**Master Thesis**

# Public Transport Network analysis

Complex network analysis of 27 PTNs from cities around the world

**Sara Cabodi**

**Supervised by**

Paolo Garza

Jari Saramäki

Final Project Report for the

Master in Ingegneria Informatica


Dipartimento di Automatica ed Informatica

Politecnico di Torino

Italia, Torino

July 2020

# Abstract

At least once in a lifetime, everyone has taken a means of public transport. Buses, subways, trains, etc., are a part of our everyday life. They are how we commute to work, meet a friend for a coffee and visit or travel to different places.

In the last decades, researches have studied the topology and characteristics of Public Transport Networks (PTN) in order to understand, plan and optimize their behaviour, cost and performance. In particular, when dealing with spatial networks such as PTNs, complex network theory plays a huge role in analysing and understanding their properties.

In this thesis we focus on the PTNs analysis of 27 cities located in three continents: Europe (19), Oceania (5) and America (3). We model each transportation network as a graph represented by an *L-space* topology, where stops and stations represent nodes and their connections edges, *e.g.*, a bus going from stop A to stop B. This work aims at finding possible relations/patterns involving the city features, such as area and population, and the properties of its PTN.

We collect basic static measurements for each city, such as number of nodes and edges, clustering coefficient, density and diameter. We deepen the network analysis discussing assortativity and average path length. We further explore the properties of each network through the distributions of node measures, like degree and different types of centrality. We then explore the networks in order to analyse shortest paths and distances, computed by standard graph algorithms and evaluated taking into account Euclidean distances. This allows us to partially capture some geographical, topological and functional characteristics of the observed networks. We conclude our work with a frequency analysis. The goal is to display and analyse the distributions of number of vehicles throughout a typical day. The work is done separately for each type of transport present in the dataset, which allows to better compare the situation in different cities.

We use local as well as global features to evaluate characteristics of urban transportation systems according to well-known network theory, e.g. small-world and

scale-free properties. We make local and global analysis on individual and multiple urban networks, considering their basic topology as well as clustering strategies based on commonly used properties. Each analysis has its own level of details, depending on the type of measure taken into consideration. For example, in some cases it is possible to have both city level analysis and comparison among cities for all measures considered, whereas other times it is necessary to divide by type of measurement.

The results obtained from the network analysis suggest that PTNs, as many other real-world networks, are neither small-world nor scale-free. Lastly, for each of the part of the analysis performed we were able to capture some insights both at city level as well as in terms of comparisons among all cities.

# Acknowledgements

My journey towards this thesis began last year, when, in April, I was looking for a thesis supervisor. I found in Prof. Jari Saramäki the person who would guide me through this new experience. I would like to thank him for his support, patience, feedback and willingness to help. I also wish thank him for his understanding and flexibility throughout the pandemic situation. I started my Erasmus with very small and basic knowledge of complex networks and, since then, I've learned so much, even though I'm aware to be just at the beginning.

I wish to express my gratitude towards my Italian supervisor, Prof. Paolo Garza, who followed me and my work during these months, offering support and feedback.

A special thank goes to the Complex Systems group of the Department of Computer Science at Aalto University. I got the chance to meet both great minds and people. I am particularly grateful for your effort and kindness to include me in the group, even when the work was shifted remotely. Thank you all for making me learn about other cultures, foods and for making me feel included: Abbas, Ana, Anna, Mikko, Richard, Sara, Silja, Takayuki, Talayeh, Tarmo and Tuomas.

I wish to thank Laura Mursu, Study Coordinator at Aalto University, who helped and guided me throughout my Erasmus adventure in Finland.

I also would like to thank all the people who have been a part of my academic journey, both in Italy and in Finland. It has been a challenge from many points of view, but you've helped me find my way. I would like to give special thanks to Antonino, Beatrice, Davide, Ilenia and Martina.

Finally, I wish to thank my family and my boyfriend, Mattia. You have always supported me and gave me love and strength to overcome every obstacle along the way.

Torino, 19.06.2020

Sara Cabodi

# Contents

# List of Figures

# Abbreviations

| | |
|---|---|
| **N** | number of nodes |
| **E** | number of edges |
| **V** | set of nodes |
| **L** | set of edges/links |
| **A** | area in $km^2$ |
| **P** | population in thousand of inhabitants |
| **k** | degree of a node |
| **d** | diameter of a graph |
| $\langle \mathbf{k} \rangle$ | average degree |
| $\langle \mathbf{c} \rangle$ | average clustering coefficient |
| $\langle \mathbf{r} \rangle$ | assortativity coefficient |
| $\langle \mathbf{l} \rangle$ | average path length |
| $\mathbf{g(i)}$ | betweenness centrality of the node $i$ |
| $\mathbf{C_c(i)}$ | closeness centrality of the node $i$ |
| **PTN** | public transport network |
| **xTN** | x transport network, where x can be the initial letter of any type of transport (bus, tram, rail, etc.) |
| **BFS** | breadth first search (sometimes also referred as bfs) |

# Chapter 1

# Introduction, Motivations and Goals

In our everyday lives we are surrounded by numerous complex systems formed by many interacting elements. If we think about city mobility, public transport is the choice for countless people. Network science is a discipline that aims to model these systems of interacting components as networks where different entities are represented as nodes and the relationships between them as edges [1]. The specific case of public transport can be modeled as a network with stops as nodes and connections between consecutive stops as edges. Starting from the 2000s, researches have began to study the topology and peculiarities of Public Transport Networks (PTN) in order to understand, plan and optimize their behaviour, cost and performance. In particular, when dealing with spatial networks such as PTNs, complex network theory plays a huge role in analysing and understanding their properties.

In this work we analyse the Public Transport Networks of 27 cities around the world, divided between three continents as follows:

- America: Antofagasta (Chile), Detroit (USA), Winnipeg (Canada),
- Europe: Athens (Greece), Belfast (Northern Ireland), Berlin (Germany), Bordeaux (France), Dublin (Ireland), Grenoble (France), Helsinki (Finland), Kuopio (Finland), Lisbon (Portugal), Luxembourg City (Luxembourg), Nantes (France), Palermo (Italy), Paris (France), Prague (Czech Republic), Rennes (France), Rome (Italy), Toulouse (France), Turku (Finland), Venice (Italy),
- Oceania: Adelaide (Australia), Brisbane (Australia), Canberra (Australia), Melbourne (Australia), Sydney (Australia).

The work aims at analysing the PTNs to first characterize the networks in terms of static measures, topology and network models. Moreover, our goal is to find, if present, relations or patterns between city features and PTN properties. To evaluate these possible correlations, we consider area and population as city features and we compute two main types of analysis:

- Distance analysis for the area
- Frequency analysis for the population

Chapter 2 presents the theoretical background needed to support the analysis. In section 2.1, we offer an overview of the network theory. We report the significant network representation and types, followed by a brief review of useful local and global network properties. The second section (2.2) presents the features and literature background of PTNs as networks. Firstly, we describe the essential features of spatial networks, followed by specific topologies used in this context. To conclude, we offer a quick review of the previous works studied to perform our analysis.

Chapter 3 explains the data structure. Section 3.1 presents the dataset, its features, how it has been created and some basic statistics. It concludes with an explanation of our collection process, concerning area and population information. In section 3.2 we describe the networks creation, underlining their type, and the process followed to plot the networks.

Chapter 4 is the main chapter, where we describe how we performed the analysis itself and its results. It is divided in 5 threads, each one explaining a specific part of the analysis. The first section (4.1) illustrates the network basic measures analysed for each city. We talk about number of nodes, number of edges, density, diameter and average clustering coefficient. Together these measures offer a first rough interpretation of the dataset. Furthermore, an outlier analysis and a correlation one give more detailed information about the different PTNs. The second section (4.2) deepens the first one. In particular, we further the exploration of network measures analysing assortativity, average path length, average degree and degree distribution. Through this analysis, we compare some well known network models and properties, i.e. small-world and scale-freeness. Section 4.3 offers an overview of nodes centrality measures. To be more precise, we analyse four types of centrality: betweenness, closeness, degree and eigenvector. We briefly review their meaning, explain their distributions with fitting parameters and comment on some comparisons. In section 4.4, we describe the distance analysis. We illustrate our approach, choices and implementation, offering a specific subsection where we compare the breadth first search, chosen one, to the Dijskstra one. The goal of this type of analysis was

to evaluate the efficiency of the PTNs, characterizing trips as well as nodes and their mutual reachability. To do so, we compared the real distances, considered as the BFS ones, with the Euclidean distances. We furthered the analysis evaluating the distances from a geographical center and finding out the network distribution of peripheral nodes. All of the results are presented and discussed both at city level and comparison one. During the implementation of this type of analysis, it became necessary to investigate the connected components of the networks. The results are presented in the last subsection of this part. In conclusion, section 4.5 deals with the frequency analysis. When approaching this last part, we aimed at studying the distributions of vehicles frequency during a typical day. The information were gathered and presented divided by type of transport, both for individual and collective plots. Like for the previous section, we displayed the results at two different level of detail: city level and comparison among city level. We deepened the study for the bus transport networks, because all the cities had information about them. Its results are shown and discussed in the last subsection.

The last chapter, 5, presents some general and specific conclusions for each section of the analysis. In addition, we offer a few ideas for future works and some suggestions on how the work done may be used to improve PTN planning and service.

# Chapter 2

# Background

This chapter overviews theoretical aspects on networks and graphs that are at the base of the analysis tasks we performed on PTNs. We first introduce general concepts on network representation, the types of networks and the local and global properties we are interested at. We then focus on PTNs seen as spatial networks, we briefly overview the main network topologies found in literature, and we finally overview a set of related works that can be considered as most relevant to to our work.

## 2.1 Network theory

Network theory is the discipline studying graphs in order to represent sets of discrete objects characterized by pairwise relations, that can be either symmetric or asymmetric. In computer science and network science, network theory is a part of graph theory: a network is a graph in which nodes and/or edges have attributes. Network representations are exploited in many disciplines, such as physics, computer science, electrical engineering, biology, economics, climatology, and sociology.

### 2.1.1 Network representation

The study of complex networks is a specific and relatively young (originated in the early 2000) field of network theory, with applications in areas of scientific research, based on the empirical study of real-world networks such as computer networks, biological networks, technological networks, brain networks, climate networks and social networks [2].

A network is a graph, i.e.,, a mathematical structure composed by a set of

objects, in which some pairs of the objects are in some sense "related". It is used to model real-life phenomena composed by entities which interact with each other or are interconnected.

To be able to model and study these phenomena, we need to exploit a mathematical representation of the corresponding network. First of all, we need to uniquely identify the entities of our model, the graph nodes, and to label them with proper attributes. Graph nodes are thus unique, and one can store other significant information inside, for example the position or coordinates of the geographical entities they represent. The number of nodes in a graph is a widely used measure of the graph size, and as a consequence of the cost of the operations to be done on it. We will here refer to it as $N$.

Relationships between nodes are represented by edges connecting them, directed or undirected, weighted or unweighted. Network edges are represented as edge lists, adjacency matrices or adjacency lists, depending on choices to be done in terms of memory usage and/or complexity of the algorithms.

Whereas edge lists and adjacency lists can be more compact, as they just represent existing edges, the adjacency matrix is often considered a simpler/straightforward option: though representing both existing and non existing edges, it provides an O(1) access to an edge, starting from the node pair it connects. It is a square matrix ($N \times N$), where each element ($A_{ij}$) indicates whether pairs of nodes ($i$ and $j$) are adjacent or not in the graph: the information is Boolean in unweighted graphs, whereas it is the edge weight in weighted graphs.

A single edge list is a very simple representation, useful whenever an algorithm just requires iterations on all graph edges. Adjacency lists represent lists of edges on a per node basis: lists are collected either in arrays or lists of $N$ lists, each one describing the set of neighbours of the $i$-th node in the graph.

Each representation type has its pros and cons. For example, adjacency matrices are useful for procedures needing easy and fast indexing. However, they may cause higher memory consumption for sparse networks. Which representation to choose depends on the network size and density, but also on the algorithms used for processing and the memory constraints for the storage.

### 2.1.2 Network types

The great variety of real-life phenomena that can be modeled trough networks calls for flexibility in the definition of the different types. For example transport networks

have spatial constraints and require information about the coordination of their nodes. In some more complex cases, the limitations of the network estimation methods may have their effect on the network type produced, too. The thesis focuses on unweighted and undirected graphs and studies properties associated with static networks. The following subsections present some of the types of graph useful for the analysis performed.

### Simple graphs and multigraphs

The most basic type of network is the graph with no self-loops and no parallel edges, called simple graph. The maximum number of possible edges in a simple graph with $N$ vertices is

$$E_{max} = N(N-1)/2. \tag{2.1}$$

On the other hand, the so-called multigraph can contain self-edges and multi-edges, *i.e.*, edges between a node and itself, or multiple edges between the same pair of nodes.

### Weighted and unweighted networks

Some kind of phenomena need to customize the interaction with some sort of numerical attribute, which explains the intensity or the type of interaction itself. A weighted graph is a graph in which each branch is given a numerical weight. It is, therefore, a special type of labeled graph in which the labels are numbers (usually positive). As opposed to a weighted graph, the unweighted one is characterized by the fact that edges do not have any associated cost or weight. The difference between a multigraph or a weighted network, in terms of edge representation, is not always clear, especially for the adjacency matrix when the values are integers. For instance, an unweighted multigraph could be represented by a weighted simple graph, where integer edge weights represent the number of edges in the multigraph. The distinction is normally rather clear when considering the phenomena being represented.

### Directed and undirected networks

When dealing with relationships, up to now we considered two ways (symmetric) ones (an edge from $i$ to $j$ is equivalent and undistinguishable from an edge from $j$ to $i$). These types of networks are referred to as undirected. However, in some contexts, it might be necessary to separate the two directions, by creating a directed graph: within the framework of transport networks, an edge between two nodes often

represents a bi-directional connection, which means that means of transportation run in both directions. Nonetheless, this is not always true, so depending on the specific network, the undirected graph representation can be enough, or one should resort to the directed version.

### 2.1.3 Local and global properties

#### Adjacency matrix

A graph with $N$ nodes and $E$ edges can be described by its $N$ x $N$ adjacency matrix $A$, which is defined as

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{Otherwise} \end{cases} \tag{2.2}$$

If the graph is undirected then the matrix $A$ is symmetric.

#### Degree

If there is an edge $(v_i, v_j) \in L$, we can say that $v_i$ and $v_j$ are adjacent, $v_i$ is a neighbour of $v_j$ and the edge is incident to $v_i$ and $v_j$.

The degree $k_i$ of vertex $v_i$ is the number of edges it is incident to. For simple graph, this is the number of neighbours of the considered node. In case of directed graphs, one can consider separate in- and out-degrees, corresponding to leaving and incoming edges.

The average degree $\langle k \rangle$ of a network is

$$\langle k \rangle = \sum_i \frac{k_i}{N} = \frac{2E}{N}, \tag{2.3}$$

where $E$ is the total number of edges and $N$ is the total number of nodes.

The *degree distribution* $P(k)$ is one of the central concepts in network analysis. It represents the fraction of nodes having degree $k$ or, equally, it is the probability that a uniform randomly chosen node has degree $k$. That is,

$$P(k) = \frac{N_k}{N}, \tag{2.4}$$

where $N_k$ is the number of nodes of degree $k$.

Some networks, notably the Internet, the world wide web, and some social networks are found to have degree distributions that approximately follow a power law: $P(k) \sim k^{-\gamma}$, where $\gamma$ is a constant. Such networks are called scale-free networks and have attracted particular attention for their structural and dynamical properties. However, real-world networks are rarely scale-free, as it is thoroughly explained in [3].

### Clustering coefficient

The *clustering coefficient* of node $i$ is the ratio of the number of edges between its neighbours to that of the number of possible such edges:

$$c_i = \frac{E_i}{\binom{k_i}{2}} = \frac{2E_i}{k_i(k_i-1)} \qquad c_i \in [0,1], \tag{2.5}$$

where $E$ is the number of edges between $i$'s neighbours. It can also be seen as the density of the local neighbourhood of a node.

When applied to an entire network, it is the average clustering coefficient over all of the nodes in the network. It can indicate if a network shows *small-world* properties. For $\langle c \rangle$ to be meaningful, it should be significantly higher than the one obtained from a random graph with the same number of nodes.

### Diameter and average shortest path

In a network, a path is a walk where vertices are never repeated. The length of a path is the number of edges on it. The path with the minimun number of edges between two nodes is called *shortest path*. The (geodesic) distance $d_{ij}$ of two vertices is the length of their shortest path. The *diameter* $d$ of a network is the maximum distance found in it, $d = max(d_{ij})$. The average path length, together with the clustering coefficient described above, contributes to underline possible *small-world* behaviour of the network. In particular, if the average distance between the nodes is proportional to the logarithm of the number of nodes, $\langle l \rangle \sim \ln N$, then the network has small-world properties.

### Assortativity

In general the degrees at the two end nodes of a link are correlated, and to describe if one can estimate the conditional probability $P(k'|k)$. This quantity represents the probability that any edge starting at a certain node of degree $k$ ends at a node of

degree $k'$. However, the function $P(k'|k)$ is hard to estimate and one can define the assortativity. The latter is defined as a preference for a network's nodes to attach to others that are similar in some way.

$$k_{nn}(k) = \sum_{k'} P(k'|k)k'. \tag{2.6}$$

Usually, the average nearest-neighbour degree $k_{nn}$ is used instead:

$$k_{nn}(k) = \frac{1}{N(k)} \sum_{i,k_i=k} \left[ \frac{1}{k} \sum_{j \in \Gamma_i} k_j \right], \tag{2.7}$$

where $\Gamma(i)$ denotes the set of neighbors of $i$. Another common way to calculate the assortativity is to use the Pearson correlation coefficient between degrees of linked nodes:

$$r = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\sqrt{\langle k_i^2 \rangle - \langle k_i \rangle^2} \sqrt{\langle k_j^2 \rangle - \langle k_j \rangle^2}}. \tag{2.8}$$

If the value of the coefficient is positive, we find ourselves in the situation of an assortative mixing, meaning that vertices with large degrees have a greater probability to connect to similar nodes with a large degree. In general, social networks are mostly assortative, while technological networks are disassortative. However, for spatial networks, the spatial constraints usually imply a flat function $k_{nn}(k)$.

### Centrality measures

In network theory, we can find several measures of centrality that try to highlight the importance of nodes or edges based on some features.

Here, we will explain only some of them, useful for the explanation of the analysis that will follow in chapter 4.

### Betweenness centrality

Betweenness centrality highlights important nodes which work as bridges, using the number of shortest paths passing through each of those nodes. In a way, it measures the traffic or flow through a node/link, if all nodes communicate to all others via the shortest paths. Formally, betweenness centrality is the fraction of shortest paths going through node/link. In the case of multiple shortest paths, it is divided by multiplicity. This measure of centrality is very effective for spatial networks and it is probably the most significant and largely used, even though it is hard to compute for large networks.

#### Closeness centrality

Closeness centrality highlights important nodes as those close to each other. The closeness of one node to all the others is computed as follows

$$C_c(i) = \frac{1}{\langle l_i \rangle} = \frac{N-1}{\sum_{j \neq i} d_{ij}}, \tag{2.9}$$

where $d_{ij}$ is the length of shortest path between $i$ and $j$, *i.e.*, their distance; $\langle l_i \rangle$ is the avg distance from $i$ to others and $C_c(i)$ is the inverse of the avg distance. This centrality measure does not directly work for networks with disconnected components where for some pairs $d_{ij} = \infty$.

#### Degree centrality

Degree centrality, which is defined as the number of links incident upon a node, is the first one invented and probably the simplest one. In case of directed networks, we usually define two separate measures of degree centrality, in-degree and out-degree. Accordingly, in-degree is a count of the number of links directed to the node and out-degree is the number of links that the node directs to others.

#### Eigenvector centrality

Eigenvector centrality defines important nodes those which are connected to other important nodes. It is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

## 2.2 Public transports as networks

In this thesis we analyse public transport networks of 27 cities around the world. Together with airline networks, cargo ships networks, road networks, they belong to a more general category of the so-called *spatial networks*. This section first introduces general notions about spatial networks, then briefly describe existing Transport Network topologies and their representation. Finally, we overview a set of related works and references from the literature.

### 2.2.1 Spatial networks

In this section we describe some key features of spatial networks, taking as reference the paper [4], that can be considered as a very complete and thorough description of spacial networks in multiple application domains, as well as the data structures and algorithms adopted to solve the related problems.

Spatial networks are networks for which the nodes are located in a space equipped with a metric, usually the Euclidean distance, for the case of a two-dimensional space. When dealing with spatial networks, there are some key aspects to take into consideration. For example, in this type of networks the probability for an arbitrary pair of nodes to be connected by an edge typically decreases with their distance. In the case of infrastructure networks, e.g. power grids, this means also that the network is planar [1]. The latter assumption, however, is not always true, since the links might not be necessarily embedded in space, like for the airline passenger networks. Another aspect to consider is, as just mentioned, how the network is embedded in space. Even though some networks do not seem to be directly embedded in space, they might still be considered as spatial. For instance, for social networks the space factor comes up when talking about link probability, which decreases with the distance between the nodes. It might be interesting to apply a Voronoi tessellation to a spatial network. It is a way of dividing space into a number of regions, that can provide a natural representation model to which one can compare a real world network. Topology is a last aspect to consider, as another important way to characterize networks, by examining the nature, disposition and relation of nodes and edges.

Let's review some empirical results found in [4]. The distribution of degree P(k) is usually a quantity of interest, as it can display some heterogeneity, such as the ones observed in scale-free networks (see for example [5]). In addition, the authors of [5] also observed in some types of networks, such as airline networks or the Internet, node degrees are very heterogeneous. However, when physical constraints are strong or when the cost associated with the creation of new links is large, a cut-off appears in the degree distribution [6], and in some cases the distribution can be very peaked. This is the case for the road network and more generally for of planar networks, for which the degree distribution P(k) is of little interest.

[1] A planar graph is a graph that can be drawn in the plane in such a way that edges do not intersect.

### 2.2.2 Transport Network topologies

When focusing on transport networks, as a specific instance of spatial networks, routes are an important component of the network topology. A route is an intermediate notion between edges and paths: a route is the path serviced by a given mean of transport. The available literature on transportation systems proposes two major strategies to represent routes, based on the notions of L-space and P-space [7, 8]. The L-space topology connects nodes if they are consecutive stops on a given route. The degree in L-space is then the number of different nodes one can reach within one segment, and the path length represents the number of stops. In the P-space, two nodes are connected if there is at least one route between them, so the degree of a node is the number of nodes that can be reached, either directly or indirectly, on that route. In P-space, a path length represents the number of connections/transfers needed to go from one node to another.

Although in this thesis we analyse PTNs based on the L-space topology, we here briefly list the topology classification made in paper [9].

#### L-space

This type of graph topology, sometimes referred also as space L, represents each station by a node, a link between nodes indicates that there is at least one route that services the two corresponding stations consecutively. In addition, no multiple links are allowed, so just one edge will connect two nodes, even though they were directly connected on multiple routes.

#### P-space

The P-space graph representation has proven particularly useful in the analysis of PTNs. Here, the nodes are stations, like for the L-space, but they are linked if they are serviced by at least one common route. In this way the neighbours of a P-space node are all stations that can be reached without changing means of transport and each route gives rise to a complete P-subgraph.

#### B-space

A somewhat different concept is that of a bipartite graph which has proven useful in the analysis of cooperation networks. In this representation, which is called B-space,

both routes and stations are represented by nodes. Each route node is linked to all station nodes that it services. No direct links between nodes of the same type occur.

#### C-space

The complementary projection of the B-space graph to route nodes leads to the C-space graph of route nodes, where any two route nodes are neighbors if they share a common station.

### 2.2.3 Related works

In the last two decades, researchers around the world started analysing PTNs as complex network systems, based on the observation that data analysis, based on the complex network theory, could be a key step in planning, decision making as well as simple performance evaluation of transport networks. This section presents an overview of the related works that we studied, as a preliminary step, before proceeding with the analysis of our dataset.

Chen et al. [10] investigated the bus transportation networks of four major cities in China (Hangzhou, Nanjing, Beijing and Shanghai). They showed that both the degree and the number of bus routes a stop joins distributions follow a power-law with exponential decay. On the other hand, the distributions of the number of stops in a bus route follow asymmetric, unimodal functions.

Sen et al. [7] studied the Indian railway network and discovered small-world properties and exponential degree distribution. In particular, the study of the mean distance of the network showed its goodness to measure the connectivity of the network. Indeed, the observation of its logarithmic variation with the number of nodes together with a high value of the clustering coefficient led to the discovery of small-world properties of the networks.

Sienkiewicz and Holyst [11] collected and analyzed the data of the PTNs of 22 cities in Poland and found that the degree distributions in L-space follow a power-law, while in P-space they are exponential. In addition, small-world behavior was observed in both topologies, but it was much more pronounced in P-space, where the hierarchical structure of the network was also deduced from the behavior of clustering coefficient.

Ferber et al. [9] studied the PTNs of 14 major cities around the world. They analysed different topologies and found that the networks have strongly correlated

small-world structures and that the degree distributions follow a power-law with various exponents giving strong evidence of correlations within these networks. However, for the properties of degree distributions as well as for features of these networks, such as clustering, assortativity and others they found considerable diversity in their expression. Lastly, they also proposed an evolutionary model of growth of PTNs.

Hàznagy et al. [12] analyzed the urban public transportation systems of 5 Hungarian cities, considering directed and weighted links, where the weights represent the capacities of the vehicles (bus, tram, trolleybus) in the morning peak hours. They discovered that, indipendently of the morphology of the cities, the PTNs have a few high-degree nodes where many lines cross, but most of the nodes have a low degree resulting in a fat-tailed degree distributions. They highlighted both similarities and differences between the cities and managed to identify the most sensitive routes and stations of the networks.

Xu et al. [13] analysed the bus-transport networks (BTN) of three cities in China. They explored scaling laws and correlations that may govern intrinsic features of the analysed networks. They observed distributions of degree, strength and weight in a weighted representation of the networks. In accordance with other researches, they found that degree distribution and distribution for the number of lines that service each station obey power laws while the cumulative degree distribution in P-space follows an exponential distribution. Moreover, small-world behavior was observed in both topologies, but it was stronger in the P-space topology. Lastly, they observed a heavy tailed power law for the weight distribution and a linear dependence between the strength and degree.

Zhang et al. [14] analysed the bus transport network of Beijing using both L and P-space topologies. In the L-space analysis they discovered that the network is scale-free, is assortative and has 46 communities. With the P-space topology, they investigated the property of transfer, discovering an average transfer time of 1.88 and that two pair nodes is reachable within 4 transfers.

Shanmukhappa et al. [15] studied the topological behavior of the bus transport network structure of three cities: Hong Kong, London and Bengaluru. In 2017, they proposed a novel approach called supernode graph structuring for modelling BTNs to combine geographically closely associated nodes based on a specific criterion, resulting in a more compact representation. It is observed that the supernode concept has significant advantage in analyzing the inherent topological behavior. For instance, the scale-free and small-world behavior becomes evident with supernode representation as compared to conventional or regular graph representation for the Hong Kong

network. Furthermore, they created weighted networks, assigning node weights based on the POI (Point of Interest) density and the population distribution in the city over various localized zones in order to obtain a better estimate on the dynamic behavior of the network. Lastly, they evaluated topological efficiency through end-to-end travel delays, finding out that Hong Kong is the most efficient among the three.

The same authors, together with Wu and Dong, [16] published another article which describes how they modeled the public transport network structure of the London city, using the "supernode" (set of geographically closely associated nodes) graph structure representation. The bus transport and the metro transport network structures are analyzed by treating them as independent mono-layer or multi-layer network structures, using a method of spatial amalgamation to integrate the two layers. Lastly, a node weight analysis method is presented and it is noticed that the node weights differ between the mono-layer and multi-layer analyses, which indicates that neglecting the interaction between the transport layers may bias our understanding of the overall network behavior considering the real-world usage of the network.

Another interesting work by Shanmukhappa et al. [17] brings together the recent development in the field of public transport analysis from a graph theoretic perspective with a focus toward bus transport network (BTN) and metro transport network (MTN). They found that the notion of supernodes offers practical and more insightful perspective to understanding the actual network behavior, which is difficult to be captured by conventional graph representations. Furthermore, adding static weights to nodes and edges has been found to be effective in capturing the significance of nodes and links in PTNs. In addition, they suggested that merely representing the PTN structure as a graph and analyzing various network parameters may not lead to practically useful conclusions because the purpose of the public transport systems is to meet travel needs of the community being served, which requires the consideration of more practical network parameters. To summarize, this work offers a recent collection of the different techniques, topologies and parameters used to analysed PTNs and their advantages and differences.

Soh et al. [18] analysed the weighted networks of travel routes on the Singapore rail and bus transportation systems, using both topological and dynamical analysis. The results tell that the second approach adds information to the topological analysis, giving a richer view of complex weighted networks. In addition, inspection of the weighted eigenvector centralities highlighted a significant difference in traffic flows for both networks during weekdays and weekends, suggesting the importance of adding a temporal perspective missing from many previous studies.

Zhang et al. [19] analysed the Shanghai subway network, studying its topological characteristics and functional properties in order to assess the reliability and robustness. In particular, the fraction of removed nodes of the network is discussed and compared against that for a random network, and the critical threshold of this fraction is obtained. Moreover, they proposed two novel parameters called the functionality loss and connectivity of subway lines to measure the transport functionality and the connectivity of subway lines. The results obtained indicate that the subway network is robust against random attacks but fragile for malicious attacks and that the highest betweenness node-based attacks can cause the most serious damage to subway networks among the different attack protocols.

Chatterjee et al. [20] modeled the bus networks of six major Indian cities as graphs in L-space, and evaluate their various statistical properties. Although they observed the common feature of small-world property throughout the dataset, their analysis reveals different network topologies due to significant variation in the degree-distribution patterns in the networks. They also observed that these networks, although robust and resilient to random attacks, are particularly degree-sensitive.

# Chapter 3

# Data structure

In this chapter we describe a set of data representing the public transport networks of 27 cities selected from different continents and countries. Section 3.1 first describes the cities and the information and formats available for each one. Then, it deals with the types of transport networks, offering some very basic statistics on their distribution in the dataset. Afterwards, it offers a brief review of the spatial filtering performed to the raw data, even though it is not part of this work. Lastly, it presents the process of collecting city features data: area and population. Section 3.2 introduces the graph-based representation of PTNs. We briefly describe the choices and process followed to create and plot the networks.

## 3.1 Data

This section introduces the set of data used. It briefly describes the which cities are present and the data available for each one. We then introduce the Public Transportation Networks (PTNs) considered presenting the types available and how the information was spatially filtered. In addition, the section offers some very basic statistics on the dataset, *e.g.*, on the differentiation of type of transport and how many each city has. Finally, we briefly discuss how we gathered data on population and area and their possible relationship with transportation data and analysis.

### 3.1.1 Subjects

In this thesis, we worked on a dataset of 27 cities from different continents and countries, created by a research group on complex networks at Aalto University,

Espoo [21]. The dataset, thoughrouly explained in [22], includes cities from different areas of the world, distributed among three continents as follows:

- Europe (19): Czech Republic (1), Finland (3), France (6), Germany (1), Greece (1), Ireland (2), Italy (3), Luxembourg (1), Portugal (1)
- Oceania (5): Australia (5)
- America (3): Canada (1), Chile (1), USA (1)

Data were collected with the aim of covering cities of different sizes, located in different continents, and representing various geographical characteristics, such as location, area, population, landscape, etc. Figure 3.1 gives a high level representation of the geographical distribution of cities: though it is clear that the majority of cities are in Europe, the set includes instances from Australia, North and South America. It is obvious that an expansion to cites in Asia and Africa would improve the global coverage of the dataset.

For each city, the dataset contains information about its Public Transport Networks (PTNs) in multiple files and formats, that include:

- network nodes list
- network edge lists
- temporal network event lists
- SQLite databases
- GeoJSON files
- GTFS data format

The dataset is obviously far from complete and fully representative, as the final selection of cities was heavily affected by the availability of data, together with the licensing terms for the source data. Just for further clarity, the cities reported in [22] do not include Athens and Antofagasta for licensing problems, but they are included in this thesis. In spite of the mentioned limitations, we believe the set of data is an interesting starting point for a modern approach to data analysis of urban PTNs, as the techniques described in this thesis could be easily applied and expanded to a larger dataset, when available.

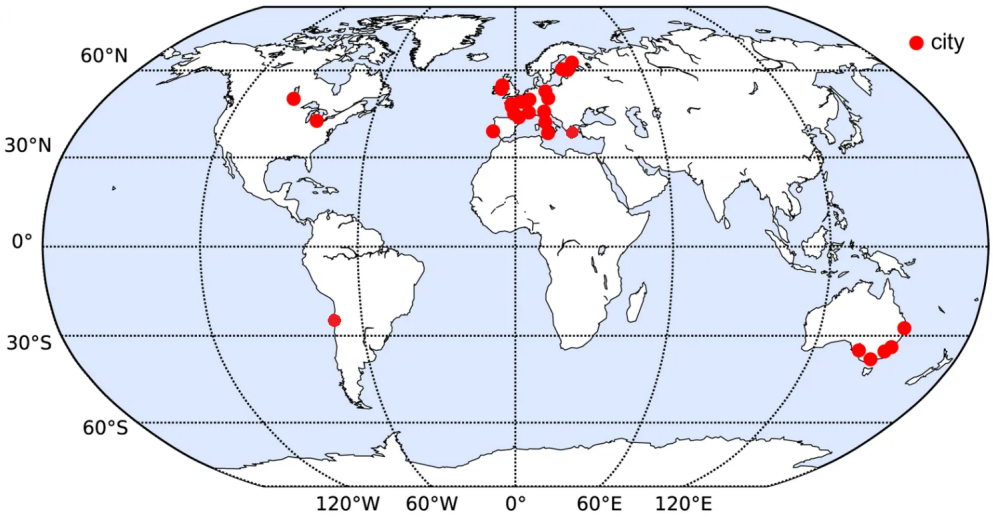



Figure 3.1: Map of database cities around the world

### 3.1.2 Types of PTNs

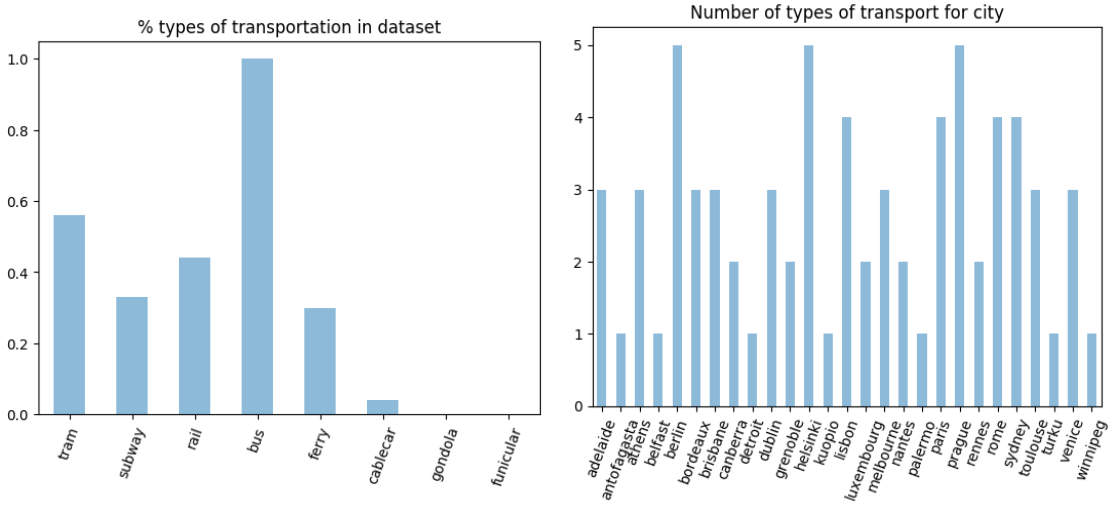
The dataset includes eight different types of means of transport, that can be roughly grouped in:

- most common and widely available: *bus*, *tram*, *subway* and *railway*
- occasionally found, such as *ferries*, available in some of the coastal cities or cities with big rivers
- rarely found, as tied to very peculiar characteristics of cities: *cablecar*, *gondola*, and *funicular*.

As shown in Figure 3.2a, we can see that all of the cities have bus transportation networks. Tram is the other broadly present type of transport, available in over 50% of the cities. Among the remaining types, we find the rail with 44%, subway and ferry both with 33% and lastly cabelcar with just 4%.

Besides adapting to geographical characteristics of cities, we can also rank and evaluate urban PTNs of a city, based on diversity, *i.e.*, number of available different PTNs. Figure 3.2b shows that the only three cities having five different types of transportation are Berlin, Helsinki and Prague (only one with cablecar in the whole dataset). It is also interesting to notice that these are not the biggest cities (in terms of area) in the dataset. Furthermore, we can observe that some of the biggest ones, like Melbourne and Sydney, have only three and four distinct types of PTNs respectively. In light of these simple observation, we can come to a counterintuitive

conclusion: the correlation between the area of a city and the number of its different PTNs is weak.




(a) (b)

Figure 3.2: Statistics on types of transportation in the dataset

### 3.1.3 Spatial information

Let's now move to the concept of urban area and urban PTN. Cities are usually identified in terms of municipality, urban and/or metropolitan areas, where the terms, though widespread and universally understood, can have different meanings and ties to political, administrative and social environments. Our work focuses on the intermediate notion of urban area, that is not always easily identified in publicly available data on transportation networks. Our dataset is the result of a preprocessing step, that was not performed in this thesis, but is worth being mentioned and understood, before starting the description of our main contributions.

Original data on urban PTNs often includes transportation stops and links from metropolitan and regional areas, that are almost impossible to filter out, unless with a heavy a difficult manual work, requiring additional area-specific data. So an empirical filter was adopted, based on a geographical and topological definition of the urban area of a given city, and of a stop, that could include a cluster of nearby an very close physical stops.

Firstly, the stops that are less than one meter apart were aggregated in one. Then, for each city they defined a central point, usually corresponding to the central railway station, and a radius. Doing so, they managed to define an area for each city. All the stops inside were kept, the links between in and out stops were lost

and, if the trip returned inside the delimited area, it was split in two parts in order to differentiate, but still keep the same trip identification. Lastly, they recorded the information about stops latitude and longitude, so they can be geographically plotted and it is possible to compute distances between them. In Figure 3.3, there is a visual example on how the spatial filtering was made.

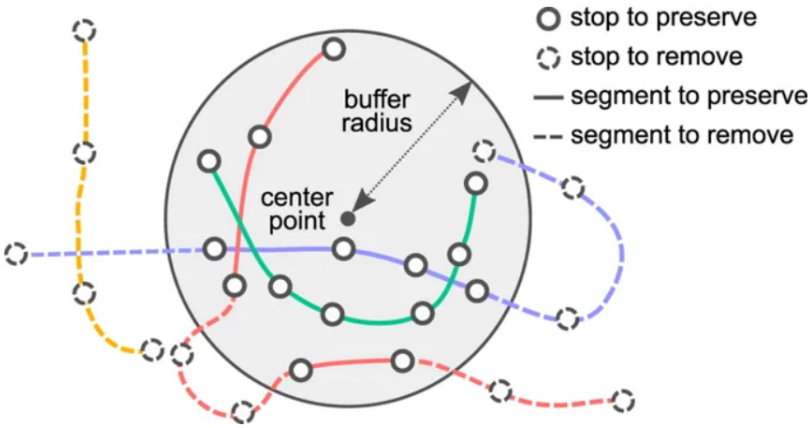



Figure 3.3: Spatial filtering from [22]

### 3.1.4 Area and population

One of the goals of this thesis is to discover possible correlations between city features and their PTNs, and if existing, how to characterize them. In order to do so, we started by gathering information about area and population, chosen as features to be analysed.

In the spatial filtering of the preprocessing, for each city a radius R was defined. It served the purpose of deciding the limit for each network and so it was calculated approximately and not to define a precise area for the PTNs. Moreover, the information about the radius did not come together with one of the population inside that designated area. For these reasons, we decided to collect our own information about both measure through several public sources. Given the non uniformity of data (population is a very dynamic statistic, urban area in not always well defined) we collected data and we compared them among different web sources, in order to come up with a unique number for area and population.

Concerning the area, we applied a manual approach. We consulted two main sources, [23] and [2], and compared the images of the areas represented in the [23] website with those of the actual PTN network for each city. This choice was made because there is not a unique definition of city area and the boundaries change based on the context. In our research, we found three different definition of area:

- municipality
- urban
- metro/metropolitan.

We found that from the first to the last definition we have increasing borders considered. In particular the metropolitan area usually includes a wide surface, which includes most of adjacent and close towns. Given the choice of the network boundaries, defined in the previous paragraph, we decided to visually compared the areas and take the more accurate one, even though it might not be perfectly fitted for the network.

As for the population, in addition to the previously mentioned websites, we compared the results also from [24] and [25] websites.

In both cases, the information was not easily collected and verified. The major problems were related to different concepts of area, the year of reference of the information and the lack of some recent or accurate data. We overcome these obstacles by manually validating those data mostly through comparison of different sources and visual comparison of maps. For example, Adelaide was one of the tricky cases. The German website [23] suggests an area of 837 $km^2$ (together with a map representation) and a population in 2016 of 1165639 inhabitants. However, Wikipedia [2] presents 1295 $km^2$ and a population in 2010 of 1203873 inhabitants. Since the first information is more recent and has a visual feedback on the actual area considered, that is quite close to the one of the network, we chose to take this one into account for our analysis. Another example of contrasting results is Rome. In this case, [23] gave 424 $km^2$ and [2] 1287. We did an additional verification through Google Maps, taking an approximate radius on the map (given the round shape of the city), that led us to confirm the first number found. On the other hand, we had cases where the information were quite similar. Helsinki has an area of 683 $km^2$ in the website [23] and of 770 $km^2$ in [2]. Another example is Detroit with a surface of 359 $km^2$ in [23] and 370 $km^2$ in [2]. In the end, we tried to get the most recent and reasonable data with the available resources, adopting a scientific approach.

## 3.2 Network creation

This section first describes the representation of transportation networks as unweighted undirected graphs, then discusses how to plot a given network, in order to provide a meaningful visualization of its main characteristics.

### 3.2.1 Unweighted and undirected network

During the development of our work, we based our static analysis tasks on a simple graph representation. For each city, we created an unweighted and undirected graph with the information about all the types of transportation combined. In particular, we created a network in a L-space topology. This means that each node represents a stop of a given mean of transportation. Two graph nodes (two stops) are connected by an edge if they are consecutive stops on a given route. The degree of a node in L-space is then the number of different nodes that one can reach within one segment (a graph edge). Given two nodes, the related shortest path length represents the number of intermediate stops[1] necessary for mutual reachability. We used the information inside *network_nodes.csv* for the nodes and those in *network_combined.csv* for the edges.

It is important to mention that during the preprocessing, all stops less than one meter apart were collapsed together and given a single id. Furthermore, since an edge can be common to several routes, the creation of an undirected graph with *networkx* python package allows to remove duplicates and it adds to the network just the first copy.

### 3.2.2 Visualizing the network

Once the network has been loaded and internally represented as a graph, its visualization is an important step, in order to get a first, though very high level view of its shape, density and variability. At the very beginning, we wanted to plot it over a map, but this turned out to be a tricky task. We explored options for a python implementation based on a package called Basemap, which requires the support of Anaconda. In our specific case, though adopting a Linux environment (usually considered as the most flexible/favourable one) with all the required setups, we failed in plotting the network over a map.

However, we managed to obtain an (almost) equivalent picture in two different ways:

- through one of the functions available in the *gtfspy* Python package [26], created by the research group which collected the data [21]
- with the functions available in *networkx* and *matplotlib* Python packages

[1] Path lengths in unweighted graph are measured in terms of number of edges, so the number of intermediate stops is actually given by the number of edges minus 1.

An example of the two different results are showed in Figure 3.4a and 3.4b, respectively. The solution with the gtfspy package (a) shows a version of the network with the map underneath and visible red nodes, while the other one (b) is the graph visualization without the map and with not nodes style. The first is useful in terms of representation of the network over a map and because it shows the density of nodes in the network. For example we can see the difference between the Helsinki network, where we cannot distinguish the nodes, and the Kuopio one 3.4c, where it is clear the the number of nodes is definitely smaller and the percentage of red in the Figure is reduced. However, this type of representation is worthless in terms of possibility to show properties of the network, since it does not allow modification to the visualization. For example, when we analyse the centrality measures, it is beneficial to highlight on the map the important nodes through different colours and sizes. On the other hand, the second visualization allows these kinds of manipulation and visualization of the network, but it loses the possibility of showing the map beneath, even though it keeps the coordinates between the stops. For these reasons, it is important to take both representation into account.

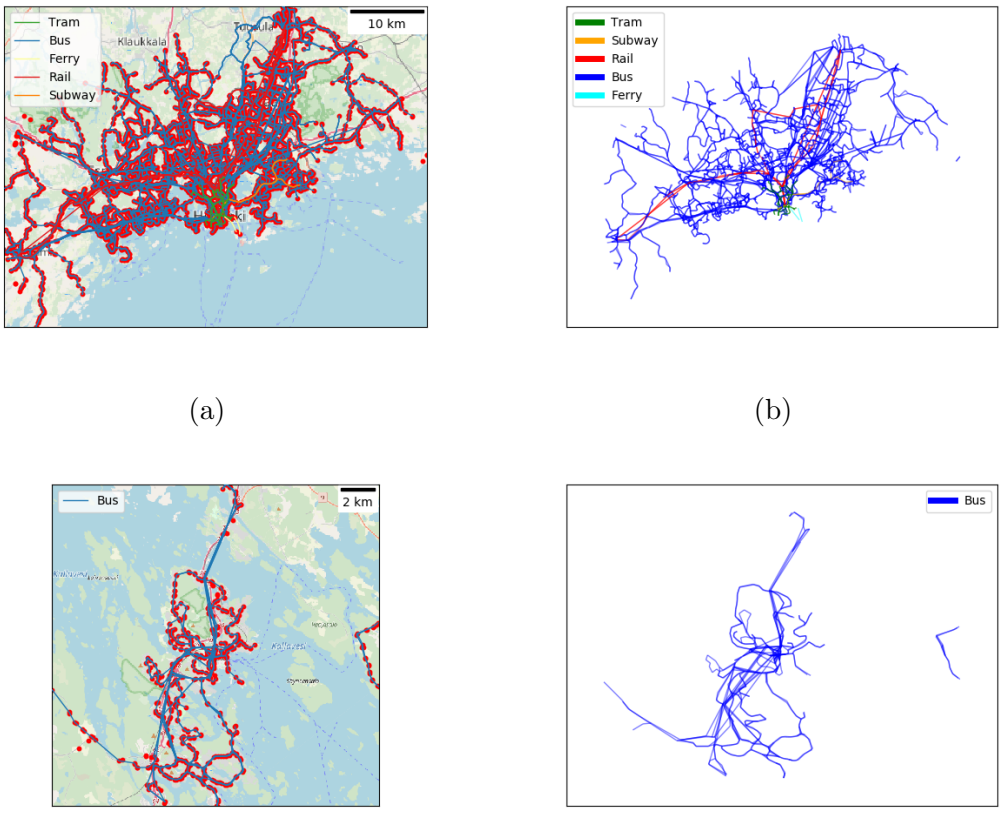



Figure 3.4: Different visualization of Helsinki and Kuopio PTNs

# Chapter 4

# Analysis methods and results

This chapter represents the core of this thesis, as it describes the analysis tasks done on the set of data described in the previous chapter (3). The following chapter goes more into details on the static analysis performed on the networks.

We describe a set of data analysis steps that we performed, starting form basic network measures (4.1) such as size, density, diameter, and clustering coefficient. These provide a very elementary and high-level information about the network properties of the PTNs in the datatset.

Going further, we analyse some other significant measures (4.2) helpful to understand the networks dynamic and topology. In particular, we chose to present discussions about assortativity, calculated as Pearson coefficient, average path length, average degree and degree distributions. As a matter of fact, this is a well known set of graph-based measures commonly used for the analysis of systems modeled as complex networks. We here discuss their meaning for the PTNs under analysis, and we use them for local characterizations of single cities, as well as for comparisons and global statistics.

Afterwards, we consider centrality measures (4.3), specifically four of the available ones: betweenness, closeness, degree and eigenvector. These measures help finding out hubs, if present, and better define the distribution of stations thorughout the PTNs.

We then present a distance analysis (4.4), which deals with pairwise node distances through breadth first search/visit of the graph, highlighting similarities and differences between cumulative distances of the stations and the Euclidean one. In addition, we present an exploration of the graphs starting from the geographical central node, which helps define reachability and discover the so-called peripheral

nodes. In particular, the latter can sometimes give information about the shape of the city, not a feature under analysis in this work, but might be interesting for future projects. Lastly, we present and comment the information about PTNs connected components, which are a crucial factor to take into account for this type of distance analysis.

The last section describes an analysis of the distribution of vehicle frequencies throughout a typical day (4.5). This work aims at describing more in depth the systems analysed and finding possible correlations between cities and their population to overall compare PTNs. In particular, we approach this analysis dividing the data for type of transport in order to be able to study the behaviour of each one separately.

In general, the dataset analysed in this work is bigger and more heterogeneous than most of the previously studied, and we agreed on the importance of representing our results at different levels:

- measure + city level (*e.g.*, frequency of bus vehicles for a specific city)
- city level (*e.g.*, betweenness centrality for a specific city)
- measure level among all cities (*e.g.*, frequency of tram vehicles among all cities)

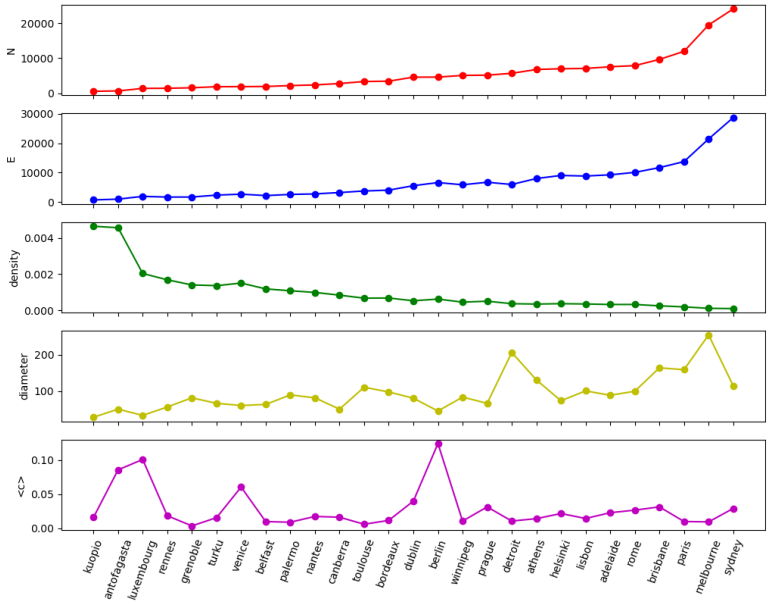
Depending on the measure analysed, some changes might be done to this model in order to better represent both individual and collective information.

## 4.1 Basic measures

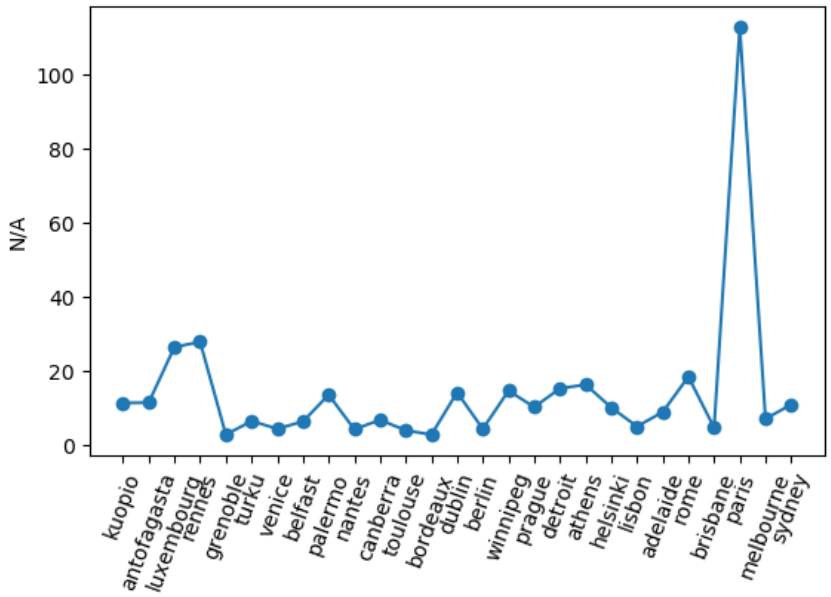
As a preliminary part of the network analysis, we started by computing some graph-based measurements, that provide useful hints for a better understanding of the dataset. According to most previous works on PTNs, mentioned in chapter 2, we chose the following measures: number of nodes, number of edges, density, diameter and average clustering coefficient.

The numbers of nodes and edges, as well as the density of the graph, provide a coarse measure of the PTN size and overall connectivity. A more detailed insight on connectivity can be obtained, as described later, by analysing node degrees and degree distribution statistics. The network diameter provides a global measure of mutual reachability, in terms of the longest shortest path between any two nodes. Finally, the clustering coefficient addresses another aspect of connectivity, by analysing the possibility to group network nodes in sets of closely related nodes.

For each city, we gathered all the previously mentioned measures. We plot them in Figure 4.1a, in order to provide a first, though rough, comparison among all PTNs on our dataset.




(a)




(b)

Figure 4.1: Basic network measures. (a) Five subplots representing: number of nodes, number of edges, density, diameter, average clustering coefficient. (b) Density of nodes per area. In both plots the cities are in ascending order of number of nodes.

### 4.1.1 Nodes, edges, and density

The two plots on number of nodes and edges show a strong similarity: the numbers of nodes range from a minimum of about 500 to 20 thousands, the numbers of edges are roughly proportional to nodes, by a factor of little more than 1. This result suggests these are very sparse networks, as confirmed by the low density ratio.

Overall, the density Figure can be a bit misleading, due to the quadratic denominator

$$\rho = \frac{2E}{N(N-1)}, \tag{4.1}$$

referring to the number of possible edges in a graph. As we can see from the plot (fig. 4.1a), given the proportionality edges to nodes, this quantity produces higher density values for smaller PTNs.

In order to better evaluate the concept of density in the analysed PTNs, we plotted in Figure 4.1b another measure, which expresses the node density per unit area. This method provides an indication of the distribution of nodes through the overall city area, feature of interest in this work.

### 4.1.2 Diameter

Concerning the diameter, measured in number of links (hops), it approximately follows the curves of nodes and edges, which means that the bigger the network, the longer the diameter, with exceptions, that are probably due to the shape and geographic characteristics of cities. Consider for instance Sydney and Melbourne, both of them in Australia, ranked among the top in terms of number of nodes and edges: Melbourne has the highest diameter (more than 200), whereas Sydney has an average value (about 100). When one looks at the shape of the two cities, it is easy to notice that Melbourne partly wraps around the Port Phillip Bay (which motivates the presence of a long shortest path), whereas Sidney has a more compact shape. Another city that pops up for its high diameter is Detroit, probably due to its polygonal shape and the orthogonal distribution of its streets.

### 4.1.3 Average clustering coefficient

Finally, the average clustering coefficient is another measure of local connectivity, partially orthogonal to other types of measures. It is partly related to the graph density, since a higher number of edges could lead to higher connectivity and, when

computed over a single node, it can be interpreted as the density of its neighbourhood.

The average clustering coefficient, defined as

$$\langle c \rangle = \frac{1}{N} \sum_{i \in G} c_i, \tag{4.2}$$

with $c_i$ presented in 2.5, tends to give higher scores to PTNs with better local connectivity, and a certain redundancy in available routes/paths: see for instance Berlin, Luxembourg, Antofagasta and Venice. As we can see from Figure 4.1a, the maximum value is 0.12 of Berlin, which means that all the PTNs have quite low values of average clustering coefficient. This result is partially in contrast to the normal behaviour of spatial networks [4], where the fact that closer nodes have a larger probability to be connected, usually leads to a large clustering coefficient.

However, as highlighted also in [11], we can see how for bus, subway and rail networks the outcome is quite different, showing low values of $\langle c \rangle$. In our case, the range is [0.0033-0.1240], though definitely higher than a $C_{ER} \sim 10^{-6} - 10^{-3}$ corresponding to a random ER graph with same parameters as the PNT one. In addition, the ratio $\langle C \rangle / C_{ER}$, which is also explicitly considered in the previous work [9], is consistently higher than 1 and presents a wide range [8.0-15647.6].

In general, having considered just the L-space topology, the density of the graph is quite low [$10^{-5}$-$10^{-3}$] and it is reasonable that the clustering coefficient does not reach high values. This conclusion is in line with previous works which have analysed and compared both L-space and P-space PTNs topology.

### 4.1.4 Outliers

Overall, we ave already observed that, though the plots in the Figure show a certain uniformity, is easy to notice "outliers" in each of the plots.

Let's consider for instance the density plot in fig. 4.1b, that shows the density of nodes vs the area of the city. It is clear from Figure that Paris is the most dense one. It's concentration of nodes it's definitely higher than all other cities. This is probably because, the area considered for the network is just around 100 km$^2$ with more than 10k nodes. The area is quite small, compared to other big cities. Even if we could also consider its geography, and maybe the availability of an old and well established "metro" and bus network, we could not give a sure answer on why this happens, apart from the logical fact that Paris is one of the biggest cities in Europe.

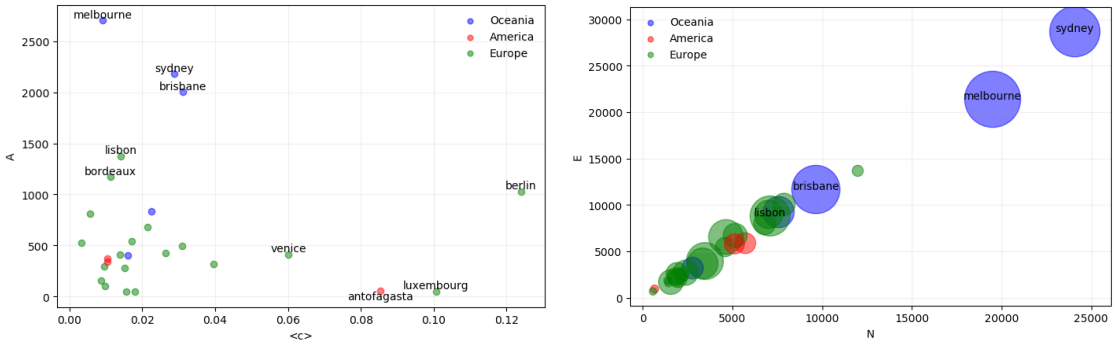
Other examples are Rennes and Luxembourg City. The first one has a density of node per area of 27.92, which is the second highest value, with a very high number

of nodes over the small area of just 50.4 km$^2$. The latter presents similar results, with a ratio of 26.54 and an area of 50.1 km$^2$. They both have a very round compact shape whereas Kuopio, a comparable city in terms of area (47.9 km$^2$) with a ratio of 11.46, less than half of the others, presents a elongated shape. On the other hand, Rennes is definitely the most populated of the three (349K) against the 119K and 118K of Luxembourg City and Kuopio, respectively.

As for the average clustering coefficient shown in Figure 4.1a, we find that Luxembourg City, again, deviates from the average trend with its 0.1 value of $\langle c \rangle$. In addition we find Antofagasta (0.09), Venice (0.06) and Berlin (0.12). The latter has a circular shape with a inland position, pretty similar to the one of Luxembourg City, even though they do not share any other similarities in terms of area, population or network measures. On the other hand, we have Antofagasta and Venice, which are both on the sea and have an elongated shape. However, also here, the two cities do not seem to share other features.

Lastly, if we look at the density plot in Figure 4.1a, there are two cities that really differ from all the others, which are Kuopio and Antofagasta. They are the only two PTNs in the dataset that have less than a thousand nodes and edges. They share a similar shape, narrow and elongated, and they both present only a bus transport network.

### 4.1.5 Correlating different measures




(a) (b)

Figure 4.2: Basic network measures correlation. (a) Area against average clustering coefficient with names for cities that have x or y values higher that a third of the respective maximum. (b) Number of edges against number of nodes. The city names showed have area higher than half of the maximum one.

As the previously described plots show one measure at a time, we also decided to represent the potential correlation between different measures by means of scatter

plots. In Figure 4.2a we compare area (y-axis) vs. clustering coefficient (x-axis): cities are represented through the dots of the scatter plots and coloured based on the continent they belong to.

In addition to single-measure plots against area, we also tried to correlate three different measures together in Figure 4.2b. For each city, the plot shows the correlation between number of nodes and edges, together with the information about the area, represented as the dimension of the dot. The first two measures are directly proportional, as expected. Moreover, we can see how the European cities are concentrated in the lower left side of the plot, whereas the biggest ones, mostly Australian, are in the middle-higher part of it. This means that also the area is quite correlated with the number of nodes and edges of the networks.

Of course, there are few exceptions to this trend, such as Paris, previously mentioned, which has a small area compared to the number of its nodes and edges, and Detroit, which, on the other hand, shows the opposite behaviour.

## 4.2 Additional measures

Next, we decided to expand the range of measures analysed in order to characterise some basic network properties such as small-world and scale free behaviour.

In accordance with some previous works analysed ([4], [10], [7], [11], [18], [14]), we added a few measures to those already collected. A summary of all the measures can be found in Figure 4.3. As one can see, the additional data include average degree $\langle k \rangle$, average assortativity $\langle r \rangle$ and average shortest path $\langle l \rangle$.

### 4.2.1 Assortativity

The assortativity, $r$, was calculated through the Pearson correlation coefficient (2.8). In general, positive values of $r$ indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degree. For all the cities analysed, the coefficient presents low positive values [0.09-0.42], with extreme values related to Grenoble and Kuopio, respectively, and 75% of the values below 0.26. This means that the networks are slightly assortative, given their positive values, but they do not show a strong similarity of connections with respect to the node degree. These results are in line with other related works, such as [11], [9], [13], [14], always concerning the L-space topology analysis.

| | N | E | A | P | d | <c> | ρ | <k> | r | <l> | <knn> | ln(N) | γ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| adelaide | 7,548 | 9,234 | 837.0 | 1,328,119 | 89 | 0.02 | 9.02 | 2.45 | 0.34 | 23.6 | 4.12 | 8.93 | 4.64 |
| antofagasta | 650 | 963 | 56.1 | 432,085 | 51 | 0.09 | 11.59 | 2.96 | 0.17 | 16.8 | 3.60 | 6.48 | 3.89 |
| athens | 6,768 | 7,978 | 412.0 | 3,154,152 | 131 | 0.01 | 16.43 | 2.36 | 0.27 | 33.5 | 3.45 | 8.82 | 5.01 |
| belfast | 1,917 | 2,180 | 297.0 | 626,760 | 64 | 0.01 | 6.45 | 2.27 | 0.19 | 24.9 | 3.00 | 7.56 | 5.24 |
| berlin | 4,601 | 6,600 | 1,030.0 | 3,556,792 | 46 | 0.12 | 4.47 | 2.87 | 0.24 | 13.6 | 5.05 | 8.43 | 4.03 |
| bordeaux | 3,435 | 4,026 | 1,172.0 | 957,383 | 98 | 0.01 | 2.93 | 2.34 | 0.16 | 32.8 | 2.90 | 8.14 | 5.87 |
| brisbane | 9,645 | 11,681 | 2,004.0 | 2,372,335 | 164 | 0.03 | 4.81 | 2.42 | 0.19 | 33.7 | 4.28 | 9.17 | 4.19 |
| canberra | 2,764 | 3,206 | 401.0 | 452,497 | 51 | 0.02 | 6.89 | 2.32 | 0.26 | 19.6 | 3.45 | 7.92 | 4.33 |
| detroit | 5,683 | 5,946 | 370.0 | 672,662 | 206 | 0.01 | 1.64 | 2.09 | 0.25 | 70.5 | 3.09 | 8.65 | 5.02 |
| dublin | 4,571 | 5,537 | 319.0 | 1,214,666 | 81 | 0.04 | 14.33 | 2.42 | 0.32 | 26.6 | 3.82 | 8.43 | 4.82 |
| grenoble | 1,547 | 1,679 | 523.0 | 524,853 | 82 | 0.00 | 2.96 | 2.17 | 0.09 | 26.9 | 2.49 | 7.34 | 4.92 |
| helsinki | 6,986 | 9,022 | 683.0 | 1,292,232 | 74 | 0.02 | 10.23 | 2.58 | 0.14 | 25.3 | 3.21 | 8.85 | 4.82 |
| kuopio | 549 | 699 | 47.9 | 118,667 | 29 | 0.02 | 11.46 | 2.55 | 0.42 | 10.6 | 4.35 | 6.31 | 3.97 |
| lisbon | 7,073 | 8,817 | 1,376.0 | 2,942,097 | 101 | 0.01 | 5.14 | 2.49 | 0.11 | 30.1 | 3.23 | 8.86 | 4.71 |
| luxembourg | 1,367 | 1,903 | 51.5 | 119,215 | 34 | 0.10 | 26.54 | 2.78 | 0.22 | 10.8 | 4.77 | 7.22 | 3.57 |
| melbourne | 19,493 | 21,434 | 2,705.0 | 4,870,388 | 254 | 0.01 | 7.21 | 2.20 | 0.22 | 75.3 | 3.23 | 9.88 | 4.94 |
| nantes | 2,353 | 2,743 | 538.0 | 669,843 | 82 | 0.02 | 4.37 | 2.33 | 0.15 | 28.4 | 2.85 | 7.76 | 4.89 |
| palermo | 2,176 | 2,559 | 159.0 | 852,454 | 90 | 0.01 | 13.69 | 2.35 | 0.13 | 30.1 | 3.13 | 7.69 | 5.01 |
| paris | 11,950 | 13,726 | 106.0 | 2,229,095 | 159 | 0.01 | 112.74 | 2.30 | 0.10 | 47.6 | 3.14 | 9.39 | 5.52 |
| prague | 5,147 | 6,714 | 496.0 | 1,308,632 | 67 | 0.03 | 10.38 | 2.61 | 0.15 | 23.8 | 3.38 | 8.55 | 4.66 |
| rennes | 1,407 | 1,670 | 50.4 | 349,759 | 57 | 0.02 | 27.92 | 2.37 | 0.14 | 20.6 | 2.73 | 7.25 | 5.04 |
| rome | 7,869 | 10,068 | 424.0 | 2,879,728 | 100 | 0.03 | 18.56 | 2.56 | 0.20 | 34.7 | 3.43 | 8.97 | 4.64 |
| sydney | 24,063 | 28,695 | 2,179.0 | 4,859,432 | 115 | 0.03 | 11.04 | 2.38 | 0.37 | 36.1 | 4.85 | 10.09 | 4.81 |
| toulouse | 3,329 | 3,734 | 812.0 | 1,010,593 | 111 | 0.01 | 4.10 | 2.24 | 0.12 | 42.9 | 2.63 | 8.11 | 5.10 |
| turku | 1,850 | 2,335 | 282.0 | 274,896 | 67 | 0.02 | 6.56 | 2.52 | 0.18 | 23.4 | 3.22 | 7.52 | 4.36 |
| venice | 1,874 | 2,647 | 413.0 | 636,244 | 61 | 0.06 | 4.54 | 2.82 | 0.26 | 20.5 | 4.46 | 7.54 | 4.13 |
| winnipeg | 5,079 | 5,846 | 344.0 | 808,419 | 84 | 0.01 | 14.76 | 2.30 | 0.23 | 29.1 | 2.96 | 8.53 | 4.92 |

Figure 4.3: Tableof network measures

### 4.2.2 Average path length

The average path length, $\langle l \rangle$, represents the average number of stops needed in order to go from one random stop to another. This value is especially important because it can help detect small-world properties. As a matter of fact, if $\langle l \rangle \sim \ln N$, then we can say that the network has small-world behaviour.

In the current analysis, we find that the values range between 10.58 of Kuopio and 75.29 of Melbourne. It is important to remind the reader that these values are calculated on unweighted and undirected graphs, where all the different types of transport networks are put together. Again, the 75% of the data are below 33.61, so the majority of the distribution is concentrated on low values. The plot of $\langle l \rangle$

vs. $\ln N$ (4.4), though showing some linear dependence (slope of 0.038), cannot be considered as a proportional relationship (the interpolation line does not cross the origin). This consideration together with the small values of $\langle c \rangle$ (4.3) do not support a small-world behaviour of the networks. All in all, this is reasonable and in line with the results found in other papers ([11], [9], [13], [14], [17]). In particular, this might depend on the topology considered, which in this case is the L-space one.

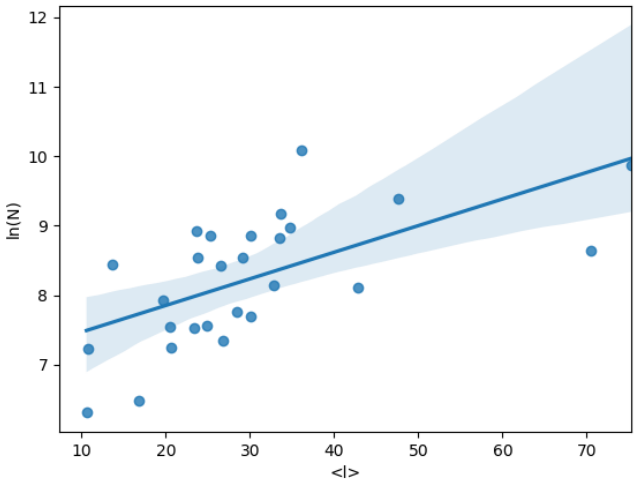



Figure 4.4: $\langle l \rangle \sim \ln N$ for small world analysis

### 4.2.3 Average degree

The average degree, $\langle k \rangle$, is the average number of edges per node in the graph, which in an undirected and unweighted network follows the simple formula

$$\langle k \rangle = \frac{2 * E}{N}. \tag{4.3}$$

Analysing this dataset, we found that this values range from a minimum of 2.09 (Detroit) to a maximum of 2.96 (Antofagasta). We analysed possible correlations between the average degree of a city and its area without finding any significant pattern. The plots created for this specific analysis can be found in the GitHub repository [27] in the directory *results/all/plots/stats/* under self-explaining file names.

In general, the fact that all $\langle k \rangle$ are above 2.0 is in line with the expectations. This is because in a PTN most of the nodes (stations) are usually connected with the previous and the next one. There are, of course, cases of the starting and terminal stations, but these are balanced out by those which are connected with more that

two other nodes. It is also important to highlight that, with a mean of 2.45, 75% of the $\langle k \rangle$ are below 2.55. This means that three quarters of the values are concentrated in the first half of the distribution and, therefore, very few cities present an high average degree. It is interesting to notice that these cities (Antofagasta, Berlin, Luxembourg and Venice) all have a different number of types of transport: 1, 5, 2 and 3 respectively, as it can be seen in 3.2b.

### 4.2.4 Degree distribution

The degree distribution, P(k), is often analysed to determine the model of the network. As presented in the work of Albert and Barabási, [28], there are three main class of models: random graphs, a variant of the Erdős-Rényi model, small-world models and scale-free ones. In particular, the latter focuses on network dynamics and tries to explain the sources of the power-law tails and other non-Poisson degree distributions that characterize a lot of real systems.

The degrees range from a minimum of 1, shared by all the cities, to a maximum of 28, shown in the city of Brisbane, Australia. In particular, more than 75% of the cities have a maximum degree equal or below 16, which is very close to the half of 28. The only cities above this threshold are Adelaide (21), Berlin (21), Brisbane (28), Luxembourg City (22), Sydney (20) and Venice (19), for a total of 6 cities over the 27 analysed. This means that most of the distributions are concentrated in the lower half of the range of values considered.

As we can see from fig. 4.5, we analysed the degree distributions using log-log plots, better solutions for fitting than semi-log ones, which we tried but are not presented in the final work. Points with $k = 1$ are peculiar since they represent routes' starts and ends. On the contrary, the rest of the distributions seems to show power law trend, represented by the formula

$$P(k) \sim k^{-\gamma}. \tag{4.4}$$

In order to have a better understanding of the distributions, we analysed both individual and collective behaviour, fitting city level curves and the average one in the summary plot. More precisely, the fitting line shown in Figure 4.5 is obtained averaging all the probability values for each $k$ and interpolating the resulting average distribution. Observed characteristic exponents $\gamma$ are between 3.57 and 5.87, with more than 25% of the cities above 4.34. The $\gamma$ value obtained from the average fitting process is 4.94. These results are quite far from the value $\gamma = 3$, characteristic of the Barabási-Albert model of evolving networks with preferential attachment [28].

As explained in [11], one can suppose that a corresponding model for transport network evolution should include several other effects. In fact, various models taking into account the effects of fitness, attractiveness, accelerated growth and aging of vertices [29], or deactivation of nodes [30, 31] lead to $\gamma$ from a wide range of values $\gamma \in [2, \infty)$. It is important to notice though, that the requirements for scale-freeness comprehend both very high $N$ and large range of $k$-values. PTNs and, more in general, spatially constrained networks do not meet these requirements. Therefore, they cannot be considered as scale-free networks, but they do have broad degree distributions, that, for the range of $k$ considered, are practically indistinguishable from power-law distributions.

Lastly, another important aspect to point out is the number of nodes with $k = 1$, which is definitely smaller that those with $k = 2$ (maximum probability reached for all the distributions). This means that most of the stops are directly connected to two others and that this is the typical behaviour. Moreover, we can see from Figure 4.5 that there are some nodes with degree higher than 10, which can be considered as network hubs, even though they are so few compared to the total number of nodes.

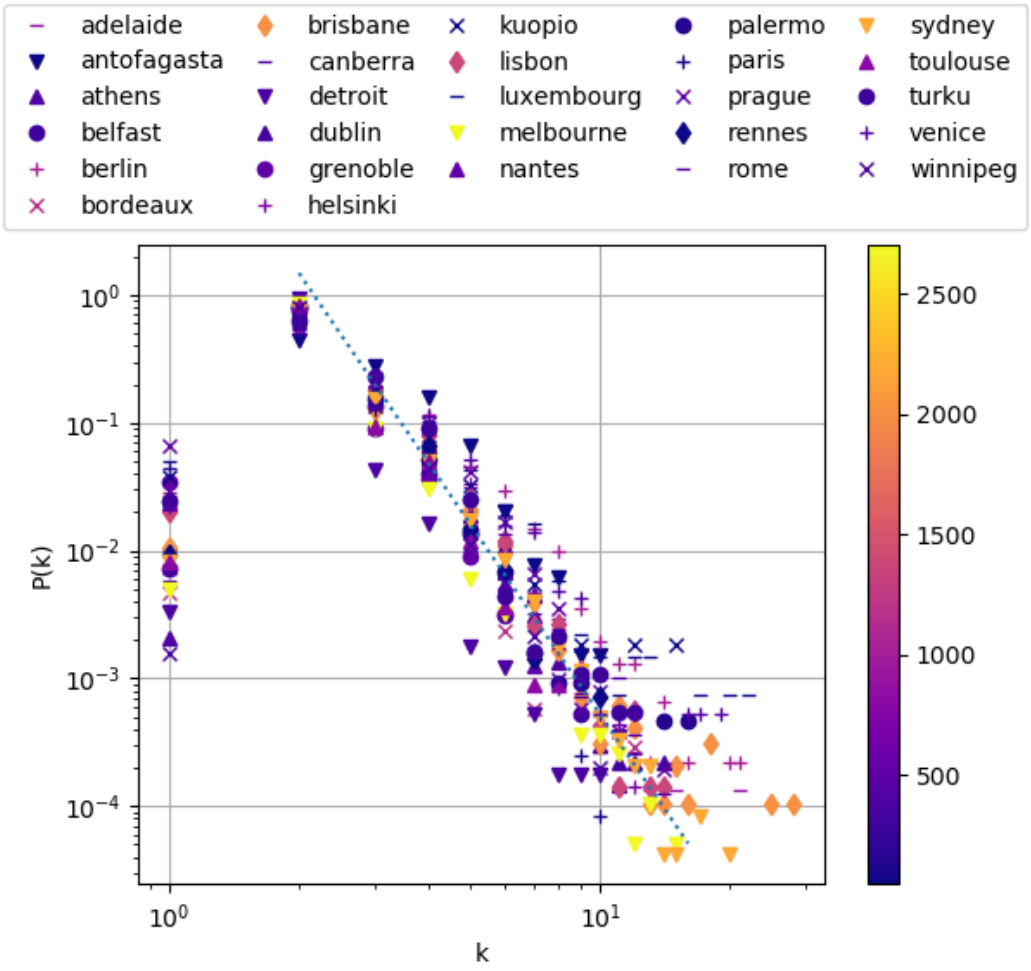



Figure 4.5: Degree distributions of all cities with colour based on their area in km$^2$

## 4.3 Centrality measures

In complex networks, the research of important nodes and edges is one of the key steps in order to analyse a graph. There are several ways of defining the importance of a network component. Usually, the centrality measures serve as a way to measure the importance of a node, even though they can be applied to edges too. *Importance* can be conceived in relation to a type of flow or transfer across the network. This allows centralities to be classified by the type of flow they consider important [32]. *Importance* can alternatively be conceived as involvement in the cohesiveness of the network. This allows centralities to be classified based on how they measure cohesiveness [33]. For example, when dealing with nodes, one can quantify their relevance measuring the number of connections, their role of bridge in the network, their closeness to other nodes and other etc. In this thesis, we decided to concentrate on the influence of nodes. In particular, we collected data on city level and plot them both locally and globally. As previously mentioned in the introduction to this chapter 4, we will present results and plots at different levels of detail. Among all the possible measures, we chose to compute the following four:

- betweenness centrality
- closeness centrality
- degree centrality
- eigenvector centrality.

For each measure we present a small section with comments on the work and results. In general, for each centrality, we offer two ways of visualisation: distribution and network-like. In addition, for each city, we gathered the information about all four distribution in one single plot. We tried to represent these information for all the cities together. However, we determined that it was not both feasible and useful to represent, since the plot would have been too crowded and difficult to interpret. Therefore, we decided to plot all the distributions for one measure at a time for all cities together, adding the information about area or population through the use of colorbars.

We now introduce one measure at a time, sorted in descending order by level of importance for the analysis of PTNs (betweenness is the most important). In addition, throughout the section we offer a couple of comparison between measures.

### 4.3.1 Betweenness centrality

The betweenness centrality of a vertex is determined by its ability to provide a path between separated regions of the network[4]. It is defined by the quantity

$$g(i) = \sum_{s \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}, \tag{4.5}$$

where $\sigma_{st}$ is the number of shortest paths going from $s$ to $t$ and $\sigma_{st}(i)$ is the number of shortest paths going from $s$ to $t$ through the node $i$.

In previous works on PTNs [11, 9, 12, 15, 17, 4], betweenness centrality is considered the most significant one. Its importance is underlined by the criterion of computation, based on the number of shortest paths passing through each node. This is significant when dealing with spatial networks, because of their construction and topology. Its relevance is confirmed by our findings, that place betweenness and closeness above the other two in those terms.

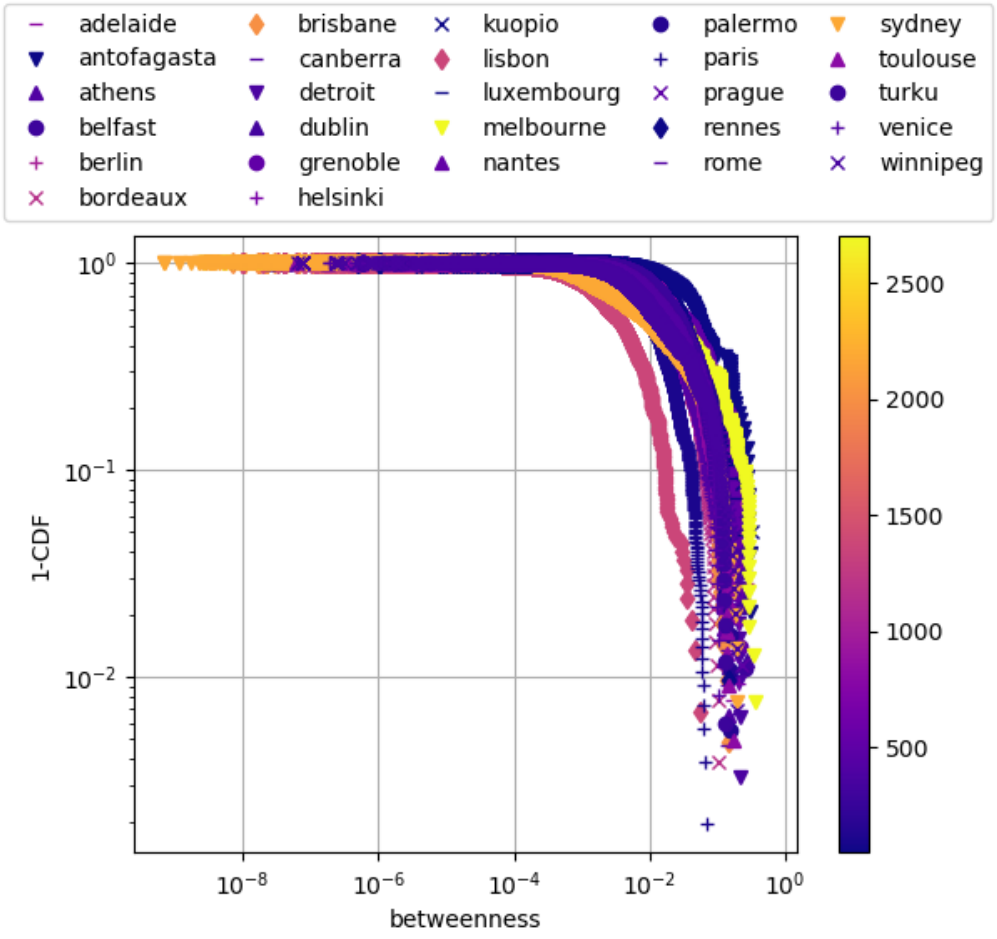



Figure 4.6: Betweenness centrality (g(i)) complementary cumulative distribution with colormap based on the AREA of the cities.

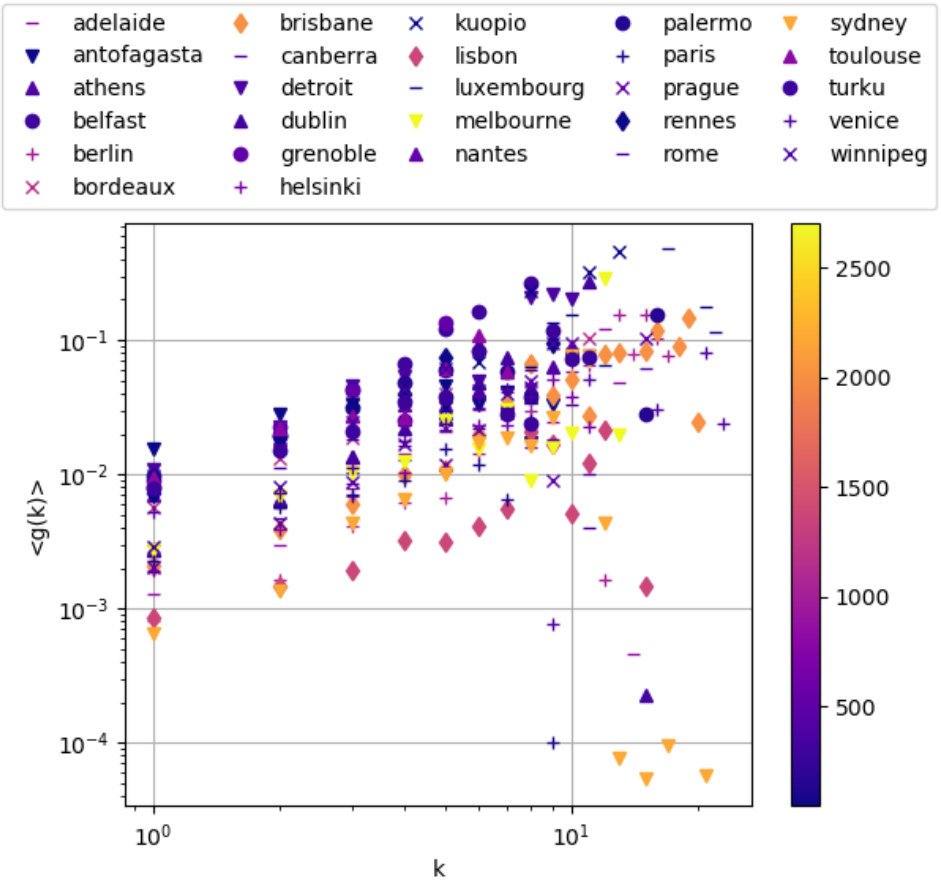
In Figure 4.6, we present the complementary cumulative distribution function of all the cities together with the colour depending on the area of the city. It is clear that the cities have similar curves, all following an exponential trend. The drop for some cities for very high degrees it's probably due to "statistical noise", since there are so few high-degree nodes. In addition, for each city, we created single plots showing the 1-CDF and fit line together (available in the repository [27]), adding in the legend the information about fitting parameters. These results are available for all the cities in the Table4.8, where the $\lambda$ column displays the exponential fitting parameter for each distribution.

Due to the nature of the betweenness centrality, in accordance with some previous works [11, 9], we analysed the correlation between the average $g(k)$ with $k$. In Figure 4.7a, we can see this relationship for all the cities together. Again, the color is related to the area of the city. There are some outliers in the bottom right part of the plot, probably because there are few nodes with high degree, that have high values of betweenness. Even with these outliers, we can see how most of the curves follow a power law

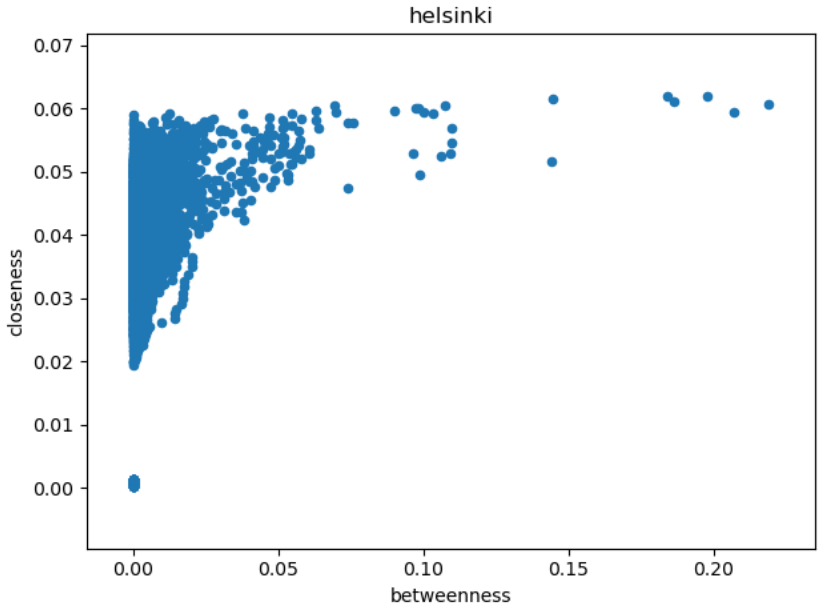
$$\langle g(k) \rangle \sim k^{\eta}, \tag{4.6}$$

where the fitting parameters are presented in the Table4.8. Comparing our results to those of [11], we find that the trend follows in both studies a power low. However, our fitting parameters do not show such a specific pattern as they do in the previous work on Polish cities. They plotted $\eta$ against $N$ and found that $\eta$ is getting closer to 2 for large networks, with a general increasing trend between the two measures. In our work, though, we did not find such strong correlation. Indeed, the two measures are negatively correlated, with a Pearson coefficient of -0.52. Furthermore, a scatter plot confirmed our hypothesis, not showing any particular trend. These results are probably due to the different nature of the PTNs considered in this thesis. The previous papers analysed 22 polish cities, which migh share more features than our 27 cities from around the world. In addition, we only analysed area and population feature, when one could introduce information about morphology and cultural differencies, which may influence the city PTN.

For a more detailed visualization, in the repository [27], we provide single plots of the correlation with the corresponding fitting line for each city. Not all the curves are well fitted by a line. In particular, scanning the values of $\eta$ (4.8) for all the cities, we see that Sydney presents a value of -0.911. If we take a look at the specific plot, we can see how the process of fitting the line is biased by a few outliers, also visible in fig. 4.7a. To points with high values of degree ($> 10$) correspond very low values of $\langle g(k) \rangle$.




(a)




(b)

Figure 4.7: (a) Average betweenness centrality against degree. The colours and markers depend on the area of each city. (b)Betweenness centrality vs closeness centrality for the city of Helsinki.

| City | max(g(i)) | <g(i)> | λ | max(g(k)) | <g(k)> | η |
|---|---|---|---|---|---|---|
| adelaide | 0.161 | 0.003 | -88.025 | 0.120 | 0.032 | 0.921 |
| antofagasta | 0.329 | 0.024 | -20.146 | 0.045 | 0.029 | 0.267 |
| athens | 0.205 | 0.005 | -77.589 | 0.091 | 0.032 | 1.870 |
| belfast | 0.266 | 0.012 | -35.050 | 0.266 | 0.086 | 2.384 |
| berlin | 0.359 | 0.003 | -48.670 | 0.156 | 0.046 | 2.098 |
| bordeaux | 0.109 | 0.008 | -73.321 | 0.103 | 0.037 | 1.723 |
| brisbane | 0.147 | 0.003 | -80.744 | 0.145 | 0.047 | 1.526 |
| canberra | 0.154 | 0.006 | -69.535 | 0.115 | 0.032 | 1.004 |
| detroit | 0.227 | 0.012 | -54.516 | 0.220 | 0.079 | 1.545 |
| dublin | 0.283 | 0.005 | -52.217 | 0.271 | 0.049 | 0.775 |
| grenoble | 0.255 | 0.012 | -34.316 | 0.135 | 0.039 | 1.974 |
| helsinki | 0.219 | 0.003 | -79.250 | 0.051 | 0.020 | 1.757 |
| kuopio | 0.451 | 0.016 | -25.574 | 0.451 | 0.120 | 1.981 |
| lisbon | 0.056 | 0.001 | -315.379 | 0.021 | 0.007 | 1.246 |
| luxembourg | 0.478 | 0.007 | -34.965 | 0.478 | 0.088 | 1.910 |
| melbourne | 0.525 | 0.004 | -42.178 | 0.286 | 0.037 | 1.358 |
| nantes | 0.150 | 0.010 | -49.664 | 0.092 | 0.031 | 1.778 |
| palermo | 0.162 | 0.013 | -47.041 | 0.153 | 0.050 | 0.633 |
| paris | 0.072 | 0.003 | -143.487 | 0.025 | 0.008 | -0.246 |
| prague | 0.243 | 0.004 | -67.469 | 0.103 | 0.029 | 1.842 |
| rennes | 0.163 | 0.012 | -53.101 | 0.094 | 0.044 | 1.652 |
| rome | 0.144 | 0.004 | -85.388 | 0.061 | 0.019 | 1.312 |
| sydney | 0.235 | 0.001 | -95.565 | 0.077 | 0.015 | -0.911 |
| toulouse | 0.193 | 0.012 | -51.185 | 0.108 | 0.039 | 1.525 |
| turku | 0.129 | 0.012 | -51.417 | 0.119 | 0.043 | 1.209 |
| venice | 0.123 | 0.008 | -59.868 | 0.081 | 0.027 | 0.767 |
| winnipeg | 0.194 | 0.006 | -59.914 | 0.094 | 0.036 | 1.868 |

Figure 4.8: Tableof betweenness centrality measures for node and degree correlation. The second and fifth columns presents maximum results for $g(i)$ and $g(k)$, respectively. Third and sixth columns show average values of node betweenness and degree betweenness, whereas $\lambda$ and $\eta$ columns show fitting parameters.

### 4.3.2 Closeness centrality

The closeness centrality is a measure of how close is a node, on average, to all other nodes. The basic concept is that the more central a node is, the closer it is to all the others. It is defined as the inverse of the sum of the length of the shortest paths between the considered node and all others in the graph,

$$C_c(i) = \frac{1}{\sum_j d_{ij}}, \tag{4.7}$$

where $d_{ij}$ is the length of shortest path between $i$ and $j$, *i.e.*, their distance.

However, when talking about closeness, people usually consider its normalized value, representing the average length of the shortest paths instead of their sum. The latter, multiplied by N-1, is computed as presented in 2.9. In this work, we adopt this last formula to compute the measure, because it allows to compare nodes of graphs of different sizes. Even though the computation does not directly work for networks with disconnected components where for some pairs $d_{ij} = \infty$, the algorithm adopted by the networkx function [34] computes the closeness centrality for each connected part separately.

As for the betwenneess, we computed the complementary cumulative distribution function for all the PTNs, saving the results both separately and altogether. In Figure 4.9, we present the 1-CDF log-log plot of all the cities together. It is important to notice that in this case the colours are based on the information about the population, instead of the area (4.6). As before, all distributions have a similar and exponential trend.

In addition, as mentioned in the introduction to this section 4.3, we provide a visual representation of the PTNs and each centrality measure. We chose to plot only the 20% most significant nodes, in order to give a clearer representation also for the bigger PTNs. In Figure 4.10, we compare two network visualization of Berlin's PTN. The first one, 4.10a, representing the betweenness centrality, highlights a dozen of hubs in the city and a lot of smaller ones. In particular, we can see that the distribution of the *important* nodes is well balanced throughout the network, with the most significant one around the central area. These information together with the round shape of the city let us assume that the PTN is well constructed. The second picture, 4.10b, represents the closeness centrality information. One can definitely spot the difference between the two: the distribution here is more clustered, with a big one in the center; the colours are brighter, because the values are better spread throughout the range of values, even though the latter is smaller than the on of the betweenness. In particular the values here range from 0 to 0.125 more or less, whereas for the betweenness the upper value is around 0.35.

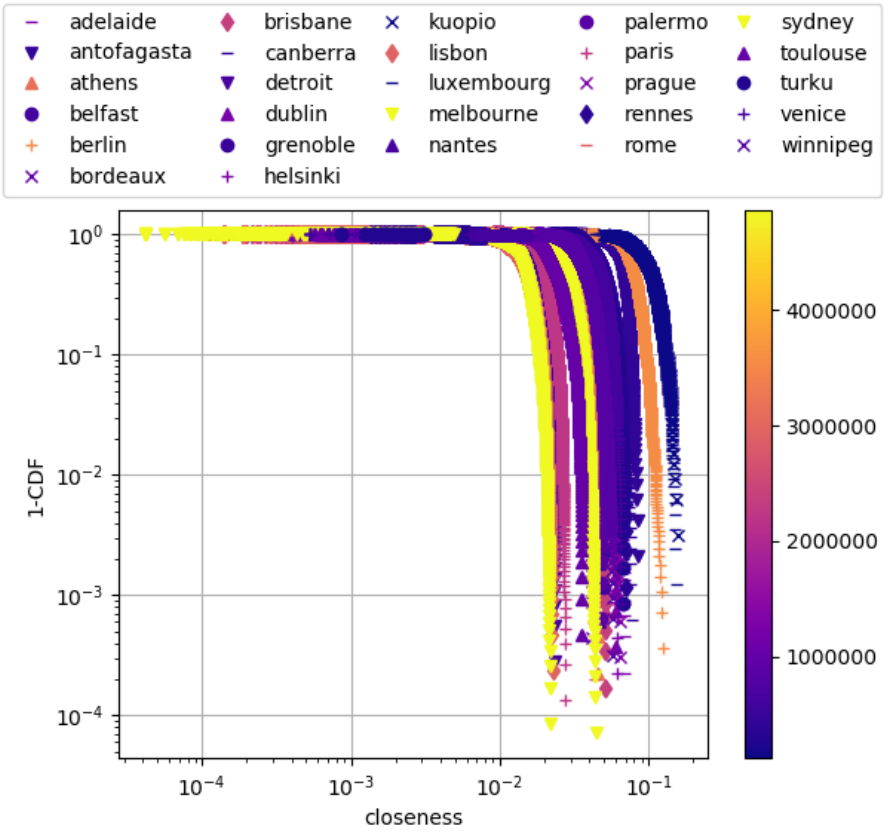
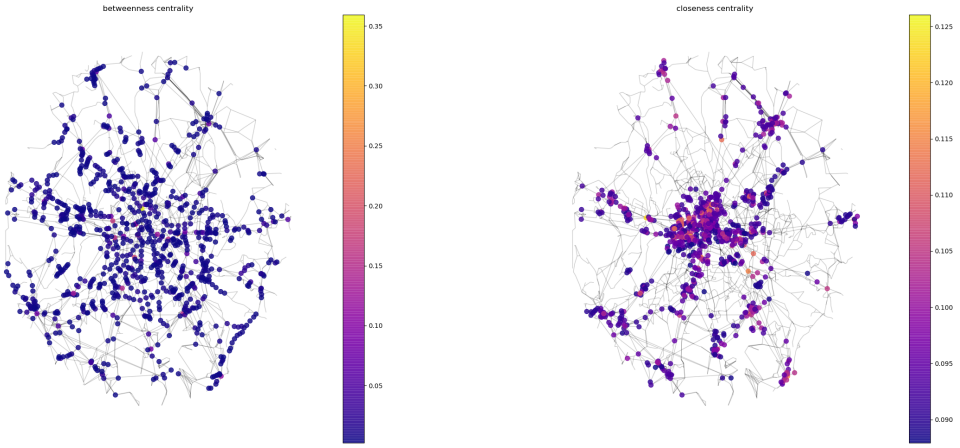



Figure 4.9: Closeness centrality ($C_c(i)$) complementary cumulative distribution with colormap based on the POPULATION of the cities

To further the comparison between the betweenness and closeness, in Figure 4.7b we report the scatter plot for the city of Helsinki as example. From the result, we can say that, as expected, high values of betweenness correspond to high values of closeness (high values for the range considered). Concerning low values of betweenness, the situation changes. From 0 to 0.05 of betweenness we find quite a wide range of closeness values, from 0.02 to 0.06. This happens because some nodes might have a short distance from many others (so be close to them), but they do not serve as bridges (low betweenness). In general, it is expected to have many nodes with low values of betweenness, even though they might have a central/close position to many others. We still find few nodes with values close to 0 for both measures, that probably represent peripheral nodes. In conclusion, we can say that the betweenness centrality helps understand if and where the hubs are located in the network and it gives a visual understanding of the connectivity of the graph. On the other hand, the closeness highlights the areas of the PTN with a high concentration of nodes close to each other.




(a) (b)

Figure 4.10: (a) Betweenness and (b) closeness centralities of Berlin's PTN. The colours depend on the value of the centrality considered. The brighter the colour, the higher the value for the node.

### 4.3.3 Degree centrality

Degree is a simple centrality measure that counts how many neighbours a node has. The networkx function computes for each node of the graph the fraction of nodes it is connected to [35]. In particular, the values are normalized by dividing them by the maximum possible degree in a simple graph N-1.

Even though we have already presented the analysis on degree distribution and average degree, we wanted to offer an analysis from a different perspective. In particular, this approach allows us to compare the degree, as measure of importance of a node, with others, first among all the eigenvector centrality. Indeed, these two give a local perspective on the centrality features, whereas the betweenness and closeness, taking into account shortest paths, give more a global view.

Since degree and eigenvector centrality give very similar results in this context, we postpone the discussion about comparison in the next subsection. In particular, we approach the degree-eigenvector correspondence (fig. 4.13) and the betweenness-eigenvector differences (fig. 4.11).

### 4.3.4 Eigenvector centrality

The concept behind the eigenvector centrality is that important nodes are connected to other important nodes. It has high values for nodes that have many neighbours which, in turn, have many neighbours. It can be interpreted as a measure of the influence of a node in the graph.

In general, these two measures give similar results just in a different range of values. They analyse similar properties of the nodes and, in the case of spatial networks, output similar results. Even though they are not the most used centrality measures in this type of networks, we decided to include them in order to give a complete analysis.

In Figure 4.12, we offer a comparison between the two complementary cumulative distributions. It is clear that there are strong similarities between the two plots, whereas they are very different from the betweenness one (4.6) and the closeness one (4.9). On the other hand, the latters have similar trends represented by an exponential function. In the case of degree and eigenvector, the trend resembles one of a power law.

We decided not to average the points that share the same centrality value to underline the expected behaviour that many nodes share low values of the centrality, whereas high values of centrality are more sparse. In addition, the information about the area allows us to underline the differences between the cities. In particular, it is clear that bigger cities tend to be on the left side of the plot. This is due to the fact that the values are normalized by N-1 and, therefore, bigger networks will have a smaller x-range overall. Moreover, the bigger the area the wider will be the length of the distribution, because these cities have more nodes. Indeed, the direct relation betweeen area and dimension of the PTN is confirmed for most of the cities.

In the Figure 4.13, we offer a comparison based on network visualization of the two measures. As done before for betweenness and closeness, we plot the network with 20% of significant central nodes, based on degree (fig. 4.13a) and on eigenvector (fig. 4.13b). It is interesting to notice that the two representations look exaclty the same, apart from the range of values of the colormap. In particular, we see how the first one ranges approximately from 0.001 to 0.0045, whereas the second one from 0.2 to 0.09. Comparing these plots to the one of betweenness, shown in fig. 4.10a, we can see how the high centrality nodes are better distributed throughout different part of the network, whereas those of 4.10a are concentrated in the central area. The closeness plot (4.10b) offer a different comparison. The nodes are more clustered and, in general, share high values of the centrality measure. On the contrary, the

majority of nodes in Figure 4.13 share lower values of centrality and they do not form such visible clusters, even though the central area has a higher concentration than the others.

As anticipated in the degree centrality section, we now discuss the comparison between eigenvector centrality and betweenness centrality. At a first glace, we can see that the majority of the nodes are concentrated in the left part of the plot. More specifically, up to 0.10 betwenness we find nodes with eigenvector values from 0 to 0.05 (almost maximum one). However, most nodes with low betweenness have low eigenvector too, whereas only a few ones have high eigenvector values. This means that hardly any node with low betweenness and many neighbours tend to be connected with other nodes with many connections as well. On the other hand, the right part of the plot is fairly sparse. Nodes with high betweenness seems to have average or low eigenvector values. This means that bridges nodes do not seem to be connected to other nodes with many neighbours.

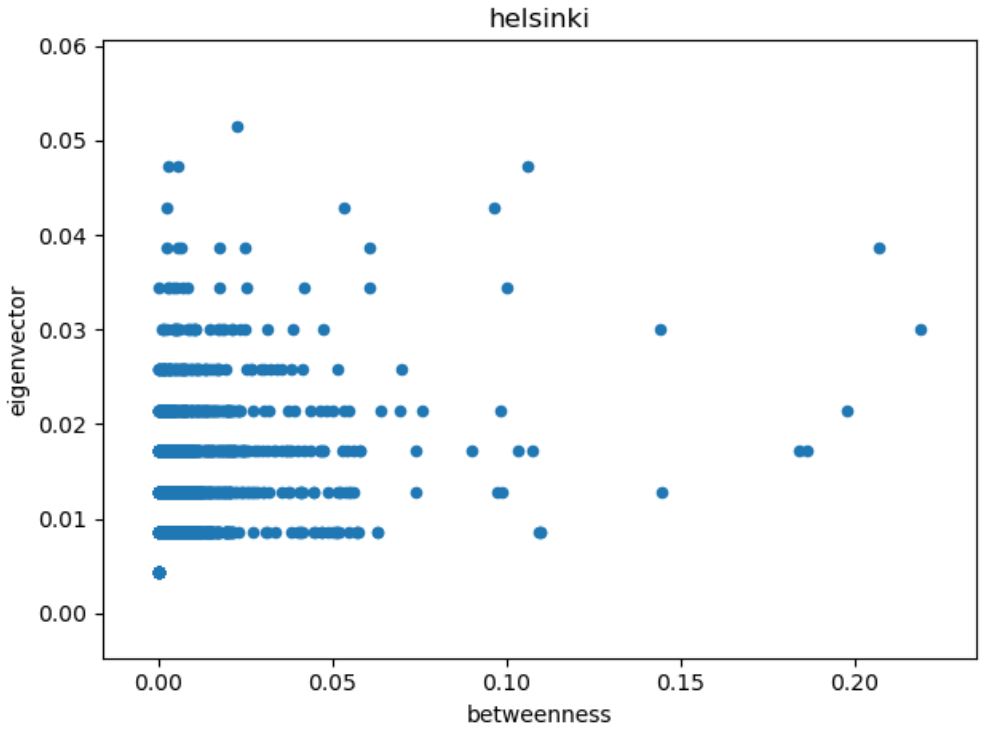
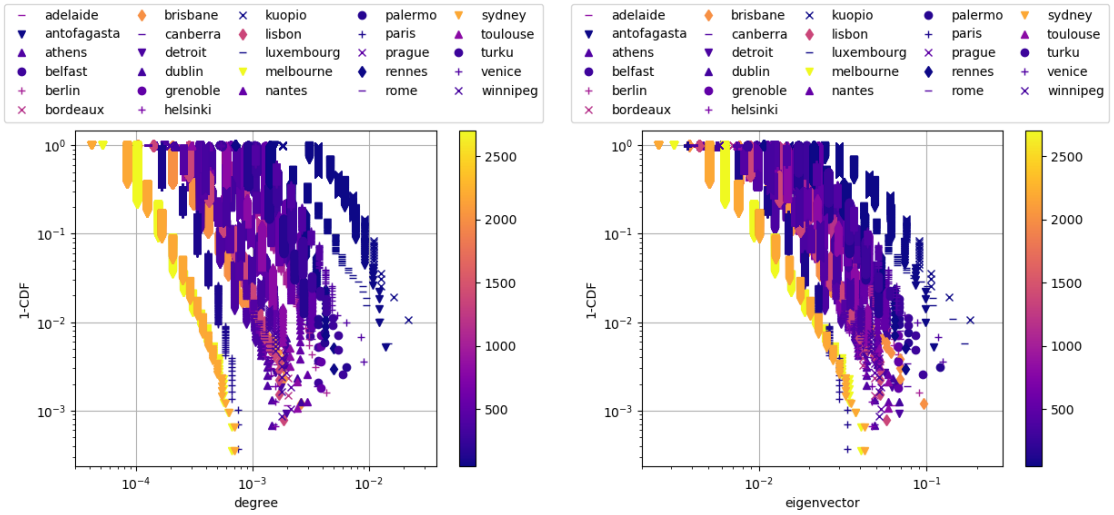
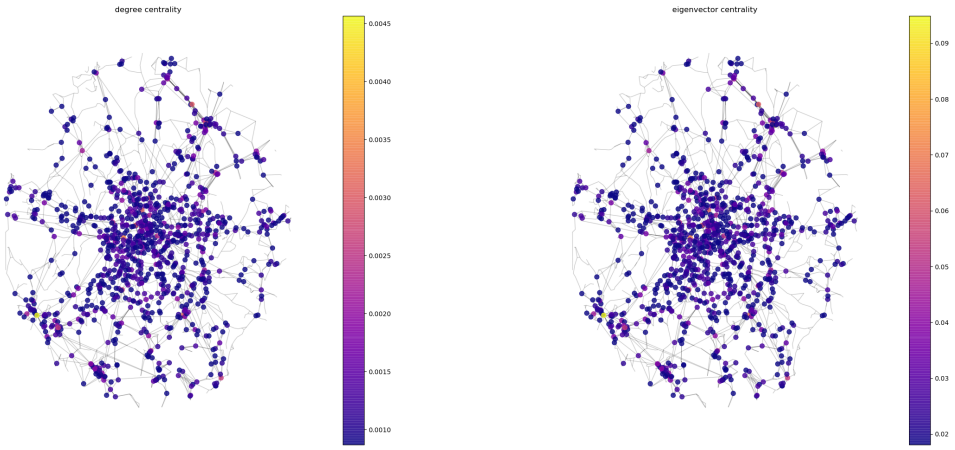



Figure 4.11: Betweenness centrality vs eigenvector centrality for the city of Helsinki.




(a) (b)

Figure 4.12: (a) Degree and (b) eigenvector centralities 1-CDF log-log plots of all the cities. The colours are based on the city area




(a) (b)

Figure 4.13: (a) Degree and (b) eigenvector centralities network representations for the Berlin's PTN. The colours in (a) and (b) depend on the area of the city (blue for small area and yellow for big one). In (c) and (d) the colour of the nodes depends for (c) on the degree value and on (d) on the eigenvector centrality one.

## 4.4 Distance analysis

When dealing with spatial networks, computing shortest paths is a key step in evaluating and characterizing trips, as path lengths are considered as relevant components of inter-node reachability analysis. Given any two nodes, for which we aim at studying their mutual reachability, shortest paths are the obvious solution to most problems of best path computation. There can be different ways, and related algorithms, in order to measure path lengths:

- Breadth-First Search considers path lengths in terms of number of edges
- Dijkstra and Bellman-Ford algorithms work on weighted graphs, with edge weights given by either distance or time, based on the desired optimization criterion

Defining a clear, unanimously accepted and simple path cost/weight is an uneasy task in PTNs. Time, *i.e.*, trip duration, is an obviously important aspect to be minimized, but one could also evaluate cost, comfort, length, number of transportation means and stops/changes, timed schedules, transit times, as well as other factors. Furthermore, one could consider the graph as dynamically weighted, in order to take into account factors that change on a daily, weekly, or even monthly/seasonal base.

In this section we limit our analysis to static graphs, we do not attempt to evaluate trip duration, as we assume it related to the traveled distance. We compare travel distances to Euclidean distances, as a way to take into account geographic and territorial aspects and shapes of transport lines: the closer path lengths are to Euclidean distances, the higher the chance of finding good, almost straight, transport paths. So part of our analysis was oriented to study probability distributions of path lengths with respect to Euclidean distances.

Concerning path optimality, we compared pure path length minimization, computed by the Dijkstra algorithm, to minimizing the number of graph edges, so using a BFS traversal. Though with general weighted graphs BFS does not guarantee shortest paths, it can be optimal, or nearly optimal, whenever edge weights are rather uniform. In our graphs, edges represent trips between stops, which could be considered, under certain conditions, as a rough estimation of trip duration, and they are often used by travellers/commuters as an empiric measure of a trip length. We thus use BFS measures as our primary optimum path evaluation strategy, while providing a comparison to shortest paths computed by the Dijkstra algorithm.

In order to properly visualize some of the collected measures, it became essential to create groups of cities simply to declutter the plots and facilitate the reading of

the results. We decided that a manual clustering was the best option for our work. In particular, we divided the 27 cities in 5 groups based on their area:

- **0-100 km$^2$**: 4 cities (Antofagasta, Kuopio, Luxembourg and Rennes)
- **100-300 km$^2$**: 4 cities (Belfast, Palermo, Paris and Turku)
- **300-500 km$^2$**: 8 cities (Athens, Camberra, Detroit, Dublin, Prague, Rome, Venice and Winnipeg)
- **500-1000 km$^2$**: 5 cities (Adelaide, Grenoble, Helsinki, Nantes and Toulouse)
- **1000+ km$^2$**: 6 cities with area bigger than (Berlin, Bordeaux, Brisbane, Lisbon, Melbourne and Sydney)

In the following, we'll first address distance analysis at city level, so within nodes of a given city, then we will make a comparison among cities, based on overall statistics, such as average measures, counters etc.

### 4.4.1 City level

In this part, we explain the procedure and results obtained on city level, with the aim of studying city level probability distributions of distances, as well as an attempt to characterize central and peripheral nodes. We first analyze reachability given a starting point, as a means to characterize the overall expected distances traveled when moving within PTNs.

So first of all, we decided to explore two different criteria for choosing the starting node:

- random starting node
- central node

For each city we calculated its geographical barycenter and then record the nearest node in the graph, defining it as the central node of the graph. Furthermore, just for capital cities, we manually selected a (well known) central POI (point of interest). For most of the capitals we selected the main railway station. However, in cities with several stations we either chose one (*e.g.*, Lisbon) or we selected another POI (*e.g.*, Paris, where we chose the metro station Chatelet).

When starting from a random node, we chose it from the biggest connected component, in order to reach as many nodes as possible. In every case, we proceeded by applying a given graph visit algorithm, repeated for each of the selected starting points.

As already discussed in 4.4, we chose BFS visits as minimum hop paths are a good compromise for shortest paths in PTNs. During the visit of each node, we thus recorded its distance from the source node both in terms of hops and of distance in kilometers, viewed as the sum of all previous edges. This type of distance (in km) is referenced in the plots as"bfs distance". At the same time, we calculated its Euclidean distance from the source node, referenced as "eu/Euclidean distance". The ratio bfs distance vs. the Euclidean one gives us a clear indication of how much the considered path is far from a simple straight connection.

#### Random source nodes

We repeated the process presented above 20 times with different random source nodes collecting for each city bfs and Euclidean distances. Then, we put together all the results divided by type of measure to plot their distributions.

In Figure 4.14 we present an example of this type of plot, a probability distribution, for the city of Adelaide. As this kind of plot requires a discretized X axis, we empirically chose a number of bins given by the square root of the number of reached nodes (the size of the connected component the starting node belongs to), then divided by ten. We observed this to be a good compromise, which gave accepTablesolutions for all the cities. In the specific case of Adelaide, we can see how the results confirm the prediction. The Euclidean curve provides a distribution of node distances, which is obviously more pointed and on the left, indicating that the majority of nodes has a distance of less that 20, averaged at nearly 17 kilometers. On the other hand, the bfs curve is softer and it reaches higher values of distances, confirming the fact that the distribution is broader and lower than the Euclidean one. The ratio between mean values (31 kilometers to 17), as well as the shape of distribution curves, shows that bfs distances are nearly double than the Euclidean ones, confirming that paths are on average fare from the straight linear ones, probably related to the city area and territorial characteristics.

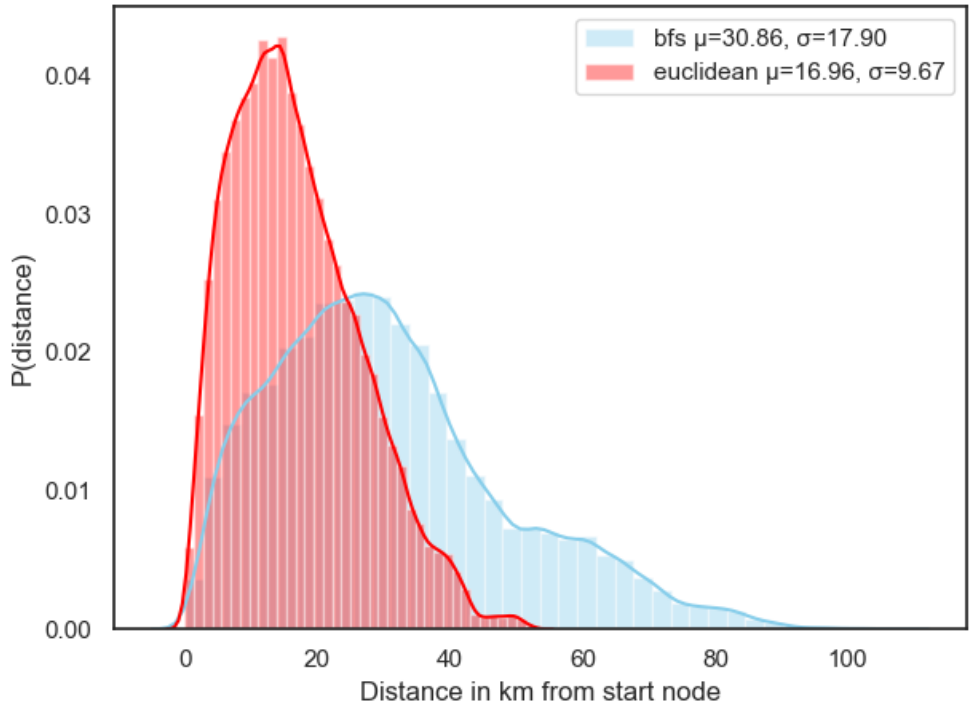



Figure 4.14: Distances distributions for the city of Adelaide. The red curve represent the Euclidean distances and the blue one the bfs one (explanation on how they were calculated can be found at 4.4.1.

### Central source nodes

Whereas the previous section studied reachability by a randomly selected starting point, we repeat here the process by starting from a conventionally accepted important node, a central one. Assuming that, on average, central nodes are well connected and reachable, we studied their reachability from two sets of nodes, those in the central area and in the peripheral one: the motivation for this choice is based on analyzing whether the center is mostly well connected to a central city area, or to suburban areas as well. We thus decided to analyse the distributions of close and far nodes starting from the central point of the network. For each city, we selected a central node as explained in 4.4.1. After collecting the information about the distances from the source node, we created two subsets, one for the close nodes and the other for the far ones. In particular, we considered a node to be near the source one if it had a bfs distance at most equal to a quarter of the maximum Euclidean one. On the other hand, we defined far nodes, those which had a bfs distance of more than half the maximum Euclidean one.

In Figure 4.15, we can see the specific case of Melbourne. Here we used a binning criterion based on the square root of the number of nodes considered in the

distribution. In the plot we notice how the result follows the expectations, having the close nodes distribution on the left and the far one on the right. In particular, the first one has a narrow and pointed shape with many small bins, whereas the second one have a broader and lower shape with wider bins. These differences highlight the fact that the distribution of close nodes is quite dense and there is a higher probability of finding a node with a low distance from the center, whereas the second distribution is quite sparse.

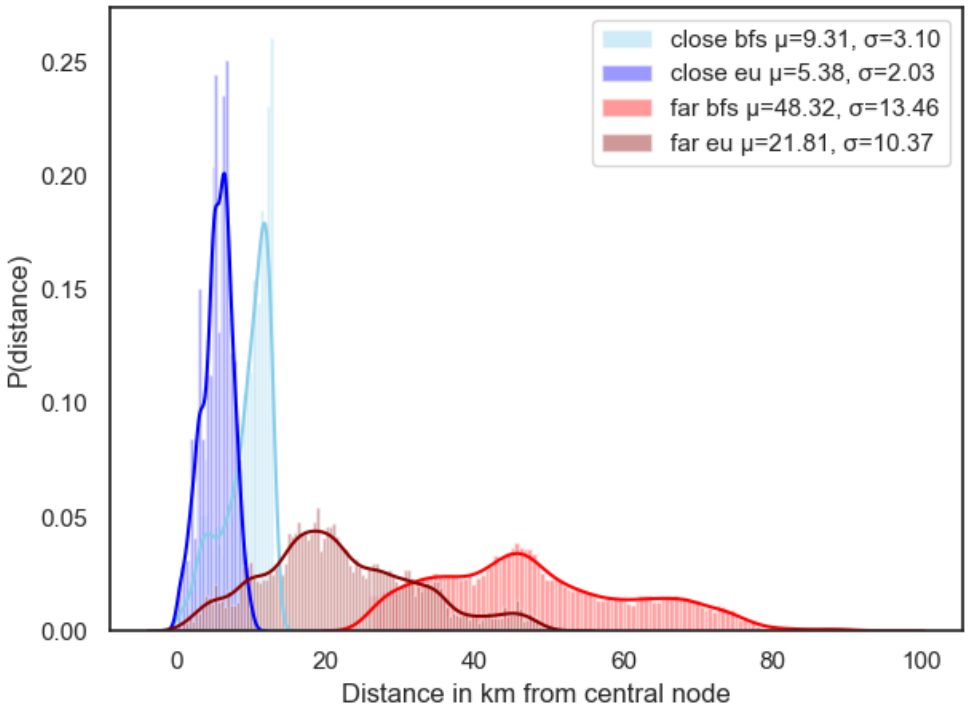



Figure 4.15: Close and far nodes distances distributions for the city of Melbourne. The blue curves represent the nodes that have a bfs distance $< 1/4$ maximum Euclidean distance, whereas the red curves those which have a bfs distance of $> 1/2$ maximum Euclidean one. The shades of colours differentiate the bfs distances from the Euclidean ones.

We followed the same process and collected for the capital cities the two subsets (close and far nodes) after visiting the graph starting from the POI central node. The information available in the repository [27] show that for most of the cities, *e.g.*, Berlin, Paris, Prague, the two distribution plots (for geographical centre and POI one comparable to the one in fig. 4.15) are very similar: intuitively this is because the two nodes are likely close to each other and have almost the same reachability. For other cities, *e.g.*, Lisbon, the two plots are quite different both in terms of trend and number of bins. In particular, Lisbon is a special case, because of its shape and the division in two main connected components (see Table4.20). The fact is that the central POI is in the second biggest connected component and it does not reach as

many nodes as the geographical centre. However, kept it this way since the shape of the city is peculiar, having the city centre in the second component and the estuary of river Tago between the two parts of the city. A last case study is represented by Athens, where the close nodes distribution is alike, whereas the far nodes one is fairly different when it come to its trend. The latter has a higher probability for the nodes that have distances close to the threshold for the geographic central node, with a mean of 24 km. On the other hand, the distribution of the distances from the POI one is smoother and more uniform, with a higher mean of 31 km.

In addition, we created complementary cumulative distribution plots for distances from central nodes. This decision was made to offer a city level representation that could be exploited a for better comparison among cities, given the nature of the normalized plot. For all the dataset, we plotted the complementary cumulative distribution of the distances obtained starting from the geographical centre, whereas for the capital cities we added the one about the POI centre. In both cases, we plotted the 1-CDF distribution both for the bfs and the Euclidean distances from the central source node. If we compare the two representation for the capital cities, we find that for most of the cities, 9 out of 10, the two plots are quite similar, showing an exponential trend. Just for Lisbon, the trend is still similar, but the density of the distributions is quite lower when starting from the central POI. This is due to the city peculiar shape and connected components definition, already explained in 4.4.1.

For each city, we exploit relative distances from the central point in order to determine a set of peripheral nodes. As one can grasp, with *peripheral* we identify nodes at the boundaries the connected component where the central node is located. We call a node peripheral whenever farther from the city center, relative to a set of nearby nodes. Given the set of nodes picked for comparison, a node could be "peripheral" on a mode local basis, or an absolute one, within the city. So peripheral nodes could help us track the city shape, as well as to identify transportation localities. Let us use $c$ for the index of the central node. In order to identify peripheral nodes, we adopted a two step algorithm:

- we first looked for nodes farther to the center, with respect to all their adjacent nodes. We labelled all reached nodes as peripheral ($\forall_{i \neq c} peripheral[i] = 1$). Then we iterated over all edges, by comparing the Euclidean distances of their two end nodes. Let us use $ed$ for Euclidean distance. Given an edge $(i, j)$, we compare $ed_{ic}$ to $ed_{jc}$: if $(ed_{ic} < ed_{jc})$, then $peripheral[i] = 0$, otherwise $peripheral[j] = 0$, so the closer to the center, between node $i$ and $j$ is labelled as non peripheral.

- as the previous step could identify nodes with all adjacent nodes closer to the

center (*e.g.*, terminal nodes of bus lines), this step generally labelled too many nodes as peripheral, so we applied a second filter, on nodes close enough (though not directly connected), to other nodes farther from the center. We empirically identified a reference distance $ed_{ref}$, given by 1/10 of the maximum Euclidean distance from the center: $ed_{ref} = (MAX_i ed_{ic})/10$. Then for every couple of nodes $i, j$, still labeled as peripheral ($peripheral[i] = 1$ and $peripheral[j] = 1$), such that $ed_{ij} < ed_{ref}$, we compare $ed_{ic}$ to $ed_{jc}$ and we reset the peripheral flag of the node closer to $c$.

In Figure 4.16a we show the representation of the peripheral nodes obtained for the city of Paris. The set is composed by 39 nodes out of the total 11950 of the graph and the 10644 of the biggest connected component (visible in Table4.20). These nodes seem to delimit quite nicely the shape of the city, which turns out to be circular.

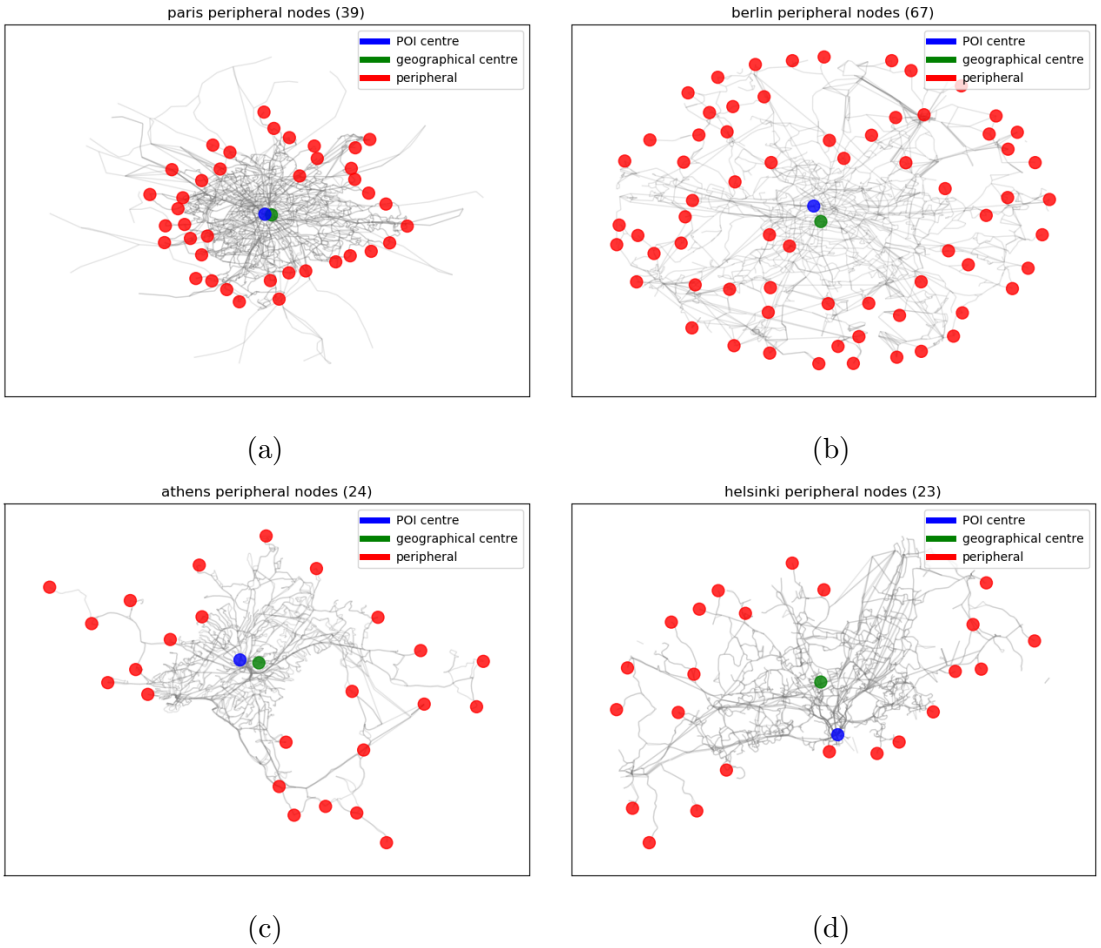



Figure 4.16: Peripheral nodes representation. Red nodes are the peripheral ones, green nodes are the geographical centre and blue nodes are the POI centre for capital cities. (a) Paris, (b) Berlin, (c) Athens and (d) Helsinki. In the parenthesis there is the number of peripheral nodes for the current city.

### 4.4.2 Comparison among cities

This section provides an inter-city comparison, based on probability distributions of distances, with the aim of understanding how well optimal (shortest) path lengths are close to the ideal best provided by Euclidean distances. To facilitate the comparison for some plots, we grouped cities in clusters, where the criterion we adopted for clustering has been explained in 4.4.

As a first comparison, we plot the fraction between the Euclidean distance and the bfs one. Though we could obviously take the inverse fraction, our choice has the advantage to provide values included between 0 and 1. The goal of this representation is to highlight how different the two measures are. The closer the fraction is to 1, the more the related bfs distance nears the Euclidean one. We represented how different the two are for all the cities, under the 1 to many (reachable nodes) distance computation scheme. In particular, we plotted the information gathered by a repeated random selection of the source node. In Figure 4.17 we present an overview of the results for all the clusters. In every plot, curves, and the related distance distributions, can be evaluated by observing

- how much they span on the right part of the graph, as a 1 value means bfs distance equals the Euclidean one. To this respect, "good" curve peaks can be found close to the 0.8 value;
- peak height and curve thickness around the peak, as tight and high peaks mean highly probable values close to the peak.

In general, it is reasonable to think that a pointed distribution well located on the right of the plot will correspond to a city's PTN more efficient than another where the curve is smoother/fatter and left-centred.

Analysing one cluster at a time, we can see how the first one (fig. 4.17a), which contains cities with an area between 0 and 100 $km^2$, have results in pairs. Antofagasta and Rennes share similar shape of their distributions, indicating that their PTNs might be quite efficient as they are. On the other hand, we have Luxembourg City and Kuopio, which share a lower curve with a smaller mean. These differences are likely related both to the shape of the city and, more probably, to the morphology of the territory. The second cluster (100-300 $km^2$), shown in Figure 4.17b, is fairly homogeneous. We can see how all the cities (Belfast, Palermo, Paris, Turku) share a similar trend, suggesting that in this case the dimension of the city might be a good grouping parameter. The majority of the nodes in all PTNs show path-based distances not far from the Euclidean ones, indicating fairly efficient planning of the

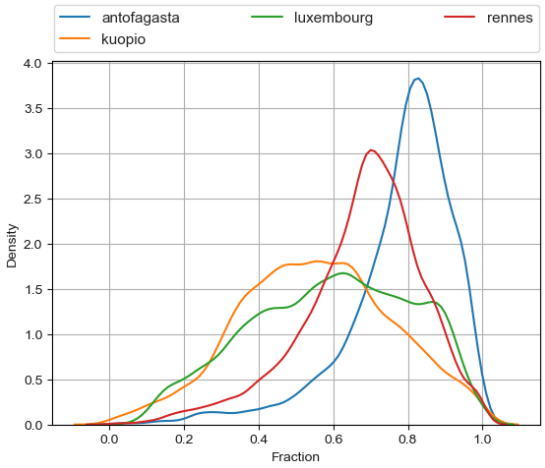
networks, especially in the cases of Palermo and Turku, with peak values of about 0.8. The third cluster, fig. 4.17c, is the biggest one, with its 7 cities with area from 300 to 500 $km^2$. Here, 6 out of 7 cities have similar trends, where the curves are not too sharp (maximum density of 2.0 to 2.5), with the peak value in the 0.5 to 0.7 range. Rome behaves as a partial outlier, with a sharper peak not far from 0.75. The situation for the fourth cluster (500-1000 $km^2$), fig. 4.17d, is quite similar to the previous one. In this case, 4 cities out of 5 have a similar trend, with slight variations in the peak location and curve tightness. However, here the different city, Adelaide, characterized by a smoother and more oscillating trend, somehow missing a real and dominating peak. This could be related to the geography of the city, a flat coastal land, crossed by multiple rivers and/or streams. As a consequence, the nodes could be decomposed in reachability clusters, witnessed by the multiple peaks and valleys in the curve. The last cluster, with cities bigger than 1000 $km^2$ and shown in Figure 4.17e, seem to have one city that stands out from the others, Lisbon, have a wide and oscillating trend, with peaks and valleys. An explanation similar to Adelaide could apply, based on the fact that, again, the territory is characterized by reachability areas/clusters, due to the river Tago, splitting the city in two.

Overall we can say that most of the city clusters have similar trends and they are pretty homogeneous, with few outliers. This might mean that city clustering by area, for this specific analysis, was a fair choice. The differences evidenced by outlier cases are generally related to the city's morphology, which in this work is not a parameter under analysis.

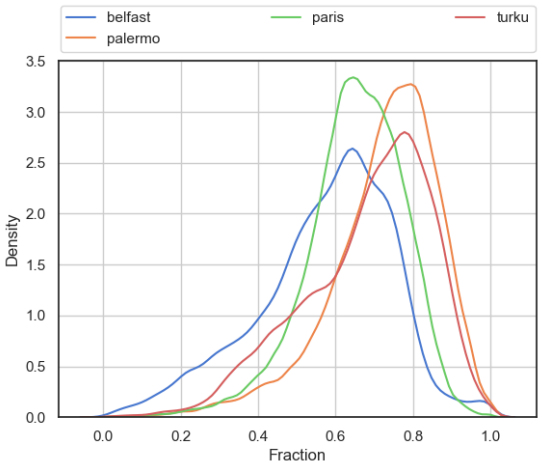
In addition to the plots explained above, we also compared the mean and standard deviation of all the cities, in order to obtain a higher level and global view. This allowed us to compare all cities in a single plot. We sorted cities by increasing mean values, and we also explicitly plotted city areas, so that we could better capture the possible correlation between distances and the size of the city. In particular, we plot two different representations, one for the bfs distributions and one for the Euclidean ones, all obtained by the randomized visit of the graphs. These are shown in Figure 4.18, where the cities are ordered by ascending mean value of bfs distance. Over the bar plot, we added the information about the city area to better compare the results. On a first look, it is clear that there are some discrepancies from the smooth and increasing trend of the mean in the first plot, 4.18a, and the trend in the second one, 4.18b. The most significant outlier here is Lisbon. In the first plot, it clearly shows a higher standard deviation, relative to the mean value, followed by Athens, meaning that their spread of values is wider than other cities'. When considering also Figure 4.18b (Euclidean distances), Lisbon is a clear outlier, with a ratio Euclidean vs. bfs distances much lower than the rest of the cities. As we already noticed, the peculiarity of this city and the results confirm that its PTN is

not, and probably cannot be, distance efficient.

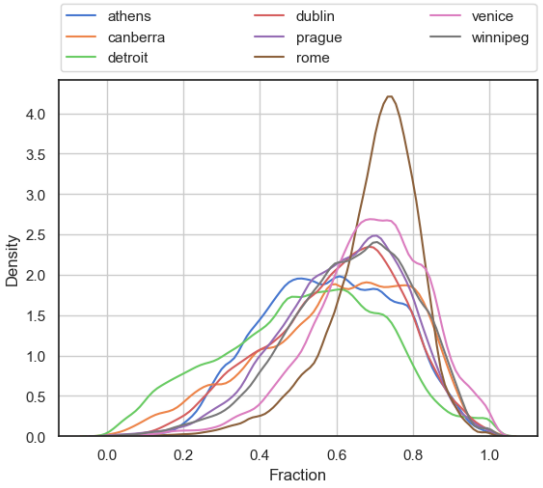
Another aspect to notice is the distance range. As we can see in the first plot, Melbourne mean reaches 50 km distance from the source node, whereas in the second plot the maximum Euclidean distance reached is 25 km, half of the first one. This is in line with the expected result, but it is still important to quantify the difference from bfs and Euclidean ranges. Concerning the area, we see that the more or less increasing trend of the distances is not shared by the area. A first evident outlier is Bordeaux. It has one of the biggest area of the dataset, but we find it in the left side of the distance plots. With its bfs mean between 20 and 30 and its Euclidean one a bit above 10, we might be surprised that the average distances are quite close to each other compared to such a big area. An explanation could be found in the choice between the urban and city areas: we chose for Bordeaux the urban area, hosting a little less than 1 million people. As the public transportation network covers an intermediate (not easily identified) area between the two, the urban area is clearly somehow big, whereas the city area would be small, resembling cities such as Paris and Luxembourg.



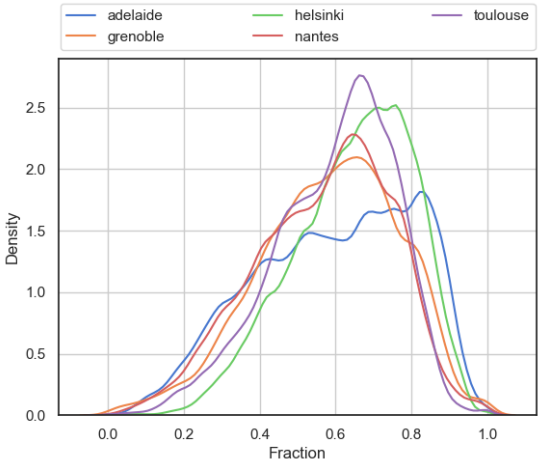

(a)



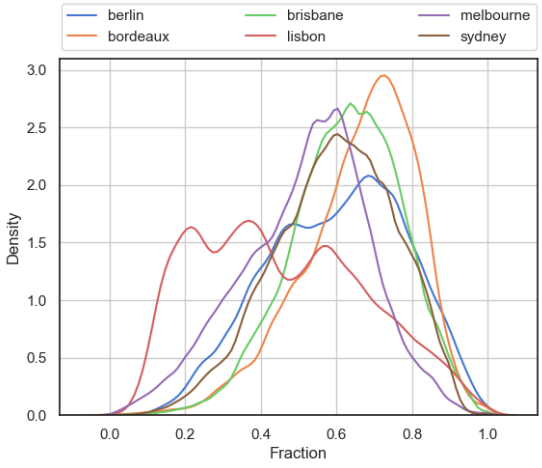

(b)




(c)

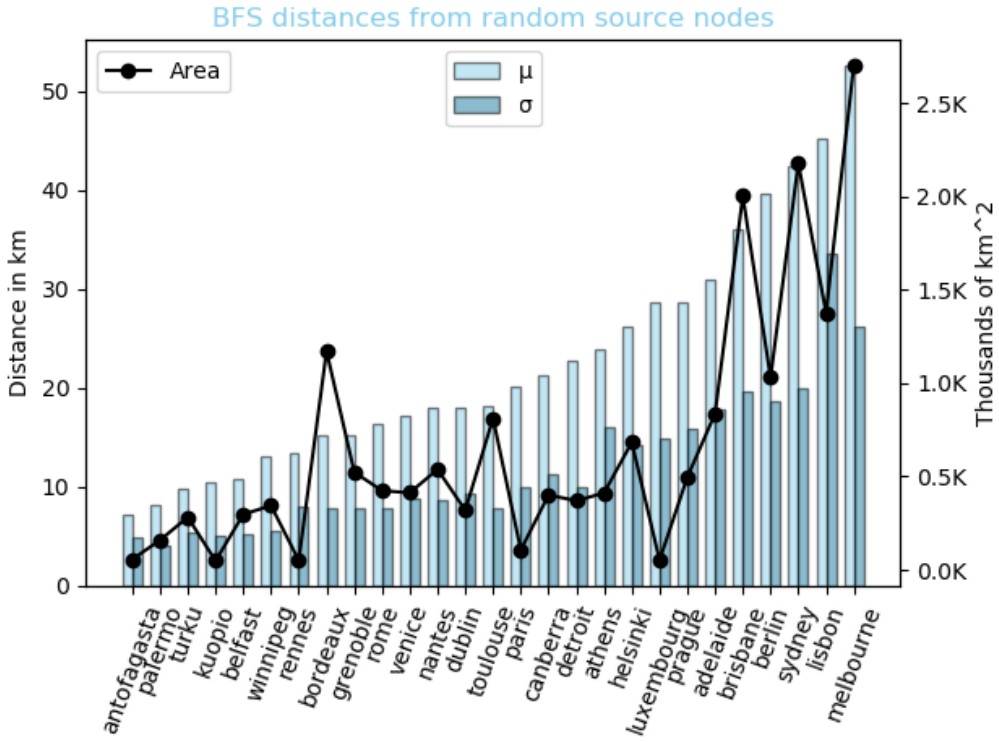



(d)

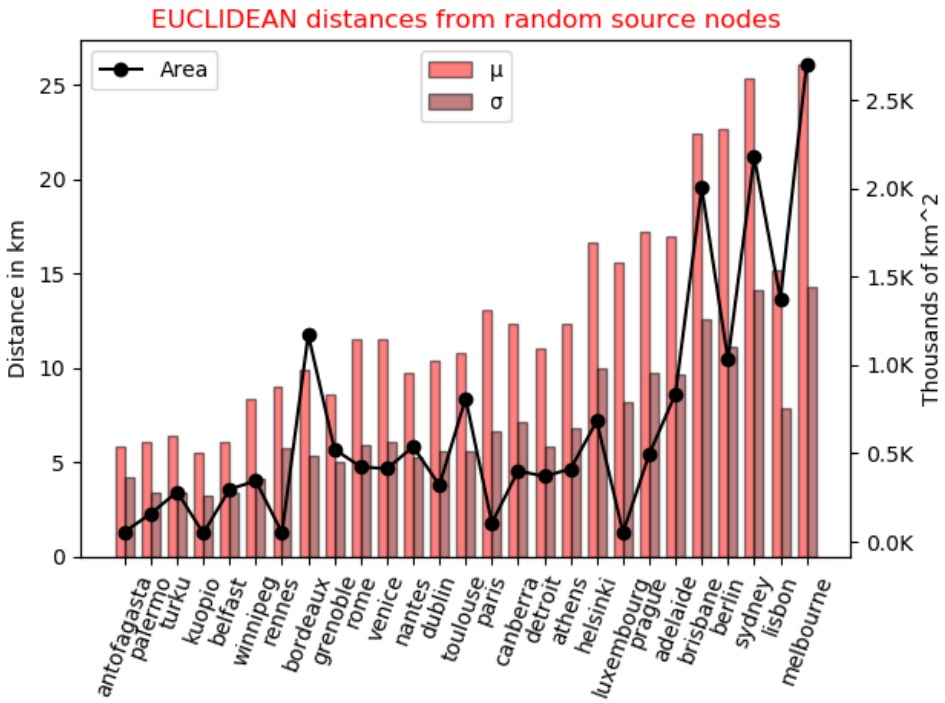



(e)

Figure 4.17: Fractions of Euclidean distances over the bfs ones. Each Figure correspond to a cluster defined as explained in 4.4. The figures are ordered by the clusters range of area, starting from the smallest range, 0-100, for the (a) Figure to the biggest one, 1000+, for the (e) one.




(a)




(b)

Figure 4.18: Bar plots of comparison between mean and standard deviation of distances distributions. (a) Bfs distances and (b) Euclidean distances. For both plots the order of the cities is by bfs mean ascending and the information on the right side is the area of the corresponding city.

### 4.4.3 Shortest paths vs breadth first

In this subsection we present a comparison between the shortest path lengths computed over a weighted network and those obtained through a breadth first search of the graph. This comparison was made to support and further explain our choice to consider the breadth first approach. To reach this goal, we created a weighted and undirected graph for each city, where the weights on the edges represent the Euclidean distances from the two nodes in km. Then, for each city we used the random source node approach (20 times) computing the paths from each source using Dijkstra algorithm and recorded their lengths. We computed the same breadth first visit approach explained in 4.4.1. In order to compare the two measures, we calculated the fractions between shortest path length and bfs distance and plotted together the results for all the cities together. The latter are visible in Figure 4.19, where we can see that the majority of the bfs distances correspond to the Dijkstra ones. All of the cities have a similar trend, even if these vary on maximum density reached.

Even thought the bfs distance might not always be the shortest one, since it is based on hops and not on the weight, we find that it fairly approximates the Dijkstra one. In light of our considerations and results, we consider our choice to be suited for the type of analysis explained throughout the section.

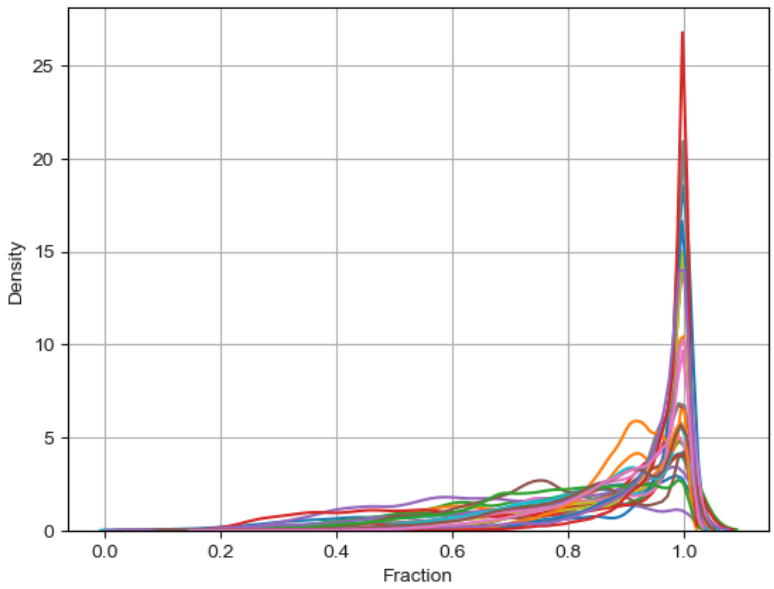



Figure 4.19: Fractions between shortest paths distances and bfs ones for all the cities with the random source nodes process.

### 4.4.4 Connected components analysis

As we approached the distance analysis, we came across the problem of connected components. In particular, in order to explore a graph through one of the available algorithms, it is important to do so on a connected component. We thus thought that an analysis on the networks connected components was necessary to better comprehend all the results shown before and the network topology. We remind here the reader that for this work we created unweighted and undirected graphs with all the types of transport together.

In Table4.20 we present an overview on the information about connected components in the dataset. For each city we show the number of total nodes in the graph, the number of nodes in the first four connected components, if present, the percentage of nodes includes in the first four components, and then we show the total number of connected components for each graph. Taking a first look at the table, we see that only 5 out of 27 cities have one connected components and all of them have just the bus transport network. For the other 22 cities the components may be mixed, having different types of transport in it, or divided. It is important to notice that a strong majority of the cities that have more than three connected components have a very high percentage of coverage by those components. On average, we see that the first four components cover more than 97% of the nodes in the graph. Only for Paris the percentage reaches 93, which is still a pretty good coverage. We did not find a common pattern between all these cities apart from the fact that, usually, the biggest ones are again the whole or a part of the bus transport network.

In addition, for all the cities we plotted each connected component with the colour of the edges in accordance with the type of transport. If one is interested, the plots may be found in the repository [27] in the directory *results/city_name/connected_components/plots/*, where the components are ordered by number of nodes, starting from the biggest one.

| City | N | comp 1 | comp 2 | comp 3 | comp 4 | % 4 comp | n comp |
|---|---|---|---|---|---|---|---|
| adelaide | 7548 | 7237 | 175 | 108 | 28 | 100 | 4 |
| antofagasta | 650 | 650 | | | | | 1 |
| athens | 6768 | 6639 | 61 | 49 | 19 | 100 | 4 |
| belfast | 1917 | 1917 | | | | | 1 |
| berlin | 4601 | 4593 | 4 | 2 | 2 | 100 | 4 |
| bordeaux | 3435 | 3212 | 89 | 72 | 57 | 99 | 5 |
| brisbane | 9645 | 9279 | 297 | 27 | 19 | 99 | 6 |
| canberra | 2764 | 2520 | 241 | 3 | | | 3 |
| detroit | 5683 | 5683 | | | | | 1 |
| dublin | 4571 | 4361 | 102 | 32 | 32 | 99 | 6 |
| grenoble | 1547 | 1292 | 164 | 51 | 31 | 99 | 5 |
| helsinki | 6986 | 6879 | 75 | 19 | 3 | 99 | 9 |
| kuopio | 549 | 528 | 21 | | | | 2 |
| lisbon | 7073 | 2730 | 2173 | 2042 | 50 | 98 | 11 |
| luxembourg | 1367 | 1365 | 2 | | | | 2 |
| melbourne | 19493 | 18757 | 352 | 86 | 86 | 98 | 9 |
| nantes | 2353 | 2228 | 76 | 47 | 2 | 100 | 4 |
| palermo | 2176 | 2176 | | | | | 1 |
| paris | 11950 | 10644 | 350 | 97 | 49 | 93 | 38 |
| prague | 5147 | 4980 | 34 | 24 | 24 | 98 | 18 |
| rennes | 1407 | 1304 | 74 | 29 | | | 3 |
| rome | 7869 | 7562 | 48 | 29 | 27 | 97 | 15 |
| sydney | 24063 | 22659 | 559 | 522 | 69 | 98 | 18 |
| toulouse | 3329 | 3230 | 52 | 21 | 19 | 99 | 5 |
| turku | 1850 | 1817 | 28 | 5 | | | 3 |
| venice | 1874 | 1644 | 111 | 110 | 3 | 99 | 7 |
| winnipeg | 5079 | 5079 | | | | | 1 |

Figure 4.20: Tableof connected components information. The N column indicates the number of nodes of the starting network. Comp i columns indicate the number of nodes in the i-th connected components, where the first one is the biggest. The fifth column shows the percentage of nodes covered by the first four connected components, if the city have four or more.The last column shows the number of connected components for the current city.

## 4.5 Frequency analysis

Lastly, we present an analysis over the behaviour of vehicles frequency during the hours of a typical day. We decided to conclude offering a different perspective on the dataset. The aim of this sort of work is to highlight differences and similarities between peak hours and other hours of the day, with the goal of establishing which PTNs are more likely to have a well distributed load and which ones are not. Furthermore, we try to correlate the results found with the city population.

For each city, the information used to perform this analysis is contained in the *network_temporal_day.csv* file obtained from [21]. In the file, for each line we have information about

- start stop id, which identifies the starting node
- end stop id, which identifies the ending node
- departure time in Unix time (number of seconds after 1.1.1970 00:00:00 UTC)
- arrival time also expressed in Unix time
- route type, which identifies the type of transport (*e.g.*, bus, tram, rail etc.)
- trip id, which identifies a trip (*i.e.*, a sequence of two or more stops that occurs at specific time)
- route id, which identifies the route (*i.e.*, a group of trips that are displayed to riders as a single service).

In the following section we illustrate how we used the above mentioned information to study the distribution of the number of vehicles throughout a typical day. In the first subsection 4.5 we present how we processed the information and analysed the frequencies of vehicles for one city at a time. In the second subsection 4.5.1 we show the comparison between cities. For this type of analysis we chose to divide the comparison between type of transport. More specifically, we decided to put together the information about mean and standard deviation for the frequency distributions of each city. Moreover, we added the information about the population to compare the number of vehicles information with the number of inhabitants. In the last subsection 4.5.2 we present a more specific comparison between the frequencies of the bus transport network for all the cities. Since every PTN has BTN, which is also the biggest one, we decided to analyse more in depth this type of transport.

### 4.5.1 City level

Here we present the process to calculate the frequency of vehicles for each city and type of transport. In the first part, 4.5.1, we illustrate how we calculated and represented the hourly frequency of vehicles. In the second part, 4.5.1, we explain the choices made to compare the peak hours to the others, how we computed the node frequency. In both cases we offer some visual examples of the results.

#### Hourly frequency

First of all, the temporal information presented in 4.5 are expressed in Unix time and they do not take into account time zones. Therefore, the first operation performed was to manually record for each city their hours difference from the central meridian. Then, we were able to sum these differences and convert the time for each city, in order to have the information about the hour of the day, form 00 to 23, when a certain vehicle would leave the start stop. In fact, for each type of transport, we wanted to collect the number of vehicles that would transit in each hour of the day. In practice, to calculate the number of vehicles we added the trip id to a specific set for the type of transport and hour defined in the same row of the csv file. At the end of the file processing, we counted the elements of each set, which would create the hourly number of vehicles for the specific type of transport. The usage of a set allowed us to remove duplicates. For example, the same vehicle would result in multiple lines with a different start stop and departure time, but with the same trip id. Counting unique trip ids was our way to calculate the number of vehicles with the minimum count error.

In Figure 4.21a we offer a visual representation of the results of this process for the bus frequencies for the city of Belfast. In the bar plot we highlighted both the morning (11:00 to 12:00) and afternoon (20:00 to 21:00) peak hours to stress the difference between the other time slots. As we can see from the picture, the majority of the vehicles transit during day hours, especially from 10:00 to 22:00, where the distribution of the number of vehicles is quite stable. In particular, it is clear that buses do not transit at all from 03:00 to 7:00, hour where they slightly start moving.

In Figure 4.21b we show the distribution plot for the same city of Belfast and type of transport (bus) as before. In this case, we do not take into account the hours with zero vehicles, that is night hours. For each number of vehicles, shown on the x axis, we present its probability. The distribution curve is pretty smooth and presents two small peaks, one among low values of frequency and one among high ones. In addition, the legend offers the information about mean and standard

deviation values. With a mean of 270.55 vehicles and a standard deviation of 159.37, we can say that the low values bring significantly down the mean value and increase that of the deviation. If one considers only the slots from 10 to 21, he would probably find a higher mean and definitely a lower standard deviation.

These two plots together offer a reasonable overview on the distribution of the frequencies for the wanted city, divided by type of transport. To be complete, the colour of each plot depend on the type of transport considered. In particular, here we offer a detailed list of the correlation between colour and type of transport, which is also the same used in the network representation:

- Blue for **bus** transport network
- Green for **tram** transport network
- Red for **rail** transport network
- Orange for **subway** transport network
- Aqua for **ferry** transport network
- Yellow for **cablecar** transport network (only one city).

#### Peak and mean hour stop frequency

In the previous section 4.5.1, we presented a first visual representation of the comparison between peak and other hours frequencies in Figure 4.21a. Here, we go a bit more in depth in this comparison explaining the choices made and process followed.

In order to highlights the differences and similarities, if present, for each city and type of transport, we chose to compare two network visualisations. Each one would represent a specific hour of the day, colouring the nodes based on the number of vehicles departed from that node in the considered hour. To do so, we defined two specific hours for each city:

- **Peak hour**: hour slot with the maximum number of vehicles in the whole day
- **Mean hour**: closest hour to an ideal one with average number of vehicles, that would guarantee a homogeneous distribution of the vehicles throughout the day.

In particular, when defining the mean hour, sometimes referred also as average hour, we considered only the hour slots with a number of vehicles higher than zero. Then, we computed the average between the values and we took the hour with the closest frequency to the average one.

Afterwards, we collected the information about number of vehicles departing from each stop in both time slots. The process was almost identical to the one used for the frequencies per hour presented before in 4.5.1, but we added a counter for each node of the network. We chose to register the number of vehicles for the starting stop. We plotted the network with the nodes colour depending on the number of vehicles starting from/passing through that node in the specified hour. We chose not to represent the nodes with zero number of vehicles. Even with this threshold, the representation is often crowded. In the solution presented in this work, we decided to maintain this approach even though some cities might have a higher concentration of nodes, most in the bus representation. For further analysis, we offer a solution for adding a relative threshold which gives a less crowded representation.

In Figure 4.21c, we present the network representation of Belfast's BTN during the peak hour, whereas in Figure 4.21d during the mean/average hour. The colours of the nodes reflect the number of vehicles departing/passing through the stop. The range of the colorbar is based on the peak hour one. We chose the same one for both plots to highlight the differences between nodes frequencies. We can see how the first picture exploits all the available shades of colours, meaning that the nodes represented have number of vehicles from 1 to 50. On the other hand, we see how the second picture uses only 3/4 of the available range. Overall, the number of nodes represented is pretty similar, and what differs is the frequencies of the nodes. From our perspective, this is a good news, because it means that the territory is served either way, just with different frequencies.

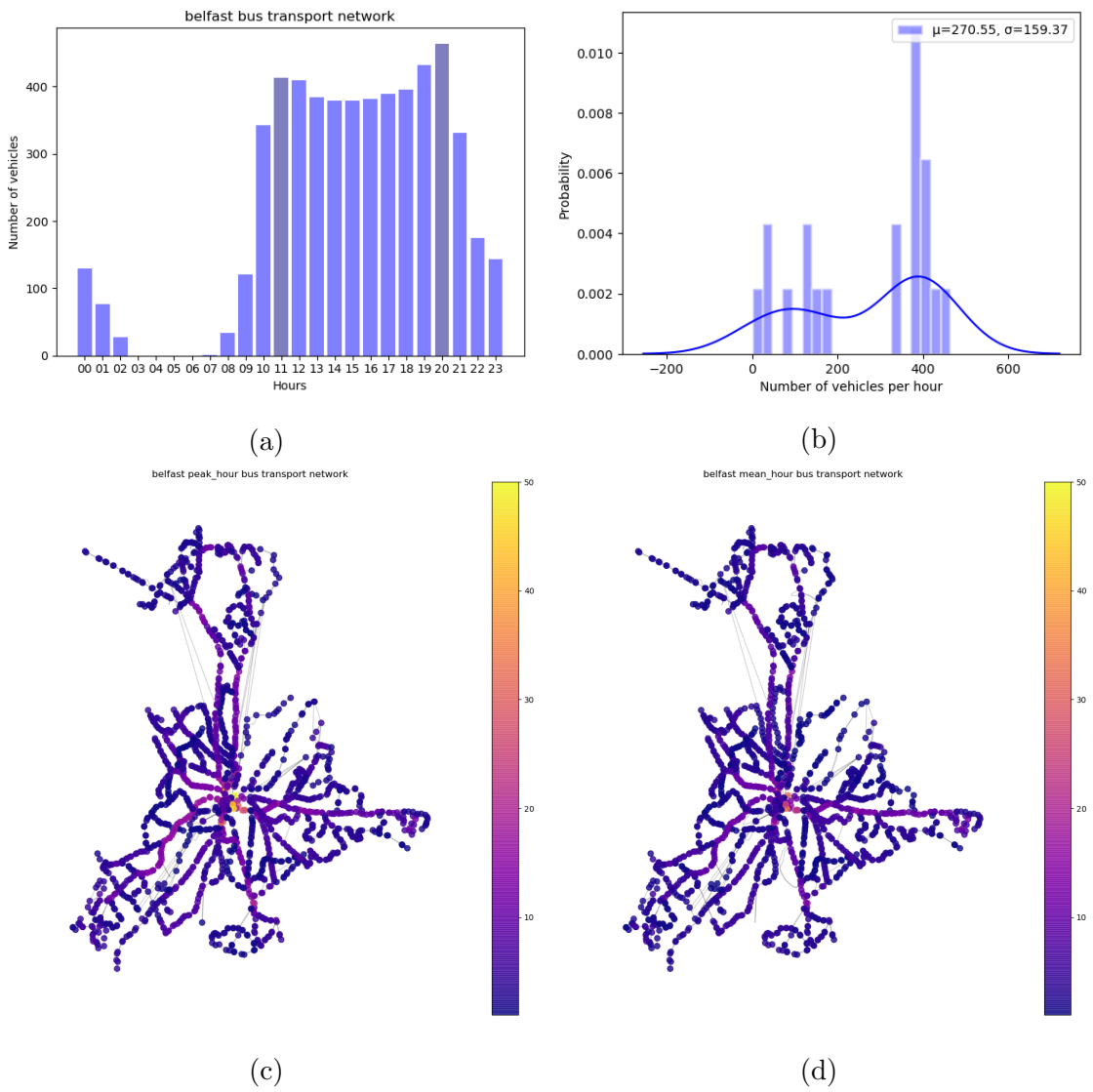



Figure 4.21: (a) Vehicle frequency for all hours of the day, (b) frequency distribution, (c) peak hour network visualization, (d) average/mean hour network visualization for bus transport network of Belfast

### 4.5.2 Comparison between cities

In this section, we present a first comparison between all the cities for the frequency analysis. To achieve this goal, we gathered together the data about mean and standard deviation for all the cities separated by type of transport and plotted them. Moreover, we decided to add the information about the population, to give a better picture of the results and allow to correlate them with that city feature.

In Figure 4.22 we show 5 out of 6 type of transport plots. We do not show here the cablecar plot, since it concerns only the city of Prague. However, we remind the reader that all the results are available at [27], where one can browse them

extensively.

In the first plot, 4.22a, we show the comparison between the frequencies of all the bus transport networks. The cities are ordered by ascending mean value. Visually, it is clear how the trend in the population does not follow those of the frequency information. The first outlier that we talk about is Athens, which is the 8th city (starting from right), but has a population of more than 3 millions people. Taking a look to its frequency plot in the repository, we can see how this city has all hours with frequency higher than zero, even night hours. This factor is probably the reason why the mean is lower than expected. If we did not consider night hours at all, Athens would probably have a more balanced frequency mean with respect to its population. The night hour factor does probably influence two other outliers: Melbourne and Sydney. Again, they have very low numbers of vehicles during nigh hours, but still above zero, which is the threshold considered in out study. Both cities share a high value of standard deviation. For the case of Melbourne this is most probably due to the night hours factor, whereas for Sydney the distribution during day hours is still quite unstable. The last case that we discuss is the one of Paris. It is the city with highest mean value, but population below many other cities. This is probably due to the fact that it is a very busy and small city, compared to others in the dataset. We remind the reader of the fact that Paris was already a specific case, having a concentration of nodes per area significantly higher than all other cities (fig. 4.1b). Therefore, this result might be explained also by that information. The first half of the cities shares a common trend between frequencies and population, whereas the second one is significantly less stable. Lastly, we can say that all cities have a fairly high standard deviation, which, in some cases, might be attributed to the contribution of nigh hours, late evening hours and early morning ones.

In Figure 4.22b we present the comparison between tram transport networks. Here, it is clear that there is no similar trend between the frequencies and the population of the cities. It is interesting to notice how a city big and richly populates as Sydney has a very small tram transport network. This is, however, compensated by one of the biggest, in terms of number of vehicles, bus transport network. Adelaide too has a very low mean value of frequency. This probably indicates that these cities have a poor tram service. On the other hand, we find the third Australian city in this plot, Melbourne, that have the second tram transport network, in terms of frequency of vehicles. Lastly, Prague is a sort of outlier in the other sense. It has quite a small population compared to other cities, but boats a very frequent tram transport network, with a around 700 vehicles during morning and afternoon peak hours.

In Figure 4.22c we show the results for the rail transport network. Given the

spatial filtering explained in 3.3, the information about this type of transport is partial and limited to the city area. However, we can still highlight some features from the plot. For example, we notice how all the cities with a rail transport networks are either capitals (8) or Australian cities (4). In fact, the latter are all present in this plot, even if the information about Canberra are not very significant. In Figure 4.22d we show the subway transport network frequency results. Here, as before, most of the cities are capitals (7 out of 9). However, none of the Australian cities is present, meaning that they prefer other types of transport and do not have this one. The most significant STN is the one of Paris, with a mean above 800 vehicles per hour. All the other cities have a mean that do not go beyond half of that value. Lastly, in Figure 4.22e we show the ferry transport network. Here, the famous Venice is, as expected, the most active. Two of the five Australian cities also share high values of vehicles per hour. All the other cities, do not exceed 30 ferries per hour.

Overall we notice how there are some cities, like Berlin and Paris, which are almost always present and usually have fairly high values of frequency. This is not a surprise, given that they are two of the most important cities in Europe. On the other hand we have the Australian cities, that tend to have all types of transport except the subway, but they have discordant behaviours depending on the type of transport. The addition of the population information helps to give a bigger picture on the specific city, but, overall, there do not seem to be any strong correlation between the population and the mean frequency of vehicles throughout the dataset.

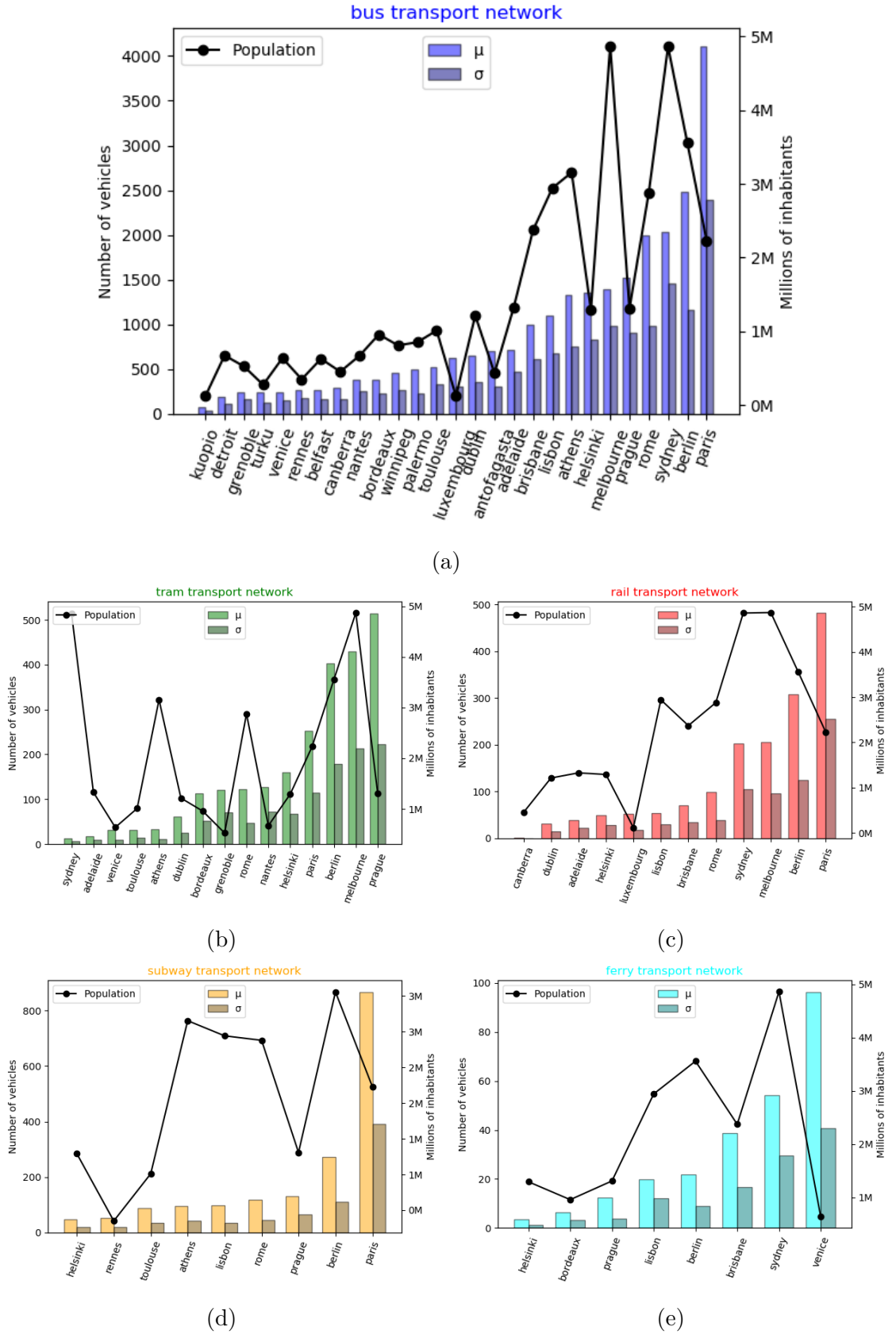



Figure 4.22: Mean and standard deviation of vehicle frequency distribution for all the cities divided by type of transport. In all the plots the cities are ordered based on increasing mean value for the specific type of transport.

### 4.5.3 Bus transport network

In conclusion to the frequency analysis, we offer a deeper study on the bus transport network (BTN) for all the cities. It is an addition to the previous subsection, where we compared general statistics of all the cities divided by type of transport. Here, we chose to expand more the discussion about the only means of transport that every city has.

Figure 4.22a, already commented in 4.5.2, offers a first rough comparison between all the cities BTNs, presenting their mean and standard deviation information, together with their population data. In addition, in Table4.23 we present an overview of some information about the BTN of each city:

- **Number of lines**: counter of all the bus routes available for a city. A route is a group of trips that are displayed to riders as a single service.
- **Peak hour**: two digits number indicating the hour where there is the maximum number of vehicles.
- **Number of vehicles in the peak hour**: how many vehicles transit during the peak hour
- **Mean hour**: closest hour to an average value of frequency of vehicles. It represents what a typical hour would be if the distribution was uniform.
- **Number of vehicles in the mean hour**: how many vehicles transit during the mean hour.
- **Total number of vehicles**: number of vehicles throughout a typical day.

Analysing the results we can say that the city with the maximum number of lines is Lisbon, with its 820 different bus routes. Furthermore, 75% of the dataset has a number of lines below half of the maximum value, this implies that on average the number of routes is between 16 (Antofagasta) and 348 (Melbourne). The number of lines is positively correlated with all the frequencies of vehicles, with a 0.71 for the peak hour number, 0.67 for the mean hour and the total number of vehicles. It is slightly correlated with the population information with a Pearson coefficient of 0.51.

Turning our attention towards the peak and mean hour data, it is interesting to underline how 7 out 27 cities have a morning peak hour (either 8:00 to 9:00 or 9:00 to 10:00). Only one city, Palermo, has a mid day peak hour from 12:00 to 13:00, probably due to different working hours in the south of Italy. The majority of the dataset has an afternoon peak hour: 2 cities at 16:00, 3 cities at 17:00, 12 cities at

18:00 and 1 at 19:00. Also in this case, only one city, Belfast, has an evening rush hour from 20:00 to 21:00. Overall, we can say that the most common peak houris from 18:00 to 19:00, which is probably expected given the most widespread working hours. The situation for the average hour is more balanced between morning and evening. We find 13 cities from 6:00 to 12:00, more specifically 1 at 6:00, 7 at 7:00, 1 at 8:00 and 4 at 11:00. Only Winnipeg has a mean hour in the first afternoon from 14:00 to 15:00 and the rest of the cities have an evening average hour (4 at 20:00, 6 at 21:00 and 3 at 22:00).

The three types of frequencies (peak hour, mean hour and total) all have a fairly spread range. The first one goes from a minimum of 124 vehicles for the city of Kuopio to a maximum of 7110 in Paris. The mean hour has a less extensive range but the two cities are the same, with 74 vehicles for Kuopio and 4382 for Paris. In both cases, the 3/4 of the dataset have values of frequencies lower or equal to 32% of the maximum one. For the total number of vehicles during the day the situation is quite similar, with a range of 1573-94377 and the majority of the cities with values less than 35% of the maximum one.

| | # lines | Peak hour | # vehicles PH | Mean hour | # vehicles MH | Total # vehicles |
|---|---|---|---|---|---|---|
| City | | | | | | |
| adelaide | 506 | 09 | 1380 | 07 | 627 | 15776 |
| antofagasta | 16 | 19 | 1008 | 11 | 736 | 13348 |
| athens | 233 | 09 | 2072 | 22 | 1312 | 31752 |
| belfast | 101 | 20 | 464 | 21 | 331 | 5411 |
| berlin | 456 | 09 | 3907 | 22 | 2291 | 59502 |
| bordeaux | 76 | 18 | 653 | 21 | 351 | 8478 |
| brisbane | 330 | 18 | 1906 | 07 | 986 | 21043 |
| canberra | 116 | 18 | 574 | 11 | 306 | 5622 |
| detroit | 39 | 17 | 360 | 20 | 193 | 4667 |
| dublin | 125 | 18 | 1088 | 21 | 606 | 14319 |
| grenoble | 46 | 18 | 482 | 07 | 227 | 5264 |
| helsinki | 463 | 17 | 2519 | 20 | 1362 | 32575 |
| kuopio | 32 | 16 | 124 | 07 | 74 | 1573 |
| lisbon | 820 | 09 | 1936 | 07 | 1148 | 26236 |
| luxembourg | 168 | 18 | 949 | 07 | 680 | 13145 |
| melbourne | 348 | 18 | 2498 | 21 | 1294 | 33574 |
| nantes | 92 | 18 | 783 | 11 | 420 | 8441 |
| palermo | 93 | 12 | 685 | 08 | 473 | 11477 |
| paris | 793 | 18 | 7110 | 21 | 4382 | 94377 |
| prague | 316 | 08 | 3071 | 06 | 1480 | 36406 |
| rennes | 48 | 18 | 487 | 21 | 266 | 5945 |
| rome | 315 | 08 | 2871 | 22 | 2079 | 47990 |
| sydney | 674 | 18 | 4263 | 20 | 2302 | 48849 |
| toulouse | 99 | 18 | 967 | 07 | 491 | 11079 |
| turku | 102 | 16 | 391 | 20 | 226 | 5294 |
| venice | 698 | 08 | 587 | 11 | 259 | 5800 |
| winnipeg | 89 | 17 | 918 | 14 | 446 | 10177 |

Figure 4.23: Tablefrequencies

# Chapter 5

# Conclusions

The purpose of this thesis was to characterise and study the network properties of the PTNs in our dataset and to find possible correlations between properties and city features, such as area and population. To achieve this goal we analysed 27 PTNs from different points of view: *network measures*, addressed in the first three sections of chapter 4 (basic, additional and centrality measures), *distance analysis* and *frequency analysis*. The various analysis tasks done can be roughly split in three main categories, or more appropriately in different levels of detail:

- city level, where we studied graphs in order to identify important nodes, probability distributions, distances and path lengths within single cities
- comparison among cities, on selected (often averaged) measures and statistics
- comparison among measures taken as overall/summary data in order to capture potential affinities/correlations and to possibly obtain hints on general features of PTNs

Here we review the most significant ones for each area of analysis and then we draw few general conclusions.

The first part of the analysis was oriented to describe and evaluate the dataset in terms of network properties and models. We started from some basic measurements: number of nodes, number of egdes, density, diameter and clustering coefficient. The analysis of results basically confirmed some rather obvious facts: as expected, the number of nodes $N$ and the number of edges $E$ have values that span over well identified ranges. They are clearly related to the ranges of area and population, and to the to spatial constraints. $N$ and $E$ are roughly proportional, and density is, overall, quite low, indicating that the networks are fairly sparse, which is typical of

PTNs as instances of spatial graphs; network diameters are generally proportional to nodes and edges, which means that the bigger the network,the longer the diameter, with exceptions, that are probably due to the shape and geographic characteristics of cities; the low values of average clustering coefficients have to do with wiring (transport routes/connections) cost and to the fact that most nodes have degree 2 (they are just transit nodes), which corresponds to a 0 clustering coefficient.

We then deepened our analysis by considering more targeted and specific measures, like assortativity, average path length, average degree and degree distribution. This part of the analysis revealed some interesting results. In particular, our PTNs do not behave as small-worlds and they are not scale-free. The first conclusion was drawn through the study of the scaling between $\langle l \rangle$ and $ln(N)$ joint with the analysis of the clustering coefficient. The scale-freeness usually requires very large networks and a large range of $k$-values, both of which are criteria not met for our PTNs. However, we still found that the degree distributions are broad and, for the range of k considered in our work, are practically indistinguishable from power-law distributions. Furthermore, it is important to highlight that a wide majority of the nodes have low degree, $k = 2$ for transit nodes and $k = 1$ for start/end stops, but there are still few hubs, usually represented by crossing stops or stations.

The last part of the network measures deals with nodes centrality measures. The study analyses four measures that are able to characterize a city network stops: betweenness, closeness, degree and eigenvector. In line with previous works, we focused more on the first two centrality measures, that are deemed more important for PTNs. In particular, betweenness centrality helped us understand if and where the hubs were located in the network. As expected, the majority of the nodes share low values of betweenness, whereas only few ones can be considered hubs and they are usually concentrated in the central area, even though there might be some exceptions where each city district has its owns. On the other hand, closeness centrality highlights PTN areas with a high concentration of nodes close to each other. Overall, in many PTNs, nodes with low betweenness seem to have fairly high closeness values, meaning that they might be central, but not a crossing stop. In our study, degree and eigenvector centrality output very close, and sometimes equal, results. In particular, there is not any specific relation between eigenvector centrality and betweenness one. Hubs do not necessarily connect to nodes with many neighbours, even though it might happen.

The second big part of the analysis is on distance analysis, where we aim at studying the mutual reachability of the nodes by comparing euclidean distances and shortest path lengths. In particular, we exploit the breadth first search (bfs) algorithm, which provides a reasonable result in term of minimum hop number.

Even though, strictly speaking it does not always give a shortest path, it is often equal or very close to the minimum path obtained by the Dijkstra algorithm. The comparison between travel distances, computed with bfs, to euclidean distances is used to evaluate the quality of the transport network. This type of analysis takes into accounts geographic and territorial aspects, shapes of transport lines and we look for relations between the results and the city area. The cluster analysis highlights some interactions between the fraction of distances and the area of the cities. Even though there are few outliers, it seems that area clustering is a fair grouping parameter for the distance analysis. Furthermore, the identification of peripheral nodes offers results that might, in some cases, underline the shape of the city. Another interesting finding concerns connected components. In this part of the analysis we study their number and distribution throughout the PTNs. More than 80% of the PTNs have more than one connected component and it does not always correspond to the type of transport division. On the other hand, the main component usually matches the first one, with the only exception of Lisbon. In general, however, given the heterogeneity of the dataset, we cannot draw strong conclusion for all the cities together.

The last section deals with the frequency analysis. The aim here was to study the distribution of vehicles frequency throughout a typical day, highlighting the differences and similarities between the peak hour and an average one. We performed the analysis divided for type of transport, with a more detailed study on the bus transport data. The results are fairly different both for distinct types of transport in the same city and for comparison among cities. In general, all the cities present two main peak hours, one during the morning and the other during the evening. Our choice for the comparison was to take under consideration the maximum peak hour and the closest hour to an ideal one, which would guarantee a uniform distribution of frequencies during the day. The main difference between the two hours, for most cities, is in terms of number of vehicles running. The two situations seem to serve almost the same number of stops, which means that the public transport network considered reaches quite the same city areas in both hours. However, depending on the city the gap between the number of stops served during peak hours and during the average one might be significant. In conclusion, this analysis does not highlight a specific trend between vehicles frequency and city population. There are many factors to take into consideration that, due to time effort and resources, we were not able to consider.

**Final conclusions and future works**

Even though our work is just a starting one, we drew some interesting conclusions that helped us capturing both expected and not so clearly visible trends and characteristics. The frequency analysis results may be a good starting point to redistribute the vehicle load throughout the day or to analyse and extend the service to more peripheral areas. Of course, these suggestions have to be coupled with data and analysis about people movements and probably political decisions about working hours. On the other hand, the results obtained from the distance analysis represents a first step in the efficiency analysis of the PTNs. One could decide to plan a different distribution of the stops and routes to enhance the fraction betwenne travel and euclidean distance. Again, one would need to deepen and customize the analysis for the city, considering for example morphology, points of interests and travel times.

In addition, we opened a path for future works, in three directions:

- improve the analysis of data that we already considered, by revising and refining the measures, possibly add more cities, improve the study of correlation among different measures;
- expand the analysis of time related and dynamic data, that could better capture the problems and peculiarities of transportation scenarios that vary on a daily, weekly and seasonal/yearly basis;
- improve the interface between data analysis and decision making applications such as public transportation planning and control at various levels.

# Bibliography


[1] M. Newman. *Networks: An Introduction*. OUP Oxford, 2010. ISBN: 9780199206650. URL: https://books.google.it/books?id=q7HVtpYVfC0C.

[2] URL: https://en.wikipedia.org/wiki/.

[3] Anna Broido and Aaron Clauset. "Scale-free networks are rare". In: *Nature Communications* 10 (Jan. 2018). DOI: 10.1038/s41467-019-08746-5.

[4] Marc Barthélemy. "Spatial networks". In: *Physics Reports* 499.1 (2011), pp. 1–101. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2010.11.002. URL: http://www.sciencedirect.com/science/article/pii/S037015731000308X.

[5] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Rev. Mod. Phys.* 74 (1 Jan. 2002), pp. 47–97. DOI: 10.1103/RevModPhys.74.47. URL: https://link.aps.org/doi/10.1103/RevModPhys.74.47.

[6] L. A N Amaral et al. "Classes of small-world networks". English (US). In: *Proceedings of the National Academy of Sciences of the United States of America* 97.21 (Oct. 2000), pp. 11149–11152. ISSN: 0027-8424. DOI: 10.1073/pnas.200327197.

[7] Subinay Dasgupta et al. "Small-world properties of the Indian Railway network". In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 67 (Apr. 2003), p. 036106. DOI: 10.1103/PhysRevE.67.036106.

[8] Yihong Hu and Daoli Zhu. "Empirical analysis of the worldwide maritime transportation network". In: *Physica A: Statistical Mechanics and its Applications* 388.10 (2009), pp. 2061–2071. URL: https://EconPapers.repec.org/RePEc:eee:phsmap:v:388:y:2009:i:10:p:2061-2071.

[9] Christian O. (Otto) von .Ferber et al. "Network Harness: Metropolis Public Transport". In: *Physica A: Statistical Mechanics and its Applications* 380 (Aug. 2006), pp. 585–591. DOI: 10.1016/j.physa.2007.02.101.

[10] Yong-Zhou Chen, Nan Li, and Da-Ren He. "A study on some urban bus transport networks". In: *Physica A: Statistical Mechanics and its Applications* 376 (Mar. 2007), pp. 747–754. DOI: 10.1016/j.physa.2006.10.071.

[11] Julian Sienkiewicz and Janusz Hołyst. “Statistical analysis of 22 public transport networks in Poland”. In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 72 (Oct. 2005). DOI: 10.1103/PhysRevE.72.046127.

[12] András London et al. “Complex network analysis of public transportation networks: A comprehensive study”. In: (June 2015). DOI: 10.1109/MTITS.2015.7223282.

[13] Xinping Xu et al. “Scaling and correlations in three bus-transport networks of China”. In: *Physica A: Statistical Mechanics and its Applications* 374 (Aug. 2007). DOI: 10.1016/j.physa.2006.06.021.

[14] Hui Zhang et al. “The Analysis of the Properties of Bus Network Topology in Beijing Basing on Complex Networks”. In: (2013).

[15] Tanuja Shanmukhappa, Ivan Wang-Hei Ho, and Chi Kong Tse. “Spatial analysis of bus transport networks using network theory”. In: *Physica A: Statistical Mechanics and its Applications* (2018). ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2018.02.111. URL: http://www.sciencedirect.com/science/article/pii/S0378437118302024.

[16] Tanuja Shanmukhappa et al. “Multi-layer Public Transport Network Analysis”. In: (May 2018), pp. 1–5. DOI: 10.1109/ISCAS.2018.8351818.

[17] Tanuja Shanmukhappa et al. “Recent Development in Public Transport Network Analysis From the Complex Network Perspective”. In: *IEEE Circuits and Systems Magazine* 19 (Nov. 2019), pp. 4880–4897. DOI: 10.1109/MCAS.2019.2945211.

[18] Harold Soh et al. “Weighted complex network analysis of travel routes on the Singapore public transportation system”. In: *Physica A: Statistical Mechanics and its Applications* (2010). ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2010.08.015. URL: http://www.sciencedirect.com/science/article/pii/S0378437110006977.

[19] Jianhua Zhang et al. “Networked analysis of the Shanghai subway network, in China”. In: *Physica A-statistical Mechanics and Its Applications - PHYSICA A* 390 (Nov. 2011), pp. 4562–4570. DOI: 10.1016/j.physa.2011.06.022.

[20] Atanu Chatterjee, Manju Manohar, and Gitakrishnan Ramadurai. “Statistical Analysis of Bus Networks in India”. In: *PLOS ONE* 11 (Dec. 2016). DOI: 10.1371/journal.pone.0168478. URL: https://doi.org/10.1371/journal.pone.0168478.

[21] URL: http://transportnetworks.cs.aalto.fi/about.

[22] Rainer Kujala et al. “A collection of public transport network data sets for 25 cities”. English. In: *Scientific Data* 5 (2018), pp. 1–14. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.89.

[23] URL: https://www.citypopulation.de/.

[24] URL: https://www.macrotrends.net/cities/20997/toulouse/population.

[25] URL: https://worldpopulationreview.com/world-cities/.

[26] URL: https://github.com/CxAalto/gtfspy.

[27] URL: https://github.com/scabodi/thesis.

[28] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Rev. Mod. Phys.* (2002). DOI: 10.1103/RevModPhys.74.47. URL: https://link.aps.org/doi/10.1103/RevModPhys.74.47.

[29] S.N.Dorogovtsev and J.F.F.Mendes. "Evolution of networks". In: *Advances in Physics* 51.4 (2002), pp. 1079–1187. DOI: 10.1080/00018730110112519. URL: https://doi.org/10.1080/00018730110112519.

[30] Alexei Vázquez et al. "Topology and correlations in structured scale-free networks". In: *Phys. Rev. E* 67 (2003). DOI: 10.1103/PhysRevE.67.046111. URL: https://link.aps.org/doi/10.1103/PhysRevE.67.046111.

[31] Konstantin Klemm and Víctor M. Eguíluz. "Highly clustered scale-free networks". In: *Phys. Rev. E* 65 (2002). DOI: 10.1103/PhysRevE.65.036123. URL: https://link.aps.org/doi/10.1103/PhysRevE.65.036123.

[32] Stephen P. Borgatti. "Centrality and network flow". In: *Social Networks* 27.1 (2005), pp. 55–71. ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2004.11.008. URL: http://www.sciencedirect.com/science/article/pii/S0378873304000693.

[33] Stephen P. Borgatti and Martin G. Everett. "A Graph-theoretic perspective on centrality". In: *Social Networks* 28.4 (2006), pp. 466–484. ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2005.11.005. URL: http://www.sciencedirect.com/science/article/pii/S0378873305000833.

[34] URL: https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.centrality.closeness_centrality.html.

[35] URL: https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.degree_centrality.html#networkx.algorithms.centrality.degree_centrality.