road usage and the associated K_{road}.

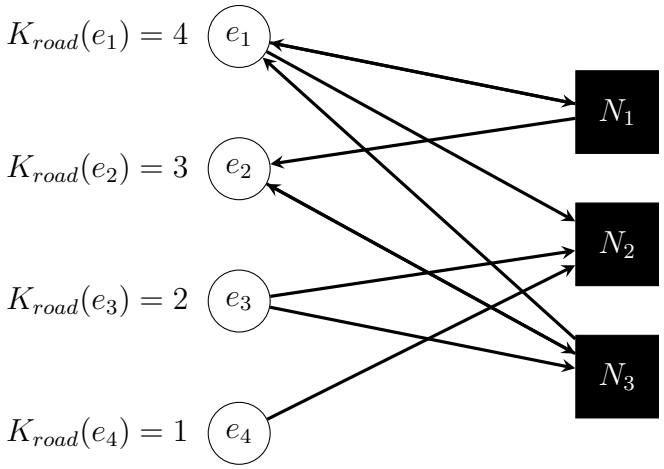


Figure 2.2: Bipartite network of road usage with edges e and neighborhoods N. K_{road} is the degree of edge nodes from and to the MDS.

The value of K_{road} represent the diversity of the origins of the drivers. Thus, an edge e with a low K_{road} shows that only a small number of vehicles use this segment. Conversely, an edge with an high value of K_{road} gathers drivers from multiple MDS, therefore indicating a high popularity for this segment.

2.1.4 Radius-fixed sampling

One of the first challenges when studying human mobility is to gather data. Most of the time, when the main research topic is about human behavior, the data are gathered by collecting GPS coordinates of real users. Conversely, if the research is focused on topological aspect of the road networks, it is mandatory to sample geographical locations across the studied city. However, these points should be well chosen to correctly represent the entirety of the city or of its users.

Among all the algorithms developed to sample locations, there is one called the radius-fixed sampling. First introduced by Lee et al. [22], this procedure consists of generating *OD* pairs on concentric circles of different radii. This process defines the centre or the circles by using the coordinates given by latlong.net[1]. Then, it creates concentric circles of desired radii around these coordinates. Once created, it generates samples regularly all around each circle. Finally, each point is paired with every other sample of the same circle to produce a set of *OD* pairs.

[1] `https://www.latlong.net/`

2.1.5 Alternative routing

The alternative routing (AR) algorithms were created to find existing paths between two locations that is not the shortest one. The reason to that is the will to reduce traffic load on overused road segments by sometimes giving different paths to drivers having the same travel demand. Most of these algorithms are still based on a shortest paths algorithm even if their aim is not always to provide the shortest paths. Indeed, most of the AR algorithms follow the same principle. First, the graph is altered according the chosen AR procedure, for instance, some edges can be removed from the graph to avoid getting path following a certain road segment. Then, a shortest path algorithm is computed on the modified graph to get a new path. Usually, the shortest path algorithm that is used is the Dijkstra algorithm.

Dijkstra Algorithm

The Dijkstra algorithm [13] is the most well-known shortest path algorithm of graph theory. Its application are very diversified. For instance, this algorithm is used frequently in telephone networks or in networks protocol such as Internet Protocol (IP). It is also the starting point of most of the routing algorithms since it can be adapted to optimised paths according to many different criterion such as length, time or toll's cost. Unlike many other shortest path algorithms, Dijkstra's one not only compute the shortest between two specific nodes. Indeed, it allows to compute the shortest path from a certain node to every other nodes of the graph.

Even though this algorithm has some intrinsic limitations, such as the fact that it can only be applied to positive weighted graphs, the Dijkstra algorithm is one of the fastest algorithm to compute shortest paths. Thus, in graph theory, most of the paths algorithms are based on this one and the following alternative routing algorithms are no exception.

Graph randomisation

The graph randomisation [8] is an AR algorithm based on Dijkstra algorithm. This algorithm is composed by a loop containing two main steps. First, the Dijkstra algorithm is applied to compute the shortest path on the current graph. The second step consists on altering the weight of the edges by adding a random noise computed as follows:

$$noise(e) = \mathcal{N}(0,1) \times (w(e)^2.\delta^2) \tag{2.6}$$

where e represents the edge, $w(e)$ its weight and δ is a randomisation parameter that controls to what extend the network is modify. Once the noise is computed, we can derive the new edges' weights by computing:

$$w_{new}(e) = max(w(e) + noise, \tau) \tag{2.7}$$

where τ is a control parameter avoiding the weights to have too small values or even null values. For instance, having $\tau = 100$ means that all new weights are at least of

value 100, and thereby preventing negative values within the network. However with $\tau = 0$, it is possible to avoid only strictly negative weighted edges without forcing the edges to have a minimal value.

After updating the weights of the edges, the algorithm loop and a new paths is computed by the Dijkstra shortest path algorithm.

Path Penalisation

The path penalisation algorithm has the same construction as the graph randomisation but is more specific in the altering part. Indeed, it only targets the edges of the shortest paths already added to the set of paths instead of the whole graph. At each iteration of the loop, the algorithm compute the shortest paths between the origin and the destination and then penalise every edges of this path by using:

$$w_{new}(e) = w(e) \times (1 + p) \tag{2.8}$$

where p is the penalisation factor. For instance, if the penalisation factor is $p = 1$, the edges' weights will be double.

2.2 Related Works

2.2.1 Paths

The diversity problem in road networks is not only a graph theory problem but also a human mobility matter. Thus, in the first part, we formalise the meaning of realistic and unrealistic trajectories from a human mobility perspective and the importance of this concept to understand the human car mobility behavior. Then, this Subsection presents some existing methods of path enumeration and their associated problems. In the second part, we present the major problematic of counting the *OD* paths in general graph theory and not only in road networks and the impact of this issue on our work. Finally, the third part introduces a paths generation algorithm to palliate the aforementioned problems and to generate huge sets of simple realistic paths between two locations.

Realistic paths

In human mobility, the notion of realistic trajectories refers to the paths following the patterns of human routing behavior defined by the literature.

The first property to define realistic paths is the concept of simple path. In graph theory, a simple path is a sequence of vertices, that can either be finite or infinite, composed by unique vertices, and therefore by non repeating edges. In human mobility, we translate this idea to the paths that never follows road portions that were already taken. Simple paths are a *sine qua non* condition of realistic path since it is very unlikely for people to follow an identical portion of street when going from one place to another.

Moreover, the existence of a strong spatial limitation was enlightened by Lima et al. [1] in 2016. After normalising the GPS traces of over 500 road users within four different cities, it appears that independently of the city layout, 95% of the overall drivers' positions are contained within an ellipsoidal bound around the origin and the destination.

However, even within this universal spatial bound, some paths remain unrealistic from drivers point of view. Indeed, it is possible without exiting the restricted area to take unnecessary detours increasing thereby both, travel time and traveled distance. In Figure 2.3, we provide a small example of the existing unnecessary detours in a road network. On this figure, we represent a road network as circles and lines cooresponding respectively to the nodes, that is to say the intersections, and the edges, namely, the road segments.

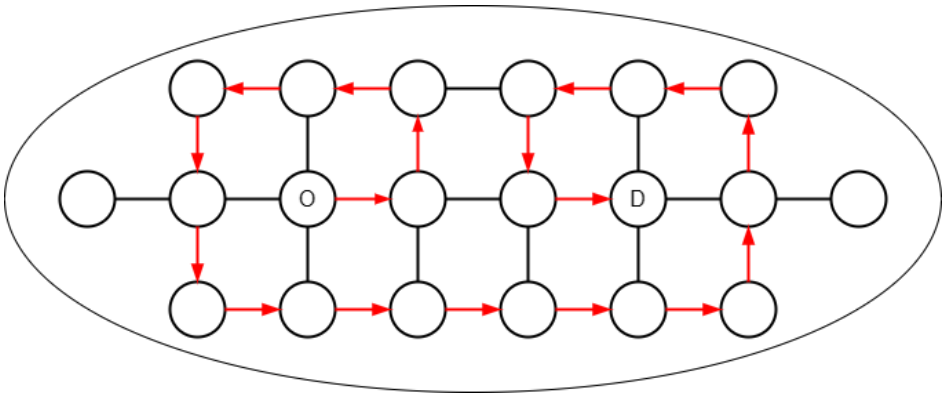


Figure 2.3: Example of unnecessary detours in an elliptic road subgraph around the origin nodes o and the destination node d. The circles represents the intersections of the road network and the edges are the road segments. The red directed path represent a trip from origin to destination with many unnecessary detours.

In this example, we analyse a small elliptical subgraph around the origin o and the destination d. To simplify the case study, we consider that each edge is bidirectional and have a length of 1. Here, the shortest path between o and d is the one following the straight line between the two vertices, thereby creating a trip of length 3. Therefore, the red path on the figure matches all the previous conditions to realistic paths. However, the length of this trajectory is equal to 17.

We can consider the detour index DI, presented in Subsection 2.1.2, as an efficiency measure. This DI is evaluated by computing the ratio between the length of a path and the euclidean distance from the origin to the destination. Thus, on Figure 2.3, the DI of the shortest path is equal to 1 since it fits exactly the euclidean distance between the origin o and the destination d, whereas the red trajectory's DI is: $\frac{17}{3} = 5.67$. Such a high detour index clearly shows the inefficiency of these detours making them unrealistic and, therefore, unnecessary to consider in our work.

To summarise the notion developed in this Section, in human mobility, realistic paths refer to simple paths contained within an ellipsoidal spatial bound around the origin and the destination and whose length does not exceed the shortest existing path

from a coefficient ϵ that needs to be fixed during the study process. Thus, to efficiently study the diversification of paths in a city, it is important to restrict the problem only to the realistic ones.

The paths counting issue in graph theory

Like most of the counting problems, the paths counting problem is known to be a #P-complete problem [24]. Those problems form a complexity class in the computational complexity theory that respect two properties. First, they have to be #P problems representing the counting problems associated with the NP problems. Formally, a function belongs to the #P class if there exists a non-deterministic Turing machine running in polynomial time such that for each instance x, $f(x)$ is the number of paths accepting x. Every #P problem, in order to be classified as #P-complete, needs to be #P-hard, meaning that every other problem #P can be reduced to this particular problem using either a Turing reduction or a polynomial-time counting reduction.

Thus, it is unrealistic to try to compute the exact solution to such problems. Indeed, the only way would be to use brute-force methods, in the counting and enumerating paths problem, either a Depth-First Search (DFS) or a Breadth-First Search (BFS) algorithm. However, both alternatives are not relevant considering the computational time they would require on road networks.

To palliate the computational limitations faced when trying to count or enumerate simple paths, some approximation algorithms were designed [27, 35, 41] either to fasten the computational time by decreasing the size of the graph or to approach the final count. In 2006, B. Roberts and D. P. Kroese mathematically formalised an estimation formula [32] via Monte Carlo simulation.

Yet, those methods are applicable in graph theory but they show some limitations when it comes to human mobility. Indeed, some of them are restrained to acyclic graphs or undirected graphs which does not correspond to the characteristics of real road networks. Moreover, some estimations do not allow any differentiation between realistic and unrealistic trajectories as presented in the first part of this Subsection because they only give an approximation of the number of paths without enumerating them.

In conclusion, the path counting problem creates major issues when studying paths diversification since one of the compulsory steps of the process is to enumerate lists of realistic paths between two locations to pursue further analysis on them.

Paths generation algorithms

As explained previous part of this Subsection, the exact number of realistic paths existing between two nodes in a graph is very time-consuming and expensive, thereby making it not feasible for real-application on graphs as large as road networks. However, it is still doable to generate sets of existing paths between two locations by using alternative routing algorithms. Indeed, these algorithms were designed to provide not only the shortest-path but also some alternative paths between the origin and the destination. These algorithms were first formalised from a theoretical aspect in graph

theory literature, but were then applied to real road networks and to human mobility.

The first naive approach would be to compute the k-shortest paths [15] for each pair of locations. This problem, known as the KSP problem, is well-studied in graph theory and has been the subject of many publications in scientific literature. Many KSP algorithms were described through time. Among them, we can find Yen's algorithm [42] that starts from a shortest path and keeps the root of this path to find potential alternatives by deactivating the edges of the shortest paths already computed. This algorithm can generate a set of shortest paths, but the paths created through this method are very similar to each other since the algorithm uses the root of the previous paths.

To guarantee a maximal diversity between all the paths computed, it is possible to used k-shortest disjointed paths (KD) algorithms [36]. For instance, the use of the path penalisation algorithm, described in Subsection 2.1.5, with a penalisation factor $p = +\inf$, will force the Dijkstra algorithm to completely avoid all the edges previously used. Thus, it will result in a set of paths fully disjointed. Instead of fixing the weight to the infinite value, it is also possible to directly discard the edges already used and recompute Dijkstra on the reduced graph [20]. Even though disjoint paths algorithms maximise the diversity of the edges used, it does not provide a sufficient set of paths since its cardinality is bounded by the number of outgoing edges of the starting node and the number of incoming edges of the ending node.

To extenuate the diversity problem of the KSP and avoid the cardinal issue of the KD, it is possible to resort KSP algorithms including a diversity aspect during the generation of paths. The first method is a filter method which, theoretically, takes the set of all the simple paths between an OD pair, sort it by length and add each path to the final set of diverse paths only if the similarity metric of the final set remains under the chosen threshold. This method is not a generation method in itself since it requires a set of paths. Yet, it can be use as a step in a larger protocol to generate short and divers paths, for instance the k-most diverse near shortest path (k-MDNSP) [19]. This algorithm computes a set of k paths with a high dissimilarity that are not exceeding the threshold defined by the user.

Yet, these algorithms were created to generate sets of paths with a relativly small cardinality, for instance, to be used as routing algorithms, especially for the k-MDNSP. However, to assess the diversity of paths, we need to analyse profuse paths and, therefore, to generate way bigger sets of paths. The chosen procedure combines two existing algorithms, the graph randomisation and the path penalisation, introduced by Chang et al. [8] in 2019 and presented in Subsection 2.1.5. The nested algorithm we developed to achieve this goal will be detailed later in Subsection 3.1.4.

2.2.2 Impact of graph's topology in human mobility

In the field of human mobility a good part of the research led are concerning the recurring patterns of human behavior when using road network such as favorite routes, geographical limitations or the adaptation of itineraries to congestion [1, 40]. However, to fully understand the difference of conduct between users of different cities, it is

required to study not only the drivers and their habits but also the topology of road networks they are using.

Thus, in this Section, we will present the results obtained by researcher studying the layout of road networks. In the first part of this Subsection, we will present an article studying the notion of K_{road} and the betweenness centrality b_c alongside to elaborate a classification of roads not only relying on human behavior regarding these edges but also on graph topology. Then, the second half of this Subsection introduces the conclusions of scientists who studied the detour index on 70 cities to understand how to assess the efficiency of a road network.

Topology and Road usage

To understand the patterns of drivers in urban area, Wang et al. [38] studied a classification of road segments from a topological and from a road usage perspective. To classify the edges from both points of view, the K_{road} and the betweenness centrality b_c are computed for each road segment. Usually, in graph theory, this centrality measure is computed on vertices to assess their centrality within a graph [16, 28]. However, the measure has been extended to edges to assess their centrality in the graph [7, 25]. The betweenness centrality of an edge b_c represents the proportion of shortest path between two vertices of the graph following that specific edge.

After assessing the popularity, with the K_{road}, and the centrality of an edge, the road segments are classified in four different classes. A road segment can either be a *connector*, a *peripheral connector*, an *attractor* or a *local*, according to the value of its K_{road} and its b_c, as explained in Table 2.1. The values are considered *high* if they are within the top 25% of this specific measure, otherwise, they are considered *low*.

		K_{road}: high	K_{road}: low
b_c	high	Connectors	Peripheral Connectors
	low	Attractors	Locals

Table 2.1: Road classes according to K_{road} and b_c.

After classifying the edges into four categories, they studied application of this classification. To do so, the free flow travel time t_f and the actual travel time t_a and the additional travel time t_e are computed. The free flow travel time is the theoretical time to travel the edge at maximum speed limit and the actual travel time is the real time necessary to travel a road segment. Finally, the additional travel time represents the difference between the free flow travel time and the actual travel time, giving that: $t_e = t_a = t_f$.

These measures can be computed for road users to study if they are following edges that are congested are not. Indeed, the higher t_e is, the more congested roads the driver uses. To generalise the results, all three of the total travel time are computed for each driver source by summing the travel times of each users starting from this source. The

results showed that some driver sources have total additional travel time that are 16 times higher than the average. What can be deducted from this discovery is that the congestion is generate only by a few driver sources. Thus, if drivers from these sources use more diversified routes, it could result in the decrease of global congestion in urban area.

Detour Index

The topology of road networks is a quite recent field of research. Indeed, the detour index *DI*, presented in Subsection 2.1.2, has been studied very recently by Lee et al. [21] in February 2023. In their work, they compare the detour index of 70 cities distributed on every continent. To better capture and understand the core-periphery structure of cities, the sampling methods used in this work is the radius-fixed method [22], presented in Subsection 2.1.4.

The detour index can also be constrained and computed with various restrictions, as shown by the Equations 2.9, 2.10 and 2.11.

$$\mathcal{D}_r = \frac{1}{N_r} \sum_{(O,D)|r_{OD}=r} D_{OD} \tag{2.9}$$

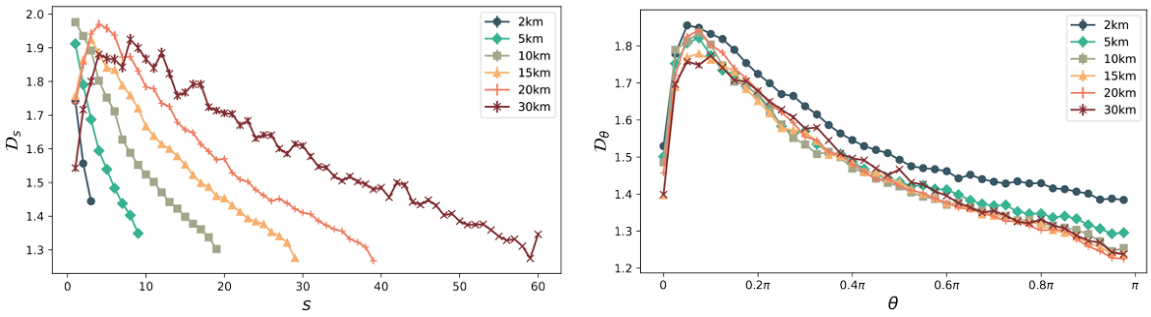
$$\mathcal{D}_s = \frac{1}{N_s} \sum_{(O,D)|s\leq s_{OD}<s+\Delta s} D_{OD} \tag{2.10}$$

$$\mathcal{D}_\theta = \frac{1}{N_\theta} \sum_{(O,D)|\theta\leq \theta_{OD}<\theta+\Delta\theta} D_{OD} \tag{2.11}$$

where r, s and θ represents respectively the radius on which the samples were created, the euclidean distance between origin o and destination d and the angular distance. All of these measures were computed for each city by averaging on every sample and, then, the curves were plotted according to the different parameters, radius, distance and angle to draw conclusions.

The first conclusion to draw is that the *DI* does not depend on the radius of the circle on which the data where sampled. Indeed, the plot of $\mathcal{D}_r$ according to the radius r is a flat curve of approximately $\mathcal{D}_r = 1.5$. Therefore, the departure and arrival neighborhoods, the city centre, the inner suburb or the far suburbia, are not impacting the DI.

Then, they explored the impact of euclidean distance s on the detour index $\mathcal{D}_s$ for samples on each radius as shown in Figure 2.4a, representing a chart they generated for their article. What is to see is that the detour index is decreasing when the distance between origin and destination increases, relatively to the radius on which they were sampled. This shows the impact of the core-periphery structure on the DI. Indeed, when the euclidean distance reaches twice the size of the radius, then the samples are on opposite sides of the ring, meaning that the shortest path passes trough the centre of the city. On the opposite, when the distance is closer to 0, the *DI* represent the periphery, the higher the radius, the further the samples are from the centre. Thus, the plot of $\mathcal{D}_s$ represents the global topology of cities across the world.



(a) (b)

Figure 2.4: Average detour index over 70 cities given by Lee et al. [21]. (a) with respect to s, the euclidean distance between the origin o and the destination d. (b) with respect to θ, the angular distance formed by the origin o, the destination d and the centre of the city.

This observations are corroborated by the plot of $\mathcal{D}_\theta$, Figure 2.4b, drawn from Lee et al.'s work [21]. Indeed, when the angle increases from 0 to π, the samples or not only moving away from each other but also moving to opposite sides of their common circle. Thus, what is noticeable is that when locations are on either sides of the city centre, the *DI* decreases to $\mathcal{D}_\theta \approx 1.25$. Therefore, showing that regardless of the distance between the departure and the arrival, the shortest path will not exceed, on average, the euclidean distance by more than 25%, thereby proving that the density of centre of cities are denser than the peripheries.

Finally, by studying this process over 70 cities located in the whole world, the conclusions drawn are applicable in every city, with the only difference being the magnitude of the curves. Indeed, even though they follow the same *DI* rules, some cities remain more efficient than the others and have lower averaged detour index.

Another meaningful conclusion that can be drawn from the experiences conducted in this article, concerns the impact of the layout of a road network in the value of the detour index. To illustrate this, we can focus on Northern American cities and their specific grid shaped road networks.

To facilitate the reasoning, we can consider a road graph shaped as a perfectly regular squared grid. On this network drivers can only follow the edges of the grid and there is no possible way to cut through the blocks. In this particular road layout, the length of the shortest path between every possible *OD* pair is the ℓ^1 norm. Also known as the *Manhattan Distance*, the value of $||(o,d)||_1$ for two points o and d is the sum of the coordinates of the corresponding translation vector, as given by Equation 2.12. Conversely, the euclidean distance $||(o,d)||_2$ of the same points represents the square root of the sum of squares of the coordinates, as provided by Equation 2.13.

$$||(o,d)||_1 = (x_d - x_o) + (y_d - y_o) \tag{2.12}$$

$$||(o,d)||_2 = \sqrt{(x_d - x_o)^2 + (y_d - y_o)^2} \tag{2.13}$$

If we consider a triangle between o, d and a point $m = (x_o, y_d)$, squared in m, the norm ℓ^1 represents the sum of the length of om and md, and the norm ℓ^2 is the length of the hypotenuse od of this triangle. Therefore, for all pair in the grid network, there is a squared triangle such that DI_{od} is the ratio between the sum of the two sides and the hypotenuse. Furthermore, if we consider the "worst case" for these triangles, that is to say an isosceles right-angled triangle with its side of length a, we have that:

$$\mathcal{D}_{OD} = \frac{||(o,d)||_1}{||(o,d)||_2} = \frac{2a}{\sqrt{a^2+a^2}} = \frac{2a}{a\sqrt{2}} = \frac{2}{\sqrt{2}} = \sqrt{2} \tag{2.14}$$

Therefore, if we consider a city with a regular grid shaped road network, the detour index of every pair of location OD will be, at most, equal to $\sqrt{2}$, that is to say approximately 1.41, as detailed by Equation 2.14.

What is striking with this result is that the theoretical value of the DI of cities having grid layouts, such as the Northern American ones, matches perfectly the empirical results obtained by Lee et al. in their article. Indeed, when plotting the average detour index of all the 70 cities studied, we notice that all the northern American cities are quite densely grouped around the value 1.4. Conversely, cities from every other continents show much more distributed average detour index. For instance, Europe have cities with average detour index from 1.3 to 1.6. East Asia shows even less stable results with DI between 1.2 and 2.0.

Chapter 3

Methods

In this chapter, we detail the procedure to analyse the diversification of routes in urban environments. In Section 3.1, we present the method to generate sets of thousands of realistic paths in an reasonable running time. In Section 3.2, we introduce the three real urban areas on which the experiments were proceeded: Milan, Rome and Florence (Italy).

3.1 Paths diversity procedure

To efficiently assess the potential diversification of itineraries in urban area, we developed a methodology to drastically reduce the complexity, and, therefore, the running time of the paths generation between an origin and a destionation in an urban road network. In Subsection 3.1.1, we explain the process to obtain the road networks of the studied urban area. Then, Subsection 3.1.2 provide detailed explanation of the sampling algorithms that are used in our work to generate *OD* pairs. In Subsection 3.1.3, we use the discoveries of Lima et al. [1] highlighting a spatial bound of individual trips to radically reduce the size of the studied graphs. Finally, in Subsection 3.1.4, we present the paths enumeration strategy we developed to efficiently generate sets of several thousands of paths.

3.1.1 Road Network

A road network may be modeled by a directed graph $G = (N, E)$ in which the set of nodes N represents the intersections and the set of edges E refers to the roads. Instances from both sets, edges and nodes, have several attributes describing them. For instance, edges have their length in meters and the free flow travel time from the starting to the ending point, that is to say, the edge length divided by the maximum speed allowed on that specific edge. On the other side, the nodes have, among others, the coordinates of the intersection.

When studying road networks, as we do in this work, the term *road* does not have the same meaning as it does in the everyday language but refers to portions of street. Indeed, when one refers to a road in the common language, it designates a street which

has a name and which crosses other streets all along its outline. Conversely, in this work, the employment of the word *road* pertains to the portions of streets between the intersection. When building the graph, these intersections are used to create the nodes of the graph and the street portions engender the edges of the net.

To build the road network of a city, it is first mandatory to collect all the aforementioned data. These data are collected from OpenStreetMap (OSM[1]) and then used to recreate the road network using SUMO[2]. SUMO is a python tool allowing to simulate real like road traffic and to visualise it with graphical user interface. However, since our work focuses on the topology of road networks, the traffic simulation tool is not useful. The only tools necessary to use are the road network downloading tool, OSMWebWizard and the function to get neighboring edges from GPS data.

Thus, after downloading the network and translating latitude and longitude coordinates to the closest edge, we build the corresponding graph using the python iGraph[3] library collection to facilitate the analysis on the road network. This library is open source and has been developed to easily manipulate networks in C language that was then enabled in python through an interface. Using an iGraph object rather than a SUMO network offers many advantages such as already implemented paths or routing algorithms. Moreover, iGraph being a more common library, the documentation is more detailed and available, and the community is bigger and more active making it easier to use.

Thus, the SUMO networks were translated to graphs while conserving the structure and the important attributes such as the length and the travel time of the edges. The nodes geographical locations were also translated from the SUMO coordinate system to the classic latitude-longitude system. However, every objects of the new graph keeps in memory the *id* corresponding to its own version in SUMO to access it easily if needed.

3.1.2 *OD* sampling

To efficiently measure the diversification of paths within a city, we need to study a set of alternative paths between an origin point o and a destination point d within the road network. Such a pair (o, d), also denoted as *OD* pair, represents a trip within a mobility demand. Usually, in human mobility, these trips are used to study flows with *OD* matrices, where each column and row represents a geographical area (e.g., a GPS point or a tile of a tessellation) and the cells of the matrix represents the flow between the geographical areas.

In our work, we study the alternative paths between two locations and not between geographical areas. Therefore, we need several origin-destination pairs between which computing the sets of paths. However, it is not possible to analyse the diversification between every *OD* pairs for computational matters, hence, we need to select a sampling method, namely, a process to select origin-destination pairs within the road network whether it is at random or by an existing algorithm. After gathering the pairs of locations, we infer the sets of alternatives paths for each of them. These sets are

[1] `https://www.openstreetmap.org/`

[2] `https://sumo.dlr.de/docs/Tutorials/OSMWebWizard.html`

[3] `https://python.igraph.org/en/stable/`

then analysed to assess the diversification of a city. Indeed, the results can be extrapolated with respect to probabilistic forecasting process such as Monte Carlo method to understand the city's diversification.

Mobility demand

The first method to select the *OD* pairs on which deepening our analysis is the mobility demand selection. The mobility demand is a set of time-ordered trips. Each trip is composed by some attributes, in particular the starting and ending edges. This selection consists of randomly picking trips from the mobility demand file and extract their starting and ending edges to add them to the set of *OD* pairs.

This algorithm consists of collection trips demand data and storing them in a JSON file. Each of them is composed by a few attributes, in particular the starting and ending edges. Once the file is downloaded, we only need to pick *OD* pairs randomly as provided by the travel demand.

This sampling can provide a good dataset to capture the diversification of paths within a city. However, it shows some limitations when it comes to comparing the diversity according to the distance between origin and destination. Indeed, this type of sampling usually shows a lack of representation of more distant locations whereas closer locations are far more represented. This deficiency of representation of long distance *OD* pairs risks to create bias and to distort the final results. Indeed, a high cardinality set can be used to extrapolate real results thanks to the law of large numbers, conversely, applying this probabilistic law can not certify good results with a set of very few samples.

Fixed-radius sampling

An alternative sampling algorithm to palliate this distribution issue is the fixed radius sampling method [21, 22]. As presented in Subsection 2.1.4, this algorithm creates concentric circles of different radii, centred on the city's centre location. Then, the algorithm samples points regularly all around each circle to produce a set of *OD* pairs starting and ending on the same ring. Ideally, the interval between two consecutive points respect a minimal angle to avoid having too close points, especially for smaller radius circles. Indeed, if we sample 36 points on a 1km radius circle, we obtain a point every 10°, meaning that the distance between two consecutive points is: $\frac{2\pi}{36} = 0.174km$, that is to say 174m. Building closer samples increases the risks of noisy data when pursuing further analysis on them.

This algorithm allows to generate large sets of *OD* pairs. Indeed, if we consider radius over the range 1, 2, 5, 10 and 20, we generate a set of $5 \times \binom{36}{2} = 3150$ *OD* pairs that can be considered in both directions, thereby representing 6300 origin-destination samples.

Unlike the travel demand selection, this algorithm have a minimal amount of samples for each distance. Indeed, this technique samples up to a chosen n locations on each circle, thereby forcing the algorithm to create at least n *OD* samples for each

diameter, since every location is paired with the one on the opposite side of the ring. Thus, it allows to ensure at least few samples, even for bigger distances.

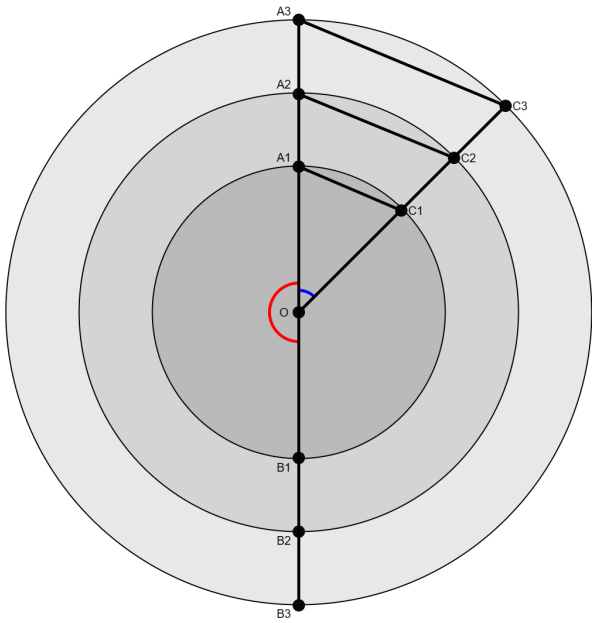


Figure 3.1: Illustration of the core-periphery structure with the fixed radius sampling. Samples on opposite sides of the circle (AB) all passes through the city center. They are measurable with an angle of $\pm\pi$ (red). Conversely, the peripheral trips are studied with smaller angles (blue). The combination of angle and radius allows to study inner suburb (A_1C_1) of further suburban areas (A_3C_3).

Moreover, the other reason that led us to use this sampling method is the inherent topological advantages provided by the data. Indeed, unlike other sampling algorithms, the fixed radius method not only allows an analysis on the distance but also on the core-periphery structure due to the angle property. As we can see on Figure 3.1, when the angle reaches 180° (red), the straight line from points A to points B always pass through the city centre o independently of the distance between the origin and the destination. Conversely, when the angle is of 45° (blue), the straight lines from locations A to locations C are going through the suburban area and even further in the periphery when the radius is increasing.

Thus, the use of the fixed-radius sampling method can provide great interest when studying the topology of a city and, especially, the core-periphery structure since it becomes analysable only by plotting results according to the radius and to the angle.

3.1.3 Elliptic subgraph

To respect the spatial boundary presented in Subsection 2.2.1, we compute the elliptic subgraph surrounding each OD pair, as shown on Figure 3.2. The use of this restricted graph allows not only to generate realistic paths between o and d but also to drastically

reduce the computational time, in particular for closer locations whose ellipses have small area.

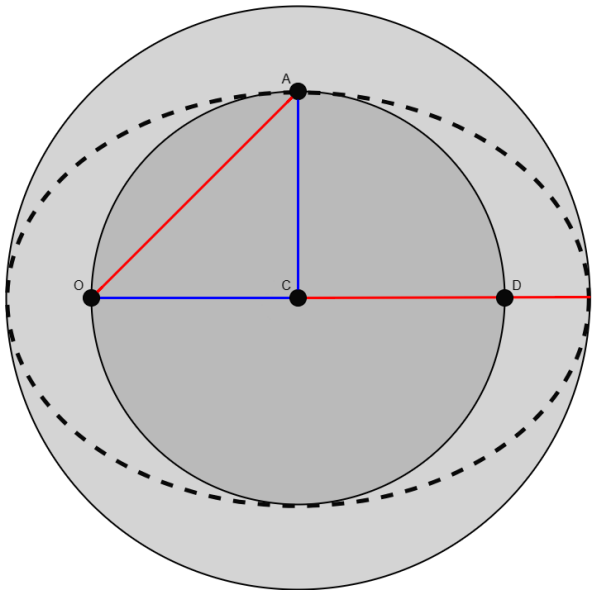


Figure 3.2: Visualisation of the eccentricity of the ellipse around the origin o and the destination d. The semi-minor axes (blue) is half the distance between o and d and the semi-major axis (red) is computed as the hypotenuse of the triangle formed by the origin o, the center c and the point A.

The Figure 3.2 shows a visualisation of the ellipse built around the OD pair and its eccentricity. The eccentricity of an ellipse e refers to the relation between the distance from the centre of the ellipse C to the focal points o and d, named f (blue) and the semi major-axis, noted a (red). To compute the eccentricity of en ellipse, we use the formula:

$$e = \sqrt{1 - \left(\frac{b}{a}\right)^2} \tag{3.1}$$

where a is the length of the major semi-axis and b is the length of the minor semi-axis. Thus, the eccentricity of the ellipse in Figure 3.2 is: $e = \sqrt{1 - \left(\frac{r}{\sqrt{2}r}\right)^2} \simeq 0.707$, where r represents half of the distance between the origin o and the destination d.

The choice of the eccentricity e may not respect the optimal value highlighted by Lima et al. [1]. The optimum eccentricity can be characterised by the value of e having the highest probability of capturing each shortest trip. If we consider an ellipse of eccentricity $e = 0.9$ and a semi minor-axis $b = r$, according to the Formula 3.1, the semi major-axis a is equal to:

$$e = \sqrt{1 - \left(\frac{b}{a}\right)^2} \implies a = \sqrt{\frac{b^2}{1 - e^2}} = 2.29r \tag{3.2}$$

This increase of the length of the semi major-axis leads to growth of the area of the ellipse. Indeed, when comparing both areas we obtain:

$$\begin{cases} \mathcal{A}_1 &= 4.43r^2 \\ \mathcal{A}_2 &= 7.19r^2 \end{cases} \tag{3.3}$$

where $\mathcal{A}_1$ is the area of the ellipse in the Figure 3.2 and $\mathcal{A}_2$ is the area of the aforementioned ellipse of eccentricity $e = 0.9$. Increasing the eccentricity of the ellipse from 0.7 to 0.9, extends the surface area by 60%, thereby increasing the size of the graph and the computational time. Therefore, choosing an eccentricity of 0.7 is a good compromise between the size of the elliptic subgraph and the computational time of the algorithm.

To characterise the restriction of the road network, we use the shapely[4] python library to describe an ellipse buffer around the *OD* pair as shown in the Figure 3.2 (dashed). Once the buffer is formalised, we iterate over each edges of the road network and check whether both $from$ coordinates and to coordinates are included in the shapely buffer. If one of the coordinates is not within the ellipse, the edge is discarded and is not included in the final subgraph.

3.1.4 Paths enumeration

To assess the diversity of paths between an origin and a destination point within a city, we need to enumerate a certain amount of paths. However, as shown in the Subsection 2.2.1, enumerating all the existing paths is not realistic in high cardinality graph such as road networks.

In 2019, Chang et al. [8] established two methods to compute and enumerate simple paths, the graph randomisation and the path penalisation, both presented in Subsection 2.1.5. Both algorithms allow to modify the weight of the edges to mislead Dijkstra shortest path algorithm into finding alternative paths. They also allow to control how diversified the set of paths generated is.

The graph randomisation process allows to control how much the network is modified. Indeed, the parameter δ (Eq. 2.6) is the randomisation factor controlling the modification. For instance, if we consider $\delta = 0.1$, the new weight of more than 95% of the edges will be within $w(e) \pm 0.2w(e)$. Theoretically, there is a 68% chance that the noise of the edges is within $\pm\delta.w(e)$ and the probability reaches 95.4% if we consider a noise between $\pm 2\delta.w(e)$. By modifying almost all the edges' weight the shortest path algorithm will more certainly find a new path as being the shortest one. Moreover, the higher δ is the higher the noise is, thereby maximising the changes of the weights and the chances of obtaining simple paths differing from the shortest one.

On the other side, the paths penalisation has the penalisation factor p (Eq. 2.8). Having an important penalisation factor causes bigger changes in the weights of the edges, thereby accelerating the alternative paths generation. Nevertheless, having a too high factor p can force the k-shortest paths algorithm to completely avoid the previously used edges, for instance by fixing the penalisation factor to $p = +\inf$ Dijkstra will avoid

[4] `https://shapely.readthedocs.io/en/stable/`

these edges and the set of paths will be completely disjointed. Thus, it is important to avoid an excessive penalisation factor to prevent the generation of too much unrealistic paths.

In their work, Chang et al. [8] analysed both algorithms separately. However, the results of the article shows that the randomisation algorithm on its own produces not so diversified paths comparing to penalisation procedure with a penalisation factor being infinite. Conversely, a path penalisation with $p = +\inf$ provides sets of alternative paths whose length are more than twice as long as the results given by the other paths generation methods.

Therefore, to drastically increase the amount of paths generated without having either sets of too similar paths or sets of too long paths, we decided to combine both algorithms. Our nested procedure, given by Algorithm 1, consists of applying first a graph randomisation and then computing n_p paths through paths penalisation. This process is cumulative, meaning that at each iteration, the new weights are computed based on the previously altered ones. Indeed, the weights are not reset after each randomisation or penalisation but they are the starting graph of the next iteration of the algorithm.

Algorithm 1: Paths enumeration procedure

Data: $G(N, E)$, origin edge o, destination edge d, number of graph randomisation n_i, number of path penalisation n_p
Result: π the set of alternative paths
$i \leftarrow 0$;
while $i < n_i$ **do**
  $path \leftarrow graphRandomisation(G, o, d)$;
  $\pi \leftarrow \pi + path$;
  $i \leftarrow i + 1$;
  $j \leftarrow 0$;
  while $j < n_p$ **do**
    $path \leftarrow pathPenalisation(G, o, d)$;
    $\pi \leftarrow \pi + path$;
    $j \leftarrow j + 1$
  end
end

This procedure is a quite good compromise between generating the shortest paths and diverse paths in terms of edge redundancy. Indeed, to compute the shortest paths it is possible to use directly Yen's k-shortest algorithm and to maximise the paths diversity it is possible to use the k-disjointed paths algorithm. Nevertheless, the combination of graph randomisation and path penalisation offers reasonable short and diverse paths.

Furthermore, this method enumerates the generated paths, thereby allowing us to recompute their length in the original unaltered graph and to discard those that are extending the shortest path too much to be viable alternatives.

3.2 Experimental Settings

After developing our experimental strategy, we apply it to three different urban environments: Milan, Rome and Florence, to understand the potential differences existing between them. However, to try to understand the impact of the structure of the road network, our analysis is not strictly limited to the administrative border of the cities but also include suburban areas. In Subsection 3.2.1, we describe the road networks for every city on which we applied the procedure presented in Section 3.1. In Subsection 3.2.2, we analyse the *OD* samples that are generated for each metropolitan area and their distribution. Finally, Subsection 3.2.3 provides description of the different sets of paths computed for each origin-destination pair.

3.2.1 Definition of the road networks

We apply the paths generation procedure detailed in the Section 3.1 to few metropolitan cities in Italy: Rome, Milan and Florence. Each studied graph covers the main city of the metropolitan area and some the neighboring municipalities.

To build the road network of Rome, we select a squared area of approximately $\mathcal{S}_R = 500km^2$ all around the centre of the city. Once the geojson[5] defined, we download the road network from OpenStreetMap to obtain the graph. The road network of Rome, noted $G_R = (N_R,\ E_R)$, is a directed graph composed by $|N_R| = 63,384$ nodes representing the intersections and $|E_R| = 31,798$ vertices referring to the road portions. Rome's road network is represented by Figure 3.3.

Figure 3.3: Rome's road network used for the experiments downloaded from OSM.

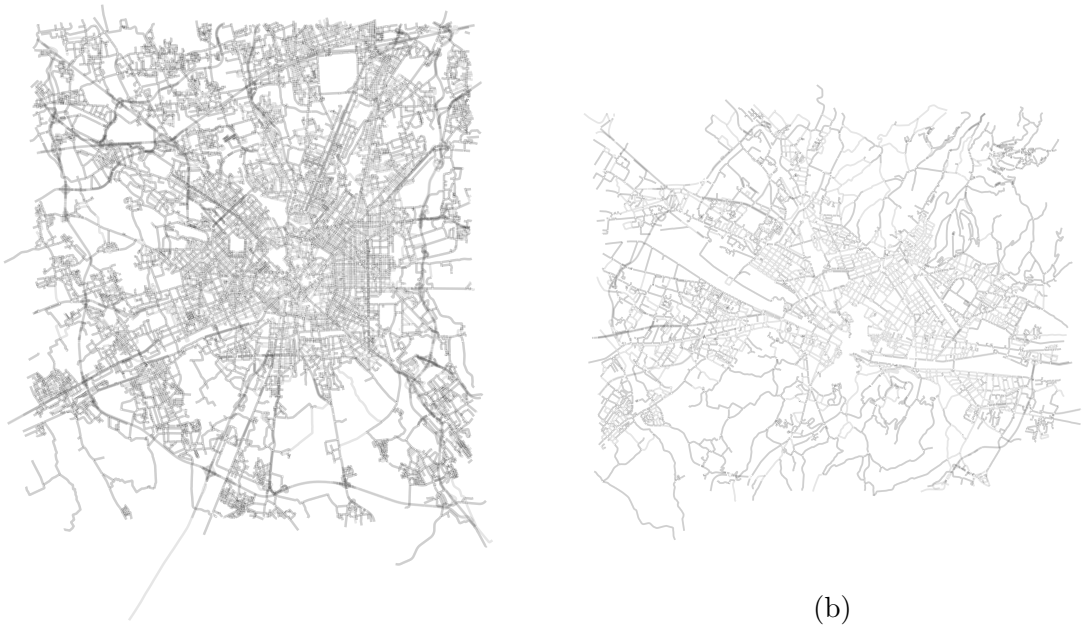
This process is extended to the cities of Milan and Florence, respectively plotted on Figure 3.4a and Figure 3.4b. The directed graphs of Milan and Florence are respectively

[5] `https://geojson.readthedocs.io/en/latest/`

designated by $G_M = (N_M,\ E_M)$ and $G_F = (N_F,\ E_F)$. For Milan's road graph, there is a total of $|N_M| = 46,488$ nodes and $|E_M| = 24,063$ edges to cover a surface area of around $\mathcal{S}_M = 350km^2$. On the other side, the road network of Florence covers approximately $\mathcal{S}_F = 90km^2$ with $|N_F| = 11,804$ crossings and $|E_F| = 6140$ road portions. All these data about the road graphs of each city are summarised in Table 3.1.

	Rome	Milan	Florence
$\mathcal{S}_i$	$500km^2$	$350km^2$	$90km^2$
$\|N_i\|$	63,384	46,488	11,804
$\|E_i\|$	31,798	24,063	6140

Table 3.1: Description of the road network of each urban area.



(a) (b)

Figure 3.4: Road networks used for the experiments downloaded from OSM. (a) Milan. (b) Florence.

The graphs are downloaded from OSM and stored in a SUMO file regrouping all the necessary data for the edges, such as length or incoming and outgoing edges, and also for the nodes, with latitudinal and longitudinal coordinates. The graphs are then translated to iGraph objects to facilitate the analysis.

3.2.2 Generation and selection of *OD* pairs

Once the road networks designed, we compute the *OD* pairs on which assessing the diversity. As explained in the Subsection 3.1.2, we use two different methods to generate locations throughout the cities. The use of different sampling techniques allow to palliate the potential weaknesses of the other set of *OD* pairs.

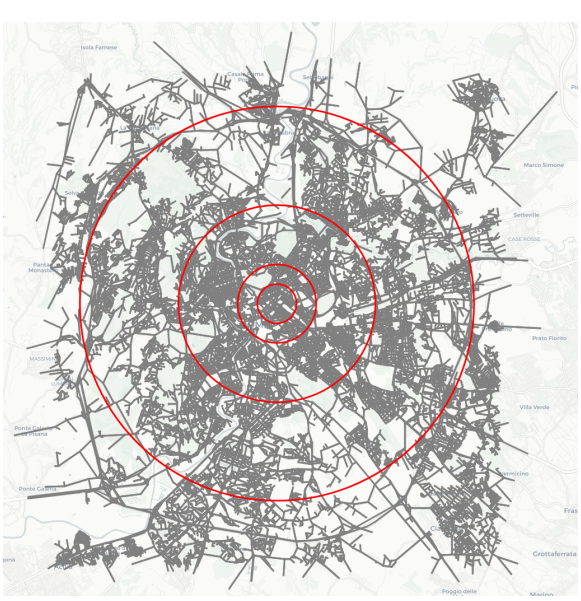
The first procedure consists of a random selection of origin and destination points from a mobility demand dictionary. The mobility demand is a file containing several thousands of trips. Each of them is characterised by a starting and en ending road segment. Usually, this type of mobility demand file is used to generate traffic flows within a city. To do so, the city is either divided according to natural administrative sections or separated into tiles to form a tessellation. A null matrix of size n^2 is also initialised, where n corresponds to the number of part of the city's division. Then, each travel demand is analysed to identify the starting and ending quarter and, the matrix is incremented consequently, namely the cell m_{ij} where i is the starting area and j the ending one is increased by 1.

In our work, the travel request are used as raw data. That is to say that the starting and ending points are exploited as provided by the file without needing any additional computations, such as tessellation. However, each city have its own mobility demand file with different size, for instance Florence's road network being much smaller than the Rome's and Milan's ones, its mobility demand provide only $10,000$ *OD* pairs. The number of samples available in each of the travel demand files are provided by the Table 3.2.

	Rome	Milan	Florence
# Travel demand	20,000	30,000	10,000
# Unique *OD* pairs	12,977	27,357	7207

Table 3.2: Number of *OD* pairs provided by the mobility demand.

The second sampling algorithm is the radius-fixed sampling method. For each city, we generate 36 points regularly spaced on each concentric circle of various radii. The limit of 36 is inherent to the chosen sampling method. Indeed, generating 36 locations on a ring produces an angular distance between two consecutive points of 10°, since $\frac{360°}{36} = 10°$. Furthermore, if the angular distance is smaller than 10°, then the risks of noisy data in the dataset is increasing, especially for smaller radii circles [22].



(a) Rome

(b) Milan

(c) Florence

Figure 3.5: Road networks and concentric circles for the fixed-radius sampling method. The circles are centred on the city centre. (a) The rings in Rome have radii of $1km$, $2km$, $5km$ and $10km$. (b) In Milan, the radii of the circles are of $1km$, $2km$, $5km$ or $10km$. (c) Florence's circles have radii of $1km$, $2km$ and $5km$.

For each city, we try to maximise the surface covered by the samples. Hence, the different radii depend on the size of the graph to study. Road networks from Rome and Milan require four radii to generate locations widespread all over the graphs. The data are generated over concentric circles and are spaced from the centre of the city by $1km$, $2km$, $5km$, and $10km$. A representation of these circles is provided by Figure 3.5a and Figure 3.5b. As for Florence, its road network being much less extensive, the samples can only be generated on three circles with radii of $1km$, $2km$ and $5km$. The Figure 3.5c

depicts the three rings required to cover the whole surface of the road graph of Florence.

Once the individual points generated, each pair of coordinates is translated to its nearest edge. To do so, we generate the list of all the edges within a radius of $200m$ around the aforementioned location. The list is sorted and the closest edge is selected. If no edge exists in its direct vicinity, the sample is discarded. The vicinity limit is set to $200m$ to restrict the potential noise created by this translation. This limitation is also a manner to filter GPS points that are outside the road networks due to the rectangle shape of the layout of the graph as shown in Figure 3.5.

After translating the samples to edges of the road networks, the *OD* couples are generated by pairing each road segment with every other edges belonging to the exact same circle. By this process, each metropolitan area can generate several thousands *OD* pairs, the exact amount of maximal samples is given by:

$$n_p = 2n_c \binom{36}{2} \tag{3.4}$$

where n_p is the maximal number of pairs and n_c is the number of circles generating samples. The total is multiplied by 2 to take into account samples going both ways, for instance a pair from a starting point A to an ending point B and another couple going from B to A. This value is maximal and not exact. Indeed, as explained previously, some locations are discarded by the translation from latitudinal and longitudinal coordinates to edges of the road network. The results of the Equation 3.4 for each city is given in Table 3.3.

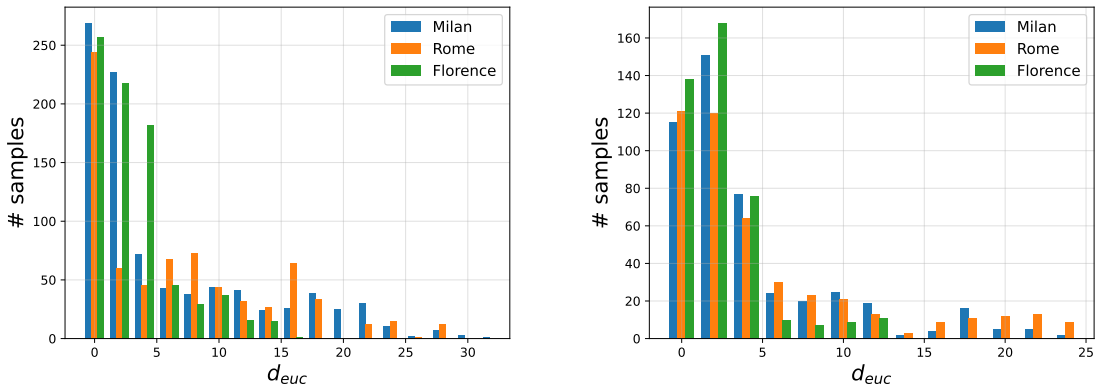
	Rome	Milan	Florence
Maximal *OD* pairs	5040	5040	3780

Table 3.3: Maximal number of *OD* pairs generated by the radius sampling method.

After generating a set of *OD* pairs through both methods, we need to randomly select a subset to deepen the analysis on. Indeed, for computational matters it is not possible to generate paths between each pair of locations. Thus, we select randomly hundreds of samples for each method to compute the diversity between these *OD* pairs. For each city, the mobility demand technique is used to select up to 1000 *OD* pairs, whereas the fixed-radius algorithm generates and pick a maximum of 500 *OD* pairs.

The gap in the number of samples is a consequence of the computational and storage challenges of this project. Indeed, the fixed-radius method requires more running time and more storage space for the paths generation. The main cause of this issue is the chosen sampling method. Indeed, the mobility demand only requires to randomly pick some samples where the fixed radius needs to generate the samples. This generation requires to sample the latitudinal and longitudinal coordinates and, then, to list and sort the nearest edges in a radius of $200m$ around the location. These closest edges need to be stored and sorted by length, thereby increasing the running time and the requires memory.

To have a better understanding of the results, we plot the number of samples for each technique by city and by distance range. Each class of distance i regroups all the OD pairs between i and $i+2$ kilometers. The Figure 3.6 shows the distribution of the samples by city and by distance for both sampling method.



(a) (b)

Figure 3.6: Count of OD pairs in each city, by distance. (a) Mobility demand selection. (b) Fixed-radius sampling.

What is noticeable on the Figure 3.6, is that the samples are slightly better distributed with the radius method (Fig. 3.6b), especially for Rome. Indeed, with the mobility demand (Fig. 3.6a), Rome has 4 times less pairs studied for a distance between $2km$ and $4km$. Furthermore, samples in Florence for the distance class of $4-6km$ are more evenly distributed with the fixed-radius sampling algorithm.

Moreover, the Figure 3.6b shows some contradiction with the sampling method. Although the fixed-radius method technically does not allow samples apart of more than $2r_{max}$, where r_{max} is the radius of the largest circle, we can note that samples exist for Rome and Milan for distance over $20km$ and over $10km$ for Florence, which respectively correspond to twice the length of the maximal radius of each city. This lengthening of the maximal distance between an origin and a destination point is due to the way the coordinates of the samples are computed. Indeed, the coordinates are computed by using the euclidean distance and the angular distance, therefore when computing the geodesic distance between the locations, a gap of few kilometers appear. The geodetic distance corresponds to the smallest distance between two points on a sphere, especially on Earth's surface. This distance along the curve of the Earth is computed by using the Haversine formula [34]. Since this formula takes into account the coordinates of the points, the distance between points can be bigger than the maximal one considering that the samples were generated with the euclidean distance and not the geodesic one.

3.2.3 Creation of the paths between the *OD* pairs

After the generation of the set of *OD* samples, we compute several sets of thousands paths for each pair by following the process detailed in Subsection 3.1.4. The paths are generated within the elliptic subgraph surrounding the origin and destination locations. The eccentricity e of the ellipse is not the optimal value enlighten by Lima et al. [1]. Indeed, for computational matter, considering an ellipse of eccentricity $e = 0.9$ almost doubles the corresponding surface to study comparing to an ellipse of eccentricity $e = 0.7$. Therefore, the chosen eccentricity of the ellipse is $e = 0.7$ as depicted on Figure 3.2.

To generate the sets of paths, we use the combination of graph randomisation and path penalisation established in Subsection 3.1.4. This process is a nested methodology, meaning that at each iteration the graph is randomised and, then, n_p path penalisation are applied on the randomised graph. At the next iteration, a new graph randomisation is computed on the already randomised and penalised graph and n_p new paths are computed. This process is repeated n_i times, therefore there are a total amount of n_i graph randomisation performed and after each one of them, n_p path penalisation are applied. Thus, the maximal number of paths that can be generated is $p_{max} = n_i n_p$, this amount of paths can be created only if every new path is not already in the set of paths.

Furthermore, to have a better overall understanding of the diversity, we generate three different sets of different size for each *OD* pair. Hence, the parameters need to be modified consequently to the maximal number of paths we want to generate for each set, as shown in Table 3.4.

	S_1	S_2	S_3
# of Graph Randomisation n_i	500	500	750
# of Path Penalisation n_p	10	20	20
Maximal # Paths p_{max}	5000	10000	15000

Table 3.4: Parameters of the graphs randomisation and paths penalisation for each set generated.

On the Table 3.4, the number of iteration n_i, the number of paths n_p and the maximal number of paths p_{max} for each one of the three different set is given.

Furthermore, the graph randomisation has two parameters to set before running the algorithm. The first one, the randomisation parameter δ allows to control how much the graphs' edges will be modified. Since we are generating large sets of paths, the value of δ can be quite high to avoid having too many duplicated paths. Indeed, with $\delta = 0.5$, we have that 68% of the edges are modified by more or less half or their initial weight and the other 32% by more than half of their weight. The graph randomisation also has a control parameter τ to avoid too small weights. In our case,

we consider $\tau = 1$, meaning that the weight of the edges are at least of 1. Indeed, with the graph randomisation and path penalisation process, having edges of weight 1 is not an issue since it will increase again later. Conversely, a weight of 0 would be problematic since both graph randomisation and path penalisation are based on the initial weight, meaning that a null weight will remain null, thereby creating biases within the sets of paths.

The different sets were computed gradually to verify if the conclusions drawn with the smaller sets were validated by sets with bigger cardinality. Furthermore, to avoid memory errors while running the analysis or overflowing data while storing, the sets can not be too large. Indeed, if each OD sample have $15,000$ paths, there are up to $1000 \times 15000 = 15,000,000$ paths generated in total for each city's largest set. In total, all the sets generated, for each city and for each sampling method, represents almost $100GB$ of data. Hence, it seems complicated to compute much larger sets or to drastically increase the number of samples studied.

After generating the sets of paths for every sample of every city, each path is analysed to define its detour category. The detour category corresponds to the excessive length when compared to the shortest path. This excess is represented by ϵ, the eponymous parameter of the *ϵ-paths* class. For instance, if the shortest path between two points A and B has a length of 1000 meters, an alternative path measuring 1234 meters is longer than the shortest one by $\epsilon = 0.234$. Thus, this path will be classified in the category of the *0.3-paths*.

In our work, the maximal ϵ considered is $\epsilon = 1$. An higher value for this detour parameter is not necessary. Indeed, considering the paths that are at least twice as long as the shortest path is counterproductive since they are in the category of unrealistic paths defined in the Subsection 2.2.1. Thus, all the paths that are more than twice as long as the smallest one are discarded.

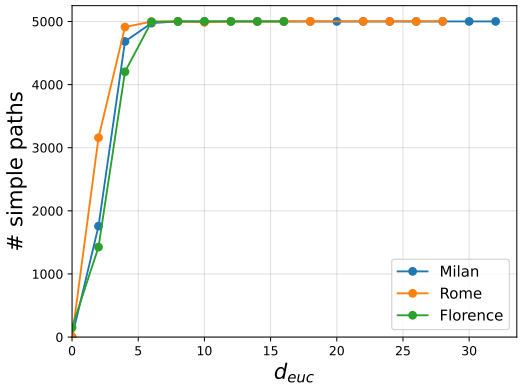
Chapter 4

Results

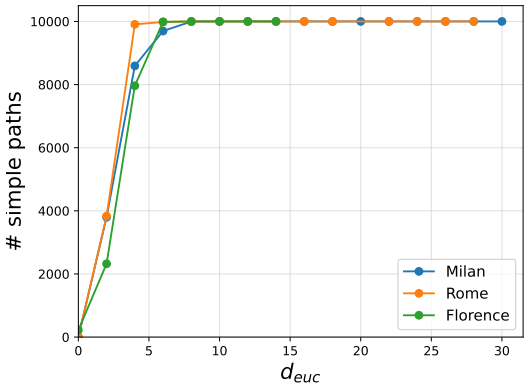
In this chapter, we present the results obtained from the experiments of paths diversification in each one of the three cities studied: Milan, Rome and Florence. In Section 4.1, we study the impact of geodesic distance on the potential diversification of paths in each city. Then, in Section 4.2, we introduce the notion of "optimal" epsilon, namely, the value of epsilon providing a certain minimal diversification threshold, and, we use it to compare the posible diversification between our three sities. Finally, in Section 4.3, we analyse how the diversification of paths changes with respect to the core-periphery structure of the road network.

4.1 Impact of geodesic distance

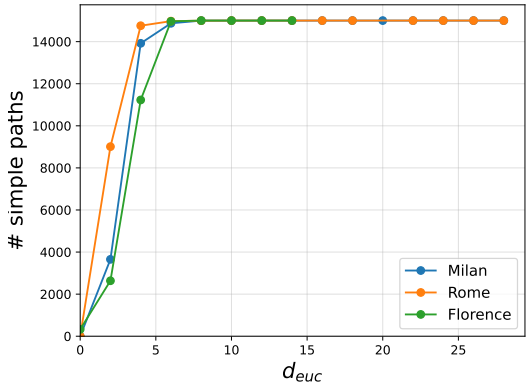
When assessing the diversity of itineraries in an urban area, the first parameter that needs to be discuss is the geodesic distance between the origin and the destination of a trip. This distance can be considered as the straight line distance on sphere surfaces, especially on Earth's surface. It is computed by using the haversine formula [34]. In the rest of our work, this distance will be denoted by d_{euc}. Understanding if the number of routes generated is limited after a certain distance is an important point to apprehend the diversity of cities. Therefore, to study the distance aspect of the diversity, we investigate the number of paths generated against their distance, on Figure 4.1. The average number of paths is done on ranges of distance of $2km$, for instance the value on distance $0km$ is the average on all the origin-destination pairs that are apart of 0-$2km$, and so on for each distance range.



(a) S_1



(b) S_2



(c) S_3

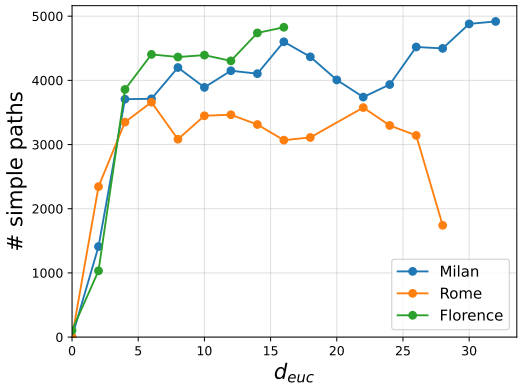
Figure 4.1: Average number of simple paths by distance for each set of paths. (a) S_1, up to 5000 paths generated.(b) S_2, up to 10,000 paths generated.(c) S_3, up to 15,000 paths generated.

What is striking on Figure 4.1 is that the curves for all the cities are collapsing for each set when the distance reaches $d_{euc} = 6km$. The second observation that can be made is that the global shapes of the curves are similar for each set, the only difference being the magnitude of the shape, 5, 10 and 15 thousands respectively on Figures 4.1a, 4.1b and 4.1c. Moreover, each curve becomes flat from the category $6km$, meaning that they reached the maximal value of paths generated. Therefore, if two points are apart of more than $4km$, the generation algorithm produced different paths at every iteration.

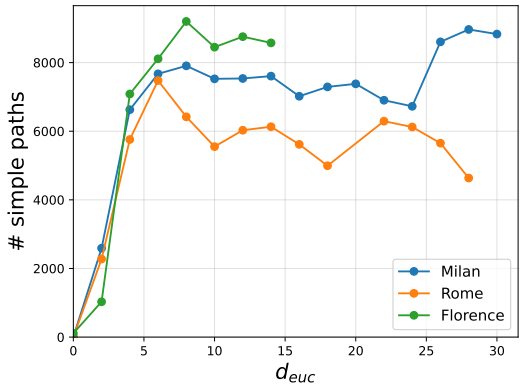
The first conclusion than can be drawn from these results, partially confirmed our first intuition that the further the origin is to the destination, the more simple paths there are in between. However, since the samples reach the maximal value for medium distances, it is not possible to fully confirmed that there does not exist a threshold at which the number of paths stagnate.

However, these analysis consider all the paths generated, especially some paths that are not realistic (e.g., they are unrealistically long), according to the definition given in

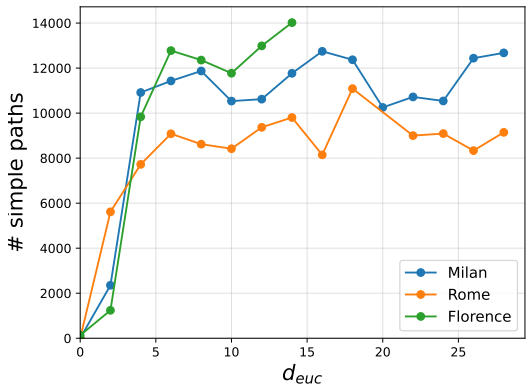
Subsection 2.2.1. Therefore, to have an actual understanding of the number of feasible paths generated, we need to discard the simple paths that are too long compared to the shortest path. The threshold of the epsilon simple paths is fixed at $\epsilon = 1$, that is to say twice the length of the shortest path. The average number of paths that are shorter than twice the length of the shortest one, is presented on Figure 4.2.



(a) S_1



(b) S_2



(c) S_3

Figure 4.2: Average number of epsilon simple paths by distance for each set of paths. (a) S_1, up to 5000 paths generated.(b) S_2, up to 10,000 paths generated.(c) S_3, up to 15,000 paths generated.

Unlike the previous results, Figure 4.2 shows significantly different results for each city. Indeed, the average number of epsilon simple paths in Florence and in Milan are higher than in Rome. This difference could be caused by the higher density of edges in Florence's and Milan's graph than in Rome's network. Indeed, if we consider the edge density as the number of vertices by squared kilometer, we obtain:

$$d_R = \frac{|E_R|}{\mathcal{S}_R} = \frac{31798}{500} = 63.596 \tag{4.1}$$

$$d_M = \frac{|E_M|}{\mathcal{S}_M} = \frac{24063}{350} = 68.751 \tag{4.2}$$

$$d_F = \frac{|E_F|}{\mathcal{S}_F} = \frac{6140}{90} = 68.222 \tag{4.3}$$

where d_i is the edge density of the city i, for $i \in [R, M, F]$ respectively corresponding to Rome, Milan and Florence, E_i its set of edges and $\mathcal{S}_i$ the surface represented by the graph.

The other interesting thing we can note is that the curves of epsilon simple paths seems to flatten after a certain distance, around $d_{euc} = 6km$, the only exception being the distance range of $30km$. However, as shown by Figure 3.6a in Subsection 3.2.2, the number of samples available for this range of distance does not allow to have reliable results for distance over $30km$. However, this flattening phenomenon can be partially explained by the upper bound of the sets of paths, as shown on Figure 4.1.

This analysis can be deepen by studying the evolution of the number of epsilon simple paths for every value of epsilon from $\epsilon = 0.1$ to $\epsilon = 1$. This generalisation of the process is represented on Figure 4.3. On this figure, the set of paths represented is only S_3, with a maximal number of paths of 15,000. Indeed, showing the results obtained for each set, would be redundant since the plots are almost identical for each different set.

Figure 4.3: Average number of epsilon simple paths by distance for each value of epsilon.

What is noticeable on Figure 4.3, is that for $\epsilon = 0.1$ and $\epsilon = 0.2$, Florence has a much higher average number of paths than Milan and Rome, especially for medium distances included in the range 4-10km. Therefore, Florence offers more near shortest alternative paths for short-medium trips. More generally, when the epsilon is smaller than 0.5, there exists a significant difference between Rome and the two other cities. The small epsilons sets of paths represent routes that are closer to the length of the shortest path. For instance, for $\epsilon = 0.5$ all the paths in the set are at most 50% longer than the shortest path. Therefore, in Rome the number of near-shortest paths is substantially lower than Milan and Florence, meaning that in Rome following alternative routes might lead to a significant lengthening of the trip. Indeed, Rome offers very few routing choices close to the length of the shortest path for medium distance trips.

However, when epsilon increases, the difference between the three cities proportionally reduces considerably. To validate this, the subplots can be reproduced with a y-axis representing the percentage of all the paths generated for each distance instead of the raw data. This new representation is given by Figure 4.4. As we can see, the shapes of the curves are quite similar to the ones from Figure 4.3 which can be explained by the fact that the values are divided by the average number of paths generated, given by Figure 4.1c. The number of all simple paths being almost identical, it explains the similarity between Figures 4.3 and 4.4.

Figure 4.4: Proportion of all the simple paths generated that belongs to each epsilon simple paths category.

As expected, Figure 4.4 corroborate the fact that when the epsilon is increasing, the relation between the number of paths decreases. Indeed, if we consider the distance

4 kilometers, for $\epsilon = 0.1$, Florence has $\frac{1.2}{0.2} = 6$ times more paths from the whole set of paths, whereas for $\epsilon = 1$, it drops to $\frac{90}{50} = 1.8$. This decrease of the gap between Florence's and Rome's curves might be due to the fact that for higher epsilons, Rome has more new simple paths. However, this hypothesis needs to be further analysed to be either validate or refuted.

To pursue this new analysis, we need to explore the non cumulative number of epsilon paths. So far, every set of epsilon paths gathers all the paths whose length are not exceeding the length of the shortest path by $(1 + \epsilon)$ times its length. To understand the number of new paths for each category, what needs to be studied is the sets of epsilon paths gathering all the paths whose length are between $1 + (\epsilon - 0.1)$ and $1 + \epsilon$. Thus, the set of 0.2-simple paths includes the paths measuring between 1.1 and 1.2 times the length of the shortest path. After, constructing the new sets of paths, we generate the charts for each value of epsilon. This plots are given on Figure 4.5.

Figure 4.5: Average number of new epsilons simple paths by distance.

Figure 4.5 shows that, from $\epsilon = 0.7$, Rome has more new paths than Milan and Florence for small distances. Thus, those new paths are slightly compensating the gap generated by smaller epsilons. In summary, what we can conclude from these charts is that Rome has a huge lack of diversity for near shortest paths when it concerns medium distance trips, whereas for short distance trips it offers more diversity than Milan and Florence. Conversely, if we consider paths that are twice as long as the shortest one for the same distance range, Rome offers more alternative routes than the two other cities, and that for almost all the distance range.

Finally, to summarise our discoveries on the impact of the distance on the existing paths diversity between an origin-destination pair, it would not be very rigorous to affirm that there is a constant increase of the number of paths between two locations

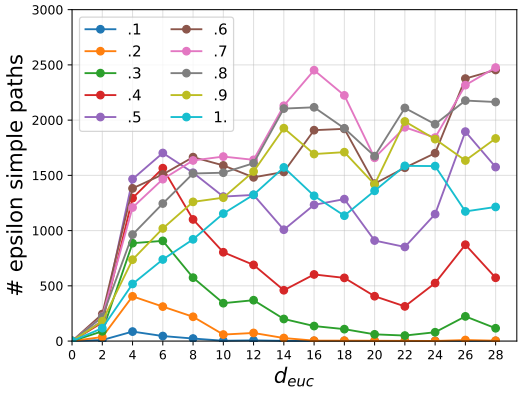
when they are moving further. Indeed, each one of our three sets reach the maximal number of paths even for small distances such as $4km$. Furthermore, when it comes to the epsilons simple paths, the results show stable results with the increase of the distance from the same range of $4km$. This flattening of the curves at the same range is too coincidental not to be a consequences of the sets' limitations. Therefore, our conclusions will not be focus strictly on the impact of the distance on diversity in on city but on the comparison between them.

The most relevant observation from all the aforementioned figures is that Rome presents a lack of epsilon simple paths, whether it is for the cumulative sets of epsilon paths or for the sets of new epsilon paths, except for the *1-epsilon new simple paths*. This gap is proportionally even more striking for nearest shortest paths, since Rome offers five to six times less alternative paths for $\epsilon = 0.1$ or $\epsilon = 0.2$. Yet, Rome has very similar initial sets of paths and this gap only appears when considering the epsilon paths. Hence, there are much less realistic paths in Rome than in Milan and Florence, especially for short paths. The conclusion that can be drawn from all this observations is that Rome has a clear lack of diversity in its road network in comparison to Milan's and Florence's ones.

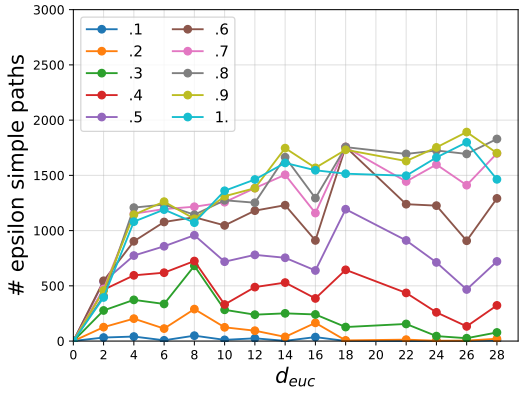
4.2 Optimal epsilon

Furthermore, when looking closely to Figure 4.5, we can note that there is a decrease of the number of new paths for each city after a certain epsilon, around $\epsilon = 0.6$. Indeed, in Florence, when $\epsilon = 0.5$, there are up to 3000 new paths, whereas for $\epsilon = 1$, the maximal values of Florence's curve is only of 1000. However, the epsilon producing the most new paths is different for each city. Therefore, studying the value of this "optimal" epsilon could be enlightening to understand the diversity of paths in urban area.

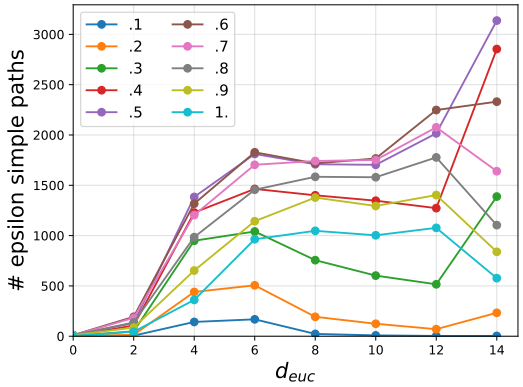
The first definition of the "optimal" epsilon that can be considered is the value of epsilon providing the biggest amount of new paths by distance range for each city. To have a better understanding of this epsilon, it is necessary to project all the curves for each epsilon by city on the same plot to compare the epsilon new paths in one city at the time, see Figure 4.6.



(a) Milan.



(b) Rome.



(c) Florence.

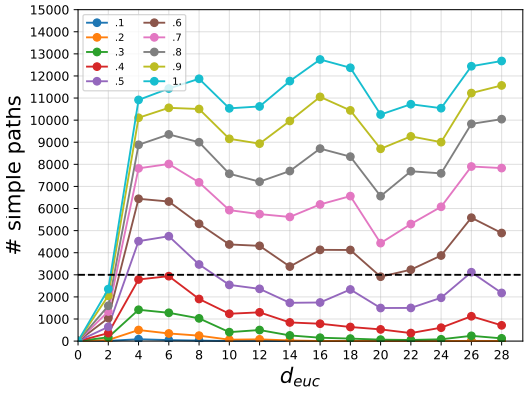
Figure 4.6: Average number of new epsilon simple paths by city.

If we consider the aforementioned definition of an optimal epsilon, that is to say the epsilon providing the biggest amount of new epsilon paths. Milan reaches its optimality at $\epsilon = 0.7$, Rome at $\epsilon = 0.9$ and Florence has an optimal epsilon of $\epsilon = 0.6$, as shown on Figure 4.6. This means that, for instance, in Milan, for $\epsilon > 0.6$, the number of new paths will decrease, meaning that the peak diversification has already been reached. This decrease might seem counter-intuitive but since we are considering the elliptic subgraph around the *OD* pair, a part of the realistic paths, in terms of length, are not considered because diverging too much.

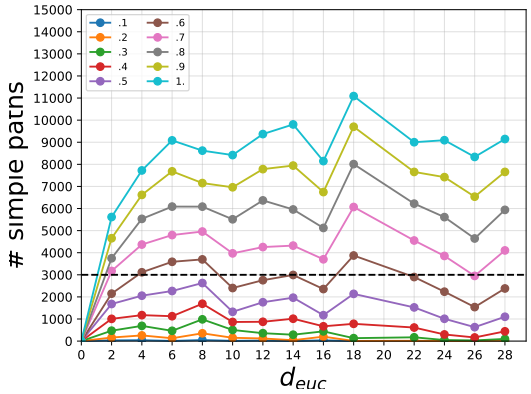
However, if we transpose these values to the cumulative sets of paths, namely the number of paths measuring at most $(1 + \epsilon)$ the length of shortest one (Fig. 4.3), we obtain sets of paths for Milan, Rome and Florence having for respective cardinal 6000, 7500 and 8000 paths. Therefore, this definition of an optimal epsilon does not represent a fix proportion of paths. For instance, Milan has an "optimal set" that represents 40% of the total paths generates, whereas in Florence it represents more than 50%. However, these sets are not only unstable but also too large, since half of the paths are included in them.

Thus, considering the set of epsilon paths with the highest cardinal as optimal might

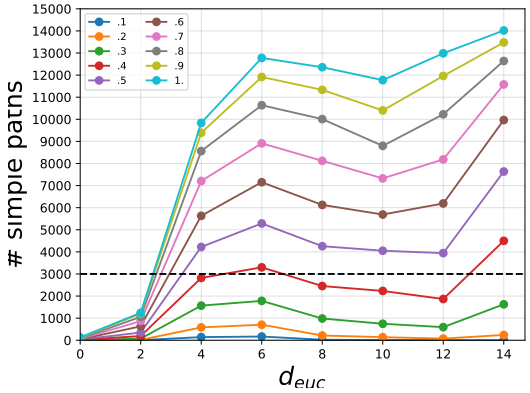
not be the most realistic definition of the notion of "optimal" epsilon. To palliate these problems, the process can be inverted. Instead of computing the proportion from the epsilon, the optimal epsilon can be determined from an arbitrary threshold. For instance, we can fix a threshold $\tau = 0.2$, meaning that the set will regroup the top 20% of shortest paths. An illustration of this this threshold is provided by Figure 4.7.



(a) Milan.



(b) Rome.



(c) Florence.

Figure 4.7: Average number of epsilon simple paths by city.

In comparison to the values given by the previous definition and Figure 4.6, all the cities have lower "optimal" epsilon. Indeed, Milan's optimal epsilon decreased from $\epsilon = 0.7$ to 0.5-0.6, in Rome it went from $\epsilon = 0.9$ to $\epsilon = 0.6$ and in Florence the optimal epsilon is now of $\epsilon = 0.4$.

The value of this optimal epsilon represents the extensions of the shortest path that are required to finally obtained sets of short paths with similar sizes. The fact that Rome has an higher value of "optimal" epsilon means that there are less near shortest paths and therefore less potential diversification for near shortest paths. Thus, this value can be considered as a comparison tool of potential diversification between different cities. In our work, the comparison of this value between the different cities, corroborate the global observations made in Section 4.1. That is to say that Rome has

more potential route diversification for short trips (i.e., d_euc = 0-2km), since Milan and Florence does not even reach the threshold for this distance range. On the other side, for medium and long distance, Rome has an higher optimal epsilon, hence, it offers less alternative paths for longer trips.

4.3 Core-periphery structure and paths diversification

In this Section, we will explore the relation between the core-periphery structure of a city and the potential of diversification of this city. The core-periphery structure of a network [33] is a type of graph having a group of nodes densely connected, the *core*, and sparser nodes, the *periphery*. Unlike other densely connected group of nodes, such as the nodes communities, the connection between the node and the periphery is more developed. In particular, road networks have a core-periphery structure since the city center tend to be denser than the peripheries.

To asses this, the sets of paths for each city has been generated using the fixed-radius sampling method, introduced in Subsection 2.1.4 and detailed in Subsection 3.1.2. This sampling is not only a way to better distribute the data in the road network but it also provides some additional information on the *OD* pairs to locate them in the city structure. Indeed, with this technique we can access to the radius on which the locations were generated and the angular distance, θ, formed by both locations of the pair and the central point of the city. By combining all of this information, we can easily determine if the pair is located in the city centre, the inner suburbia or more distant suburbs and if they belong to the same side of the city or on opposite sides, meaning that the shortest path will pass through the city centre.

The results described in this Section of our work, will focus mainly on the comparison of road structure between Milan and Rome. Indeed, Florence's road graph being much narrower, the different suburbs are not in the same range of distance, making it more difficult to compare. However, Florence still has some very interesting results, especially for the diversification of its city centre, as shown on Figure 4.8.

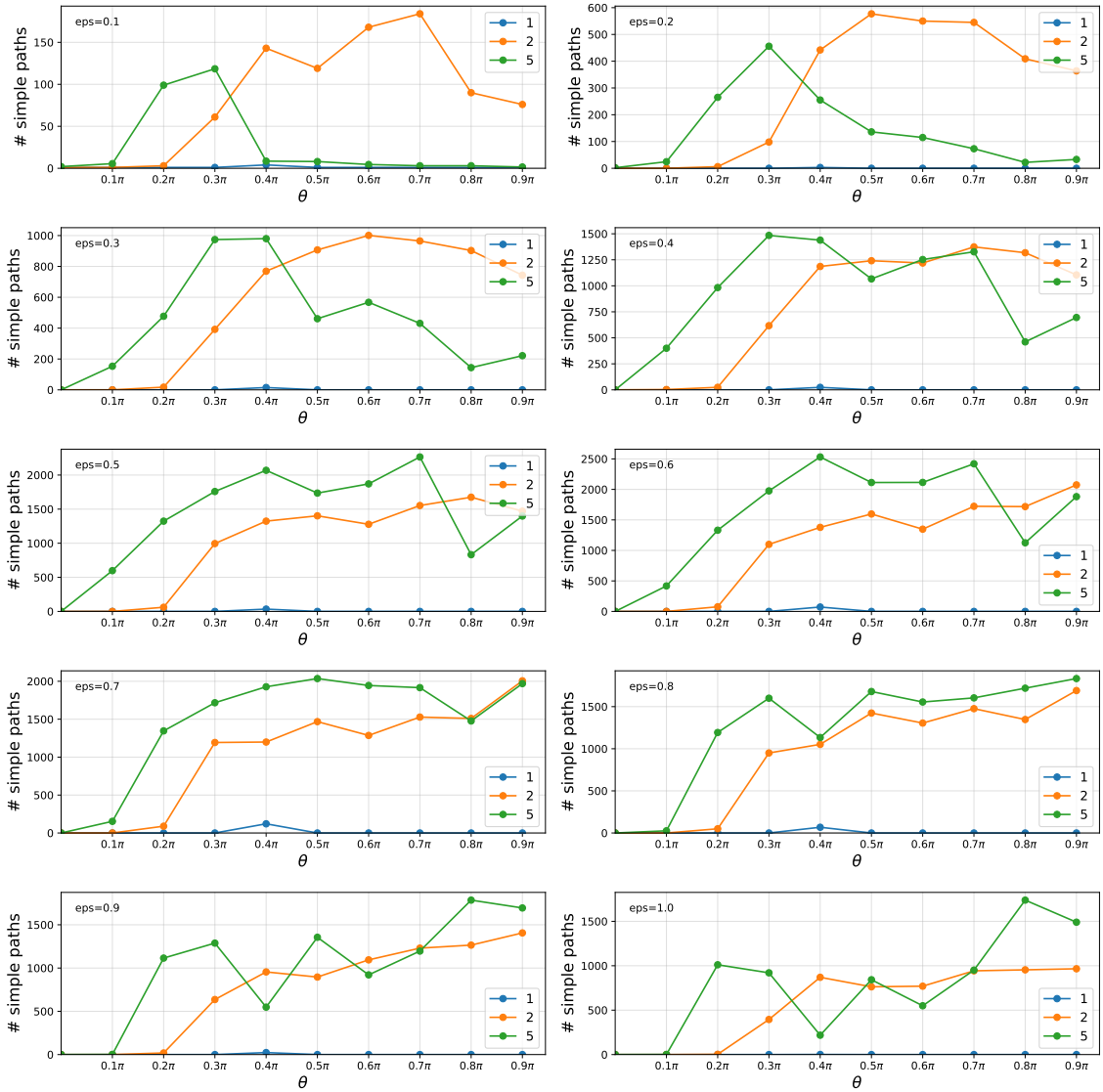


Figure 4.8: Average number of epsilons simple paths by angular distance, θ, for each radius in Florence.

The first noticeable on Figure 4.8, is that in Florence, the locations sampled on the circle of radius $1km$ have on average an almost empty set, and that regardless of the value of epsilon. This lack of paths can be explained by the fact that most of the historical city centre of Florence is a pedestrian area or a *ZTL* (*Zona a Traffico Limitato*), meaning that the traffic is restricted only to residents, taxis and buses in this area. This restricted area can easily be seen on Figure 3.5c. Indeed, on this map of Florence, the smallest circle drawn has a radius of $1km$. It appears clearly that the inside of this ring has a large area with no edges, whereas the maps of Milan (Fig. 3.5b and of Rome (Fig. 3.5a indicate much denser city centre in term of edge density.

However, the lack of paths diversity in Florence's city centre is not really the result of the core-periphery structure of the urban area in its strict sense. It is actually the

consequence of a topological choice, restricted the historical city centre to a minimum car usage to limit the pollution within the centre.

To study more precisely the impact of the structure in the potential diversification of itineraries in a city, we compare the results obtained by the radial paths generation in Milan and Rome. The results are respectively given by Figure 4.9 and 4.10.

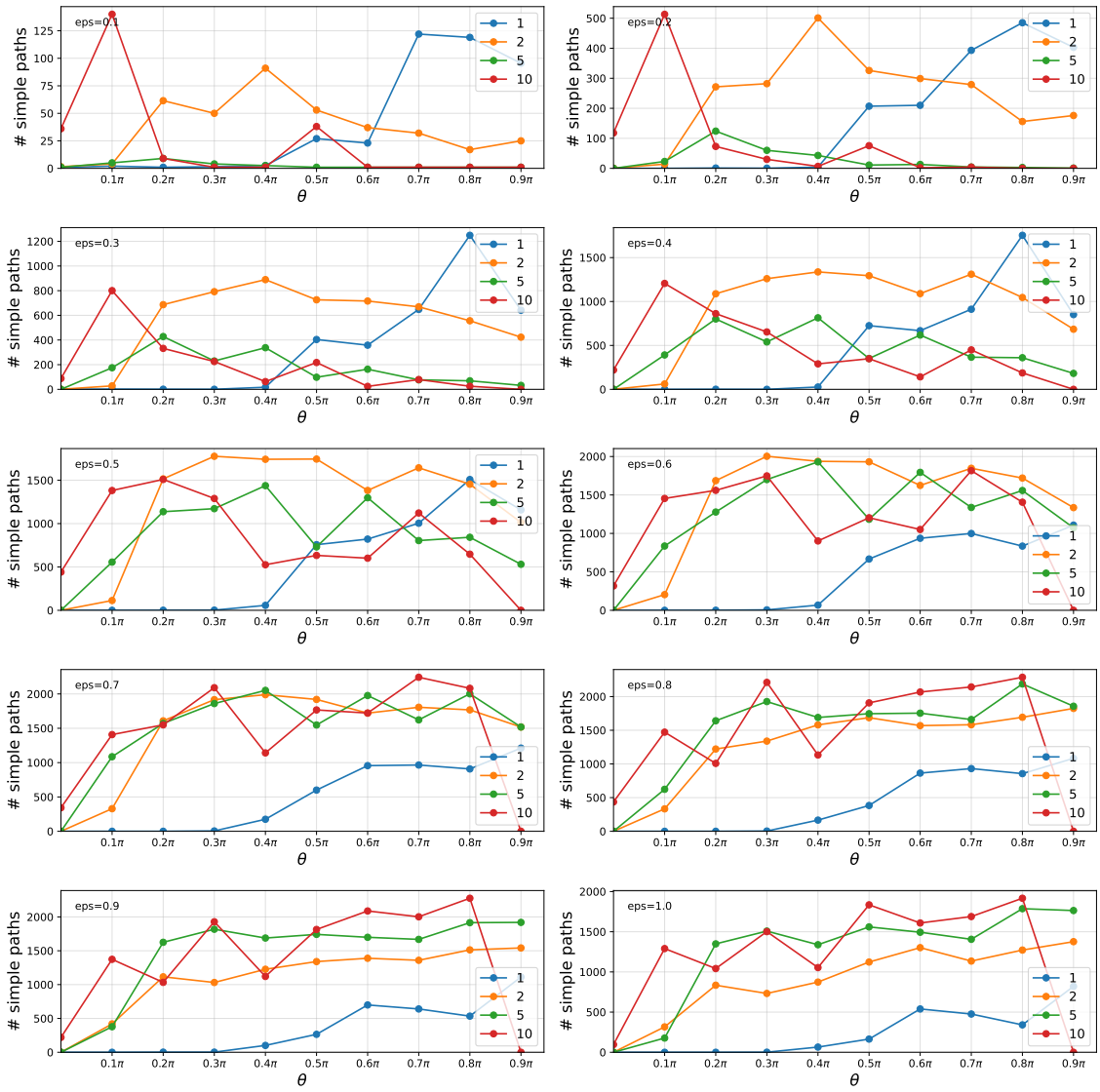


Figure 4.9: Average number of epsilons simple paths by angular distance, θ, for each radius in Milan.

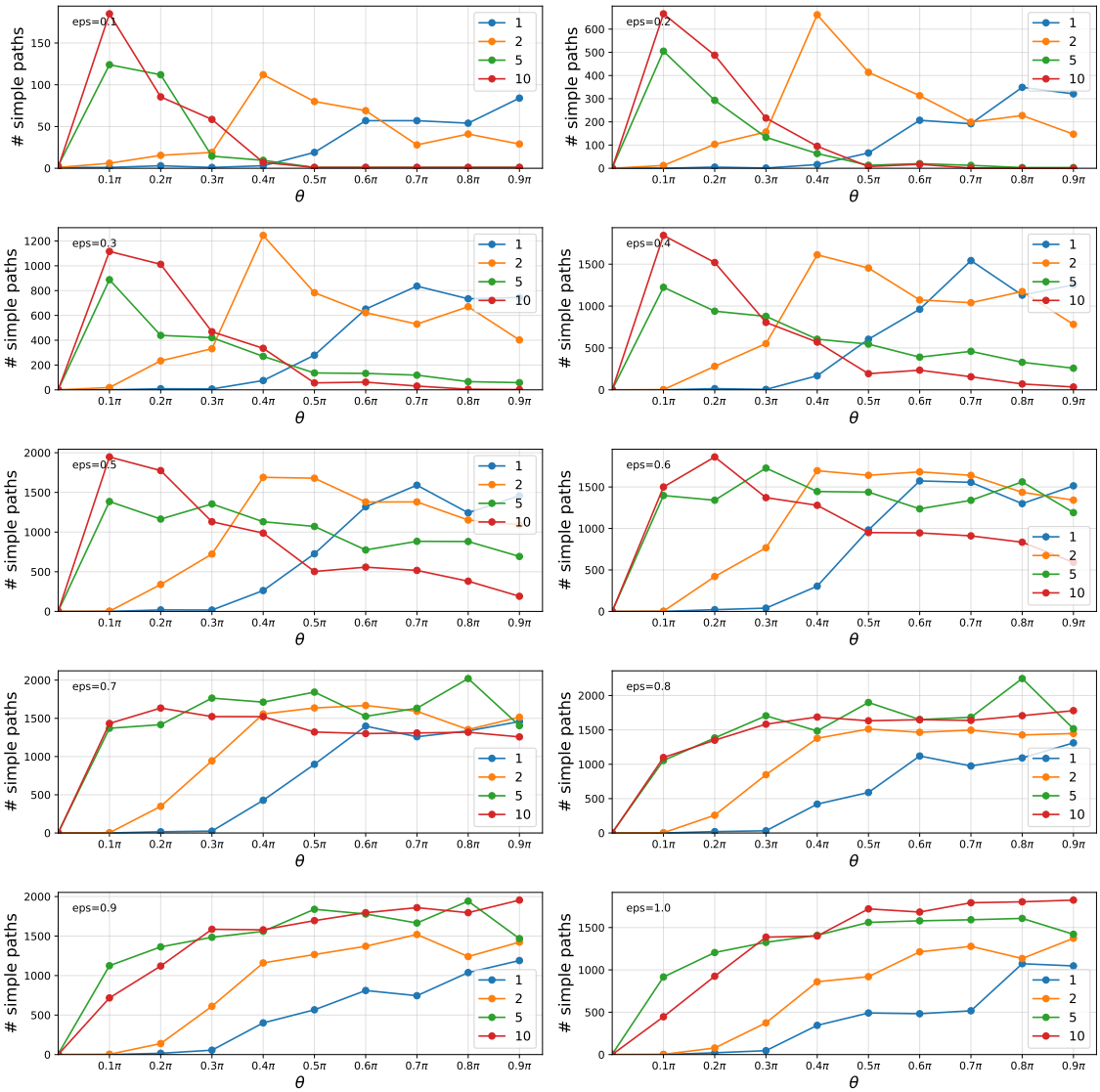


Figure 4.10: Average number of epsilons simple paths by angular distance, θ, for each radius in Rome.

Before comparing the results obtained in both cities, we can focus on the singularities of Milan's curve. First, when the epsilon reaches at least 0.5-0.6, we can notice that the samples on the $10km$ radius curve drop significantly for $\theta = 0.9\pi$. This results from the noise generated by the shaped of the road network. Indeed, when we analyse the road graph and the sampling circles, on Figure 3.5a, the wider ring is partially out of the covered area to the east and south. Therefore, the number of OD pairs on the circle of $10km$ having and angular distance between 0.9π and π, namely on opposite sides of the ring, is very low, thereby making unstable final results. Indeed, since half the circle does not cover the road graph, no samples were generated there.

The second noticeable characteristic of Milan is that for $\epsilon = 0.1$, samples on circle $5km$ have almost a null potential diversification. Indeed, the corresponding curve on

Figure 4.9 is almost flat with a maximal value around 10. However, if we compare this specificity with Rome and Florence, we notice that for the $5km$ curve, they have a significant jump of value between $\theta = 0.1\pi$ and $\theta = 0.4\pi$ before decreasing to an almost null value. This can be explained by the fact that there are less simple paths for small epsilons when the distance increase. Thus, since an increase of the angle also means an increase of the distant, this flattening of the curve is explainable. Yet, for near shortest paths in the periphery ($radius = 5km$), Milan shows very low potential diversification, unlike Rome and Florence. A potential explanation of this phenomenon is that for $\epsilon = 0.1$, the threshold is too strict in Milan and, therefore, there are no near shortest paths.

When deepening this comparative analysis between Milan and Rome, the first thing noticeable is that there is a big difference for samples on the circle of radius $1km$ when $\epsilon = 0.1$ and $\epsilon = 0.2$. Indeed, in Milan, the number of epsilon paths for OD pairs on the $1km$ radius circle is at most 125 for $\epsilon = 0.1$ and 600 when $\epsilon = 0.2$, where in Rome these values only reach 75 and 400. This gap happens mainly when the angle measures between 0.7π and 0.9π, meaning that the points of the OD pairs are located almost on opposite sides of the city centre. Thus, Milan offers more short paths diversification in its city centre than Rome. This result corroborate the observations and hypothesis made in Section 4.2 that Rome is less diversified for near shortest paths.

The main source of information given by these figures is the peak angle, that is to say the angle at which the curve reaches its maximum increase. For instance, in Rome for $\epsilon = 1.0$, the curve corresponding to the circle of radius $10km$ has a peak angle of $\theta = 0.1\pi$ because it goes from almost 0 paths to 12,000. If we compare closely each curve in Milan and in Rome, we notice that, except the aforementioned near shortest peripheral trips singularity in Milan, every one of them have almost similar shapes and magnitude, especially for high epsilons.

The major difference between them is this peak angle. If we focus on peripheral samples, namely the OD pairs sampled on circles are radius 5-$10km$, the peak angle is always of 0.1π-0.2π regardless of the value of epsilon. Thus, there is the same potential diversification in the suburbs whether it is a closer one or a more distant, otherwise the results for different radii would not have been that similar.

On the other hand, for the city centre ($radius = 1km$), the peak angle is around 0.5π-0.6π. The difference between peripheral area and city centre is mainly due to the scale of the subgraphs studied. Indeed, for small angle two consecutive samples on the $10km$ circle are 10 times further than two successive locations on a ring of radius $1km$. Thus, the gap between peak angle in suburbs and in the city centre does not necessarily result from an actual difference of potential diversification but from the range of distances.

Yet, there is one major difference between Milan and Rome when we compare these peak angles for each curve. The value of the peak angle for the $2km$ radius curve is of 0.2π in Milan and of 0.4π in Rome. This difference leads the curve to be very similar to the peripheral ones in Milan, whereas in Rome, the peak angle really happens between the peripheral one and the city centre one. The explanation of these different peak angle lies in the structure of both cities, especially the spread of the historic city centre of each city. Indeed, Rome's historic city centre, that is to say the inside of the walls

of the city, expands on $14km^2$. This surface correspond to a area of a circle of radius $2km$. Conversely, Milan's historic centre covers a much smaller area of less than $1km^2$. Therefore, the circle of radius $2km$ represents the periphery in Milan and the limit between city centre and periphery in Rome, thereby changing the peak angle.

Finally, we can conclude by saying that the peripheries have quite similar potential diversification regardless of their distance from the city centre. Furthermore, even though the distance impact the diversification as explained in Section 4.1, it is not the only parameter to take into account. The structure of cities have a major role in how much it is possible to diversify the paths in a city. For instance, Rome's historical centre being wider than Milan's one also affects the potential diversification of paths. Indeed, in peripheries, there are usually orbital roads to drive around cities or to go from a suburb to another one. However, city centres do not offer roads for this specific purposes. That is why city centres have there peak angle when trips start to pass through the central point of the city.

A potential explanation of the diversification problems in Rome lies in its unique status of "eternal city". Indeed, Rome is one of the most important city in European and World culture and its historical centre is often compared to an open-air museum. This particularity gave to Rome's centre a UNESCO World Heritage title, preventing major modifications in it. Therefore, Rome's road network was not designed to be efficient but was build all around the archaeological monuments, forcing drivers to take lengthening detours. The same phenomenon is exacerbated in Florence. Most of its historical centre is a pedestrian area and, therefore, completely closed to car. Hence, the results in Florence for diversification within the city centre are much lower than in Milan.

Chapter 5

Conclusion

In this thesis, we develop a procedure to assess the route diversity in urban environments. Our aim is to understand the potential paths diversification of cities and how the number of simple paths changes according to different factors such as its extension from the shortest path, the geodesic distance between the starting and ending point and its location within the metropolis structure.

The assessment process starts by sampling origin-destination pairs throughout the urban area to represent all the possible trips within the city. To do so, we use two different sampling methods, the mobility demand one and the fixed-radius sampling algorithm. The first one is based on real travel requests, whereas the second one consider only trips starting and ending on the same circle. By applying our procedure to two different method, we benefit from advantages of both techniques. We obtain results that represents well real everyday trips and we try to understand the impact of the city's road network organisation.

After gathering all the *OD* pairs from both sampling algorithms on which we will explore the diversity, we generate the corresponding elliptic subgraph for each pair. This reduction allows us to obtain a better representation of realistic paths and to significantly reduce the complexity, the running time and the memory cost of the procedure.

Then, we developed a paths generation algorithm, based on alternative routing algorithms, that can meet our need of generating sets of thousands of paths with a good trade off between overall diversity and length of the path. We build this algorithm by combining two pre-existing routing algorithms, the graph randomisation and the paths penalisation.

Once the sets generated for all the pairs of locations, we analyse the diversity by comparing the cardinal of these sets according to divers parameters. (i) We analyse the impact of the distance on the number of paths between two locations; (ii) we explore the notion of "optimal epsilon", to understand if there is threshold after which the possible diversification is guaranteed and; (iii) we analyse the impact of the core-periphery structure of a town on diversity by comparing the results obtained with the fixed-radius algorithm and its inherent advantages to differentiate short trips in the city centre and in the periphery.

The results of this thesis clearly indicates that there is a lack of potential routes diversification in Rome comparing to Milan and Florence, for the medium and long trips (i.e., $d_{euc} > 4$km). Conversely, for small distance trips, Rome offers more alternative paths. Furthermore, our work also highlight the impact of the core-periphery structure in the diversity of a city. Indeed, Rome's center being much wider than Milan's and Florence's ones impacts its diversity. In this work, we also provide some leads to understand and assess the efficiency of a city towards congestion from a topological aspect.

Our analysis of the potential diversification of paths in a city can be improved in different ways. The first possible improvement would be to reduce the potential noise existing in our analysis. To do so, it could be enriching to try other sampling methods, especially a technique that could palliate the lack of representation of long distance to get an even distribution of *OD* pairs by distance. This could be done by increasing the number of samples on larger circles of the fixed-radius method, by purposefully selecting more long distance pairs in mobility demand or with a completely new sampling algorithm.

Another potential deepening of our work could be to increase the number and the diversity of the city we studied, as it was done for the detour index [21]. Indeed, comparing cities of different countries and continents could really improve the understanding we have of diversification by giving new perspectives. All the cities we studied are ancient European cities that had an important part in the Italian and European history. The comparison with recent cities whose road network were conscientiously designed, such as cities in the US, could be beneficial to have an overview of the problems of paths diversity in urban environments. The main development of Northern American cities is much more recent then European ones and their road network have been designed alongside the modern architecture and the creation of buildings and skyscrapers. Hence, the road network in North America were purposefully designed as grids. Evaluating if this choice offers more diversity than less organised road networks can help understand the characteristics of an "optimal" road network.

Furthermore, our diversity analysis could be a starting point for some future works linking this topological analysis with actual road usage. In Section 2.2.2, we presented an article [38] studying both at the same time by classifying edges according to their K_{road} and their betweenness centrality b_c. In their work, they highlighted that most of the congestion was caused by only a few number of drivers' sources. A potential new research subject could be to identify these sources and compare their diversification with the other area of the cities. This new project could aim to understand if the congestion of our cities is partially caused by a lack of alternative paths for the users from this sources. Indeed, if the sources creating the main part of congestion offer less paths choices, the drivers will all use the same few edges increasing the risks of congestion on these road segments.

Another possible extension of our work would be to explore the relation between the "optimal" epsilon and the routing algorithms to exploit the global potential of diversification. What we observed in this thesis is that some cities needs to consider longer paths to reach a similar diversity. For instance, in Rome if only the nearest

shortest paths, that is to say the epsilon $\epsilon = 0.1$ simple paths, there are only very few different alternative routes and the drivers all have to use the same paths. Conversely, in other cities as Milan the diversification for those paths might be enough and does not require the exploration of longer alternative. Thus, it could be interesting to try to derive algorithms' parameters from the value of this epsilon or to try to develop brand new routing algorithms using this extension parameter.

5.1 Acknowledgments

This thesis was for me a first experience in the research world and I would like to thank my supervisors, Luca Pappalardo and Giuliano Cornacchia, for the help and availability from the thesis request to the thesis submission. Their suggestions have been beneficial and decisive in my journey at the Università di Pisa.

This work marks the ending point of my studies, started in 2018 in Paris and finishing in Pisa five years later. This double degree's opportunity between Université Paris-Dauphine and Università di Pisa deeply changed me, on the professional and personal aspect. Therefore, I want to acknowledge both universities and all the people that guided me in this process.

Particularly, I would like to thank Gabriella Pigozzi, Associate Professor in Computer Science at Université Paris-Dauphine, for the kindness she demonstrated during the lockdown and the remote classes. She took a determinant part in my application at this double degree and I am very thankful for this.

Finally, I am very grateful of my whole family, for their love and their support during the redaction of this thesis. In particular, I would like to thank my brother, Antoine, for reading, correcting and giving an external opinion on this work.

Bibliography

[1] Lima A et al. "Understanding individual routing behaviour". In: *Journal of the Royal Society Interface* 13 (2016). DOI: 10.1098/rsif.2016.0021.

[2] Dragan Amić and Nenad Trinajstić. "On the detour matrix". In: *Croatica Chemica Acta* 68.1 (1995), pp. 53–62.

[3] Matthew Barth and Kanok Boriboonsomsin. "Traffic congestion and greenhouse gases". In: *Access Magazine* 1.35 (2009), pp. 2–9.

[4] Yoann Bernard et al. "Impacts of the Paris low-emission zone and implications for other cities". In: *TRUE–The Real Urban Emissions Initiative, United Kingdom, ISBN* 1744072 (2020).

[5] M Bohm, Mirco Nanni, and Luca Pappalardo. "Quantifying the presence of air pollutants over a road network in high spatio-temporal resolution". In: *Climate Change AI, NeurIPS Workshop*. 2021.

[6] Matteo Böhm, Mirco Nanni, and Luca Pappalardo. *Improving vehicles' emissions reduction policies by targeting gross polluters*. 2022.

[7] Ulrik Brandes. "On variants of shortest-path betweenness centrality and their generic computation". In: *Social Networks* 30.2 (2008), pp. 136–145. DOI: https://doi.org/10.1016/j.socnet.2007.11.001.

[8] Dan Cheng et al. "Shortest-path diversification through network penalization: A washington DC area case study". In: *Proceedings of the 12th ACM SIGSPATIAL International Workshop on Computational Transportation Science*. 2019, pp. 1–10.

[9] Giuliano Cornacchia, Mirco Nanni, and Luca Pappalardo. "One-Shot Traffic Assignment with Forward-Looking Penalization". In: *ArXiv* abs/2306.13704 (2023). DOI: 10.48550/arXiv.2306.13704.

[10] Giuliano Cornacchia et al. "How routing strategies impact urban emissions". In: *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. Association for Computing Machinery, 2022, pp. 1–4. ISBN: 9781450395298. DOI: 10.1145/3557915.3560977.

[11] Giuliano Cornacchia et al. "Navigation Apps and Urban Sustainability". In: *Fluctuation and Noise Letters* (2023). DOI: 10.1142/S0219477524500160.

[12] Giuliano Cornacchia et al. "The Effects of Route Randomization on Urban Emissions". In: vol. 4. June 2023, pp. 75–87. DOI: 10.52825/scp.v4i.217.

[13] Edsger W. Dijkstra. "A note on two problems in connexion with graphs". In: *Numerische Mathematik* 1 (1959), pp. 269–271. DOI: 10.1007/BF01386390.

[14] Friedman EJ. "A Generic Analysis of Selfish Routing". In: (Jan. 2002). DOI: 10.1.1.619.3803.

[15] David Eppstein. "Finding the k Shortest Paths". In: *SIAM Journal on Computing* 28.2 (1998), pp. 652–673. DOI: 10.1137/S0097539795290477.

[16] Linton C Freeman. "A set of measures of centrality based on betweenness". In: *Sociometry* (1977), pp. 35–41. DOI: 10.2307/3033543.

[17] Conor K Gately et al. "Urban emissions hotspots: Quantifying vehicle congestion and air pollution using mobile phone GPS data". In: *Environmental pollution* 229 (2017), pp. 496–504. ISSN: 0269-7491. DOI: https://doi.org/10.1016/j.envpol.2017.05.091.

[18] Ante Graovac and Tomaž Pisanski. "On the Wiener index of a graph". In: *Journal of mathematical chemistry* 8.1 (1991), pp. 53–62.

[19] Christian Häcker et al. "Most diverse near-shortest paths". In: *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. 2021, pp. 229–239.

[20] T. Kihara et al. "Fast link-disjoint path algorithm on parallel reconfigurable processor DAPDNA-2". In: Nov. 2008, pp. 1–5.

[21] Minjin Lee et al. "Exploring the relationship between the spatial distribution of roads and universal pattern of travel-route efficiency in urban road networks". In: *Chaos, Solitons & Fractals* 174 (Sept. 2023). DOI: 10.1016/j.chaos.2023.113770.

[22] Minjin Lee et al. "Morphology of travel routes and the organization of cities". In: *Nature communications* 8.1 (2017), p. 2229. DOI: 10.1038/S41467-017-02374-7.

[23] David Levinson and Ahmed El-Geneidy. "The minimum circuity frontier and the journey to work". In: *Regional Science and Urban Economics* 39.6 (2009), pp. 732–738. DOI: https://doi.org/10.1016/j.regsciurbeco.2009.07.003.

[24] Valiant LG. "The complexity of enumeration and reliability problems". In: *Siam Journal of Computing* 8 (Aug. 1979), pp. 410–421. DOI: 10.1137/0208032.

[25] LongJason Lu and Minlu Zhang. "Edge Betweenness Centrality". In: *Encyclopedia of Systems Biology*. Ed. by Werner Dubitzky et al. New York, NY: Springer New York, 2013, pp. 647–648. ISBN: 978-1-4419-9863-7. DOI: 10.1007/978-1-4419-9863-7_874.

[26] István Lukovits and Marko Razinger. "On Calculation of the Detour Index". In: *Journal of Chemical Information and Computer Sciences* 37.2 (1997), pp. 283–286. DOI: 10.1021/ci960034j.

[27] Chehreghani MH. "Effectively Counting st Simple Paths in Directed Graphs". In: *arXiv preprint arXiv:2103.06102* (2021). DOI: 10.48550/arXiv.2103.06102.

[28] M.E. J. Newman. "A measure of betweenness centrality based on random walks". In: *Social Networks* 27.1 (2005), pp. 39–54. DOI: https://doi.org/10.1016/j.socnet.2004.11.009.

[29] Luca Pappalardo et al. "Future directions in human mobility science". In: *Nature Computational Science* 3 (2023). DOI: 10.1038/s43588-023-00469-4.

[30] Dino Pedreschi et al. *Social AI and the Challenges of the Human-AI Ecosystem*. 2023. DOI: 10.48550/arXiv.2306.13723.

[31] Alexis Poulhès and Laurent Proulhac. "The Paris Region low emission zone, a benefit shared with residents outside the zone". In: *Transportation Research Part D: Transport and Environment* 98 (2021), p. 102977. DOI: https://doi.org/10.1016/j.trd.2021.102977.

[32] Ben Roberts and Dirk P Kroese. "Estimating the Number of st Paths in a Graph." In: *J. Graph Algorithms Appl.* 11.1 (2007), pp. 195–214. DOI: 10.7155/JGAA.00142.

[33] M Puck Rombach et al. "Core-periphery structure in networks". In: *SIAM Journal on Applied mathematics* 74.1 (2014), pp. 167–190. DOI: 10.48550/arXiv.1202.2684.

[34] P.N. Ruane. "Heavenly mathematics: the forgotten art of spherical trigonometry by Glen Van Brummelen, pp 192, £24.95, ISBN 978-0-691-14892-2, Princeton University Press (2013)." In: *The Mathematical Gazette* 98.543 (2014), pp. 561–562. DOI: 10.1017/S0025557200008573.

[35] Chatterjee S et al. "Counting all possible simple paths using Artificial Cell division mechanism for directed acyclic graphs". In: *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. 2015, pp. 1874–1879.

[36] John W Suurballe. "Disjoint paths in a network". In: *Networks* 4.2 (1974), pp. 125–145. DOI: 10.1002/net.3230040204.

[37] Roughgarden T and Tardos E. "How Bad is Selfish Routing?" In: *Journal of the ACM* 49 (Feb. 2002). DOI: 10.1145/506147.506153.

[38] Pu Wang et al. "Understanding Road Usage Patterns in Urban Areas". In: *Scientific reports* 2 (Dec. 2012), p. 1001. DOI: 10.1038/srep01001.

[39] Harry Wiener. "Structural Determination of Paraffin Boiling Points". In: *Journal of the American Chemical Society* 69.1 (1947), pp. 17–20. DOI: 10.1021/ja01193a005.

[40] Xu Y, Di Clemente R, and González MC. "Understanding vehicular routing behavior with location-based service data". In: *EPJ Data Science* 10 (Feb. 2021), pp. 1–17. DOI: 10.1.1.619.3803.

[41] Norihito Yasuda, Teruji Sugaya, and Shin-Ichi Minato. "Fast Compilation of s-t Paths on a Graph for Counting and Enumeration". In: *Proceedings of The 3rd International Workshop on Advanced Methodologies for Bayesian Networks*. Vol. 73. 2017, pp. 129–140.

[42] Jin Y. Yen. "Finding the K Shortest Loopless Paths in a Network". In: *Management Science* 17.11 (1971), pp. 712–716. DOI: 10.1287/mnsc.17.11.712.

[43] Kai Zhang, Stuart Batterman, and François Dion. "Vehicle emissions in congestion: Comparison of work zone, rush hour and free-flow conditions". In: *Atmospheric Environment* 45.11 (2011), pp. 1929–1939. ISSN: 1352-2310. DOI: https://doi.org/10.1016/j.atmosenv.2011.01.030.

[44] Bo Zhou and Xiaochun Cai. "On detour index". In: *MATCH Commun. Math. Comput. Chem* 63.84 (2010), pp. 199–210.