Urban Structure and Mobility as Spatio-temporal complex Networks

Doctoral Thesis
by

Gevorg Yeghikyan

Doctoral Program in Data Science

Supervisor
Mirco Nanni, ISTI-CNR, Pisa

Supervisor
Angelo Facchini, IMT, Lucca

Supervisor
Marco Conti, ISTI-CNR, Pisa

Supervisor
Andrea Passarella, ISTI-CNR, Pisa

Supervisor
Bruno Lepri, FBK, Trento

Urban Structure and Mobility as Spatio-temporal complex Networks

by

Gevorg Yeghikyan

Tuesday 16th June, 2020 01:42

Submitted to the Scuola Normale Superiore di Pisa
in June 2020, in partial fulfillment of the requirements for the
Doctoral Program in Data Science

Abstract

Contemporary urban life and functioning have become increasingly dependent on mobility. Having become an inherent constituent of urban dynamics, the role of urban moblity in influencing urban processes and morphology has increased dramatically. However, the relationship between urban mobility and spatial socio-economic structure has still not been thoroughly understood. This work will attempt to take a complex network theoretical approach to studying this intricate relationship through

- the spatio-temporal evolution of ad-hoc developed network centralities based on the Google PageRank,
- multilayer network regression with statistical random graphs respecting network structures for *explaining* urban mobility flows from urban socio-economic attributes,
- and Graph Neural Networks for *predicting* mobility flows to or from a specific location in the city.

Making both practical and theoretical contributions to urban science by offering methods for describing, monitoring, explaining, and predicting urban dynamics, this work will thus be aimed at providing a network theoretical framework for developing tools to facilitate better decision-making in urban planning and policy making.

Keywords: urban mobility, machine learning, complex networks, socio-economic attributes, spatio-temporal activity, neural networks

Publications

As main author

1. Gevorg Yeghikyan, Felix L Opolka, Mirco Nanni, Bruno Lepri, and Pietro Lio. Learning mobility flows from urban features with spatial interaction models and neural networks. *arXiv preprint arXiv:2004.11924*, 2020 (to appear in the Proceedings of IEEE International Conference on Smart Computing (SMARTCOMP 2020))

2. Gevorg Yeghikyan, Leandro Tortosa, Jose F Vicent, and Mirco Nanni. Ranking places in attributed temporal urban mobility networks. *PloS one*, 2020 (under review)

As co-author

1. Manuel Curado, Leandro Tortosa, Jose F Vicent, and Gevorg Yeghikyan. Analysis and comparison of centrality measures applied to urban networks with data. *Journal of Computational Science*, page 101127, 2020

2. Manuel Curado, Leandro Tortosa, Jose F Vicent, and Gevorg Yeghikyan. Understanding mobility in Rome by means of a multiplex network with data. *Journal of Computational Science*, 2020 (under review)

3. Leandro Tortosa, Jose F Vicent, and Gevorg Yeghikyan. A centrality measure based on the Adapted PageRank Algorithm for multiplex networks with data. *Applied Mathematics and Computation*, 2020 (under review)

Conferences

1. IEEE International Conference on Smart Computing (SMARTCOMP 2020), Bologna, Italy

2. Italian Regional Conference on Complex Systems, (CCS Italy 2019), Trento, Italy

Dedicated to the memory of my grandfather, Razmik Aslanyan.

Acknowledgments

This work would not have been possible without the support, advice, and encouragement on the part of my supervisors Mirco Nanni, Angelo Facchini, Bruno Lepri, Andrea Passarella, and Marco Conti, to whom I owe very much.

I would also like to thank Dino Pedreschi - the coordinator of the joint Data Science PhD program I had the honor of being part of - for making the research infrastructure and collaborations within the PhD ecosystem possible.

Special thanks also go to other members of the Data Science PhD board as well as individuals connected to the program who have rendered support to my research activities: Anna Monreale, Fosca Giannotti, Tiziano Squartini, Diego Garlaschelli, Tommaso Cucinotta, Francesca Chiaromonte, Luca Pappalardo, Salvatore Rinzivillo, Ioanna Miliou, Vittorio Romano, and Silvia Zappulla.

The pleasure and ease of working with my co-authors from institutions outside of the Data Science PhD ecosystem - Felix Opolka, Pietro Lio' from Cambridge University, Vahan Nanumyan from ETH Zurich, and Leandro Tortosa, Jose Vicent, Manuel Curado from the University of Alicante - cannot be emphasized enough.

I would also like to emphatically thank the following people, who supported me personally during this research, and without whom I would not have completed the PhD degree: Mesrop Andriasyan from Politecnico di Milano, Francesco Grotto from the Department of Mathematics of the Scuola Normale Superiore, Giorgio Vinciguerra and Mattia Setzu from the Department of Computer Science at the University of Pisa, and my awesome fellow PhD candidates Giorgio Tripodi, Luca Insolia, Jisu Kim, Cecilia Panigutti, Vasiliki Voukelatou, Elisa Ferrari, and Tommaso Radicioni. Enormous thanks go to Asya, for all the care, warmth, and incredible support she gave in every moment of this undertaking. Finally, I thank my family for believing in me.

Contents

List of Figures

List of Tables

"If you do not care about networks, the networks will care about you, anyway. For as long as you want to live in society, at this time and in this place, you will have to deal with the network society."

Manuel Castells, 2001

Chapter 1

Introduction

Let me introduce the reader to the topic of my PhD work in the Data Science PhD program jointly held by Scuola Normale Superiore of Pisa (SNS), University of Pisa (UniPi), National Research Council (CNR), Sant'Anna School of Advanced Studies (SSSA), and IMT School for Advanced Studies Lucca (IMT).

1.1 What is urban data science?

The origins of urban data science can be tracked to the field of urban analytics which, broadly speaking, develops, utilises and exploits a set of analytical methods and big data to study, understand and predict properties and features of urban environments, formalised as urban systems [236]. However, overly focused on data analysis of huge urban data streams collected from various sources in the city (see Section 2.1.2), urban analytics often lacks a firm methodological foundation and clarity about research paradigms. Given this shortcoming and the unprecedented growth in data-informed research on cities in the past 20 years, Michael Batty calls for the need "to go beyond data analysis *per se*" and establish a theory of the urban - city science - aimed at understanding and explaining spatial and temporal variations of urban phenomena [31].

In [199], the authors establish the term "urban data science" by extending urban analytics and city science to require the incorporation of both quantitative and qualitative methods, and, most importantly, a clear research paradigm and a dia-

logue with other established scientific fields from which to borrow and with which to intertwine our understanding of cities as complex systems.

This aspect is particularly important for the present thesis, as it aims to explore, model, and understand the complex relationships between urban socio-economic characteristics and urban mobility, and will attempt to accomplish this via the scientific apparatus of ***network science***.

In what follows, I will provide a short comprehensive excursion into the essence of this work with which the reader will be offered a road map of the motivation, research questions, methodology, main contributions, and results of this PhD thesis.

1.2 What is this thesis about?

If I were to outline what the present PhD thesis contributes to and what it is essentially about in just a few paragraphs, it would be:

What is this PhD thesis essentially about?

The bulk of research in City Science - a computational understanding of urban systems - can be said to follow two main trajectories: that of *urban structure* and *urban mobility* [30]. The former studies the spatial organisation and morphology of the physical infrastructure, urban space, and the location choices of firms and individuals - also known as urban economics. The latter, on the other hand, studies the individual and collective movement patterns in cities to inform urban and transportation planning.

This thesis attempts to bridge the two trajectories, presenting a methodology for studying the intricate relationships between *urban structure* and *urban mobility* through the lens of **network science**.

Classical urban geography treats urban phenomena as spatial processes in which the relations of urban spaces and locations among each other is limited to so-called spillover effects typically decaying with distance [200]. Urban space is primarily understood in geographical or temporal space, with a locally determined part-to-whole relationships to the city. A major conceptual

view underpinning the present thesis is whether the city can be meaningfully represented in ***relational space***, and, if so, whether this representation can help us extract new kinds of knowledge about cities as systems and reevaluate part-to-whole relationships in the city. In particular, this thesis models the city as a network of mobility flows encoded with origin-destination (OD) matrices, augmented with attributes describing network nodes - city locations (e.g., population density, number of restaurants, real estate prices, etc.) and edges - the various relationships between them (e.g., road distance, travel time, public transport connections, etc.). This attributed urban mobility network is then studied from two methodological viewpoints:

- Network centrality measures
- Urban flow modelling and prediction

First, new network centrality measures based on the Google PageRank algorithm are presented. Just like how Google ranks web pages based on queries, urban planners or policy makers are given the opportunity to "search" the urban network based on specific criteria of desired urban attributes to study the spatio-temporal characteristics of city locations and to inform urban planning and policy making. Then, rankings obtained from the introduced algorithms are used to enhance the modelling and prediction of urban mobility flows. This is achieved via two novel approaches presented in the thesis:

- ***Statistical random graph regression***, in which the observed urban mobility network is considered a realisation from an ensemble of random graphs and is regressed on socio-economic attributes describing the urban environment with the aim of *explaining* the effect each attribute has on the mobility flows.
- ***Graph Neural Networks*** for *predicting* flows to/from a specific location of interest in a city. Imagine a developer aims to build a new commercial center at a given location in the city, knowing the socio-economic attributes of that location in advance (i.e., type of activity, retail volume, parking area, etc.). Can we predict the mobility to/from

that location, given the attributes of the project location and the rest of the mobility network?

Thus, this thesis presents a network-oriented methodology to describe, model, analyze, and predict spatio-temporal aspects of urban mobility flows.

1.3 Thesis Structure

The present PhD thesis is a ***cumulative thesis***. The individual thesis chapters are essentially modified versions of papers published in the course of three years of the PhD program.

In what follows, the thesis structure and a brief synoptic outline of each individual chapter are presented.

Chapter 2 - Background In this Chapter, we prepare ground for the subsequent body of work by motivating *why* studying cities and urban mobility is important, by introducing core concepts we are going to work with, and by providing an extensive review of literature in ***urban structure*** and ***urban mobility***, and baseline techniques and models we will build upon and develop.

Chapter 3 - Thesis Objectives In this Chapter, we define the objectives of our work and formulate the research questions we will attempt to answer.

Chapter 4 - Data Here we describe in detail the methodology and process of building the urban mobility network **dataset** from private car GPS trajectories, augmented with socio-economic attributes describing city locations from various open sources. We explore the dataset, discuss its main features, and set a common ground to be referred to and used by the main body of work represented in the remaining chapters of this thesis.

Chapter 5 - Adapted PageRank and Eigenvector Centrality In this Chapter, we introduce the conceptual framework for centrality measures based on the Google PageRank and discuss the motivation and importance of such centrality measures in an urban context. This Chapter is a modified version

of our paper "Analysis and comparison of centrality measures applied to urban networks with data" in which we present new centrality measures based on the Google PageRank and eigenvector centralities for networks augmented with node attributes, provide a comparative analysis, and apply the discussed measures on the constructed OD flow network in Rome. The presented algorithms offer the possibility to choose the relative importance of the attribute data with respect to the network topology in computing the rankings, thus providing the urban planner with a high flexibility.

Chapter 6 - APA centrality for Biplex urban networks In this Chapter, we build upon the Adapted PageRank Algorithm presented in the previous Chapter 5, by extending it to a biplex mobility network setting. This Chapter is a modified version of our paper "A centrality measure based on the Adapted PageRank Algorithm for multiplex networks with data" which modifies and enhances the previously introduced APA centrality to a multiplex network with the possibility to control for the importance of node attribute data with respect to the network topology in each of the network layers. The presented algorithm is then applied to a case study of the Rome urban OD network with mobility flow and bus connection layers, where different parametrisations of the relative importance of attribute data with respect to the flow network topology are compared.

Chapter 7 - APA centrality for Multiplex urban networks This Chapter is a conceptual follow-up of the previous Chapter 6 which introduced the APA centrality for biplex mobility networks, and naturally extends it to a multilayer setting with the possibility to consider many kinds of relations between city locations at the same time. This Chapter is a modified version of our paper "Understanding mobility in Rome by means of a multiplex network with data" in which we apply the presented APA centrality algorithm for multiplex networks to the same Rome OD flow network extended to include subway connections and topologically short travel distances as additional network layers. Different cases of attribute data importance in each of the layers are then

considered and the possibilities of such a multilayer network approach are highlighted. We will use the rankings obtained from this algorithm to enhance the explanatory and predictive models in Chapters 9 and 10.

Chapter 8 - Spatio-temporal APA centrality In this Chapter, we build upon the APA centrality paradigm introduced in Chapter 5, and explore spatio-temporal characteristics of the distribution of APA centrality values in cities. In particular, this Chapter is a modified version of our paper "Ranking places in attributed temporal urban mobility networks" in which we apply the APA centrality introduced in previous chapters to temporal urban OD flow networks in Rome and London. We introduce several metrics to capture the spatial distribution of "hotspots" with high centrality values across different hours of the day and days of the week in both cities, look at how different socio-economic attributes affect the spatio-temporal behaviour of "hotspots", and provide a comparison among the two cities. The Chapter both presents a methodology for studying urban space with its socio-economic characteristics within a network paradigm and a practical workflow for building and monitoring mobility in a city with specific simple metrics.

Chapter 9 - Explaining mobility from urban attributes We place this Chapter within the paradigm of human mobility modelling by considering the attributed urban mobility network as a multilayer network with each attribute - both node and edge, including the APA centrality rankings computed in previous chapters - transformed into dyadic relationships and considered as a separate layer in the network. Within this framework, the OD flow layer is considered as a realisation from a particular family of statistical random graphs. A regression model respecting the network topology is then proposed, in which the OD flow layer is regressed on the dyadic attribute layers, offering the possibility to *explain* the impact each attribute layer has on the observed urban OD flow network. A temporal regression setting is further presented, with results compared for Rome and London.

Chapter 10 - Urban Mobility Graph Neural Networks This final study Chap-

ter is a modified version of our "Learning Mobility Flows from Urban Features with Spatial Interaction Models and Neural Networks" paper in which we propose several neural network-based architectures, including Graph Neural Networks (GNN) for predicting mobility flows to/from a specific city location, the socio-economic attributes of which are known in advance. We show how the proposed models significantly outperform classical mobility models and more recent machine learning approaches, and demonstrate the impact of the previously computed APA centralities on the prediction accuracy.

Chapter 11 - Conclusion This Chapter concludes the PhD thesis by summarising the formulated research questions and the methodology proposed in the different Chapters for tackling these, discussing the main results and findings, highlighting the advantages of the presented techniques, pointing out drawbacks and shortcomings, and sketching the directions for improvement and future work.

"If a man who can't count finds a four leaf clover, is he lucky?"

Stanislaw Lem

Chapter 2

Background

2.1 Introduction and motivation

2.1.1 Why cities?

Along with the agricultural revolution, the first city-like settlements came to be approximately 10,000 years ago [184] and witnessed unprecedented growth with the industrial revolution. The first city to reach 1,000,000 inhabitants was London, at the heart of the industrial revolution, at the onset of the 19th century. This unleashed the further spread through the end of the 19th and the 20th centuries to other parts of the world. However, while western countries are already largely urban (in 2017, the US population was 82% urban, Australia's 86%, and the majority of the countries in the EU hosted around 80% of their population in cities [115]), the significant part of 'rapid urbanisation' takes place in developing countries. 2005 marked the year in which it was estimated by the U.N. that more than 50% of the world population was inhabiting in cities [186]. It is beyond doubt that urbanisation is not a contingent phenomenon in history, and that the impact cities are going to have on the world is only expected to grow. In fact, the importance of cities in the modern world is already enormous.

First, they play a disproportionately large role in the world's economy. A 2011 report by McKinsey revealed that while the USA and India had respectively 79% and 19% of urban population, their contribution to their countries' respective GDP

was 85% and 39%. NASA data show that urbanised areas cover 6% of the total land surface area in the world, comparable in size to the whole area of the European Union. Notwithstanding their relatively small spatial trace, cities have a considerable environmental impact. The United Nations reported in 2016 that cities were responsible for 74% percent of the world's CO2 emissions.
Representing but the tip of the iceberg, the above-mentioned should suffice to convince anyone of the importance to study and understand cities if we want to make the world we built for ourselves a better place. The unprecedented explosion of urbanisation in developing countries poses significant challenges. Both the cause as well as the solution of some of the most pressing problems in the world unarguably lie in cities. By improving the way cities function we can possibly have a dramatic impact on people's lives. In order to do so however, we need to understand how they function first.

2.1.2 Why data?

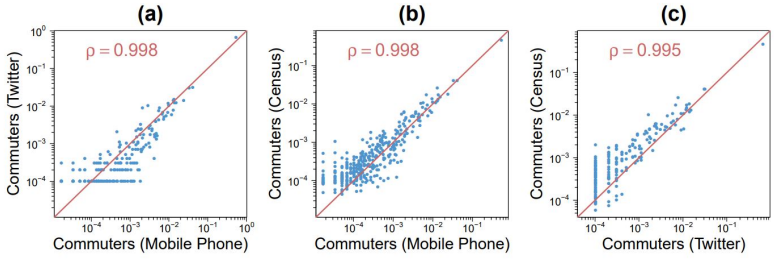
The fundamental game changer in research on cities is data. We now have at our availability huge amounts of data pertaining to virtually all aspects of urban life. Data about many different processes is available at various scales.

At short time scales, we have human mobility data originating from call detail records (CDR) that contain information on the location of individuals at the moment of making a call. Until relatively recently this kind of data was mainly obtained by surveys that had limitations with respect to time and space, whereas CDR or automobile GPS data give a much more precise and real-time overview of mobility in cities. Moreover, the use of Radio-frequency identification (RFID) in subway, bus, and private vehicle transport modes enhance this type of datasets and the growing set of different sensors in urban environments measuring, for instance, air pollution offer the possibility to extend our understanding and modeling of urban mobility.

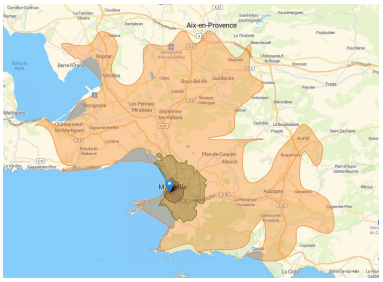
At larger time scales of several months to a year, there is socioeconomic data on such aspects as the income-location relationship, the spatio-temporal change in real-estate prices, etc. Finally, at a very long time scale, the digitization of historical documents such as maps offers the opportunity to study the long-term evolution and

change of urban infrastructure. All the mentioned types of data are indispensable in the process of modelling and studying cities and the revelation of the major forces that propel their evolution.

Another crucial issue has to do with the accuracy of the mentioned new datasets. It is necessary to test and compare them with more conventional methods of obtaining socioeconomic data, for example surveys. In [164] the authors investigated the relationship between various sources of data concerning Origin-Destination flow (OD) matrices[1] describing mobility in Spanish cities with data obtained from Twitter, CDRs, and census data. They showed a valid consistency between the different datasets. The study showed the importance of working with a multitude of different data sources, allowing for cross-checking the obtained results (Figure 2-1a).



(a)



(b)

Figure 2-1: (a) Correlations between data sources [164]. (b) The reachability within 30 minutes by foot, bycicle, and car in the city of Marseille, France [196].

2.1.3 Why mobility?

Mobility is undoubtedly a critical phenomenon in urban environments. In fact, it can be considered as one of the most important mechanisms underlying the structure and dynamics of contemporary cities. Indeed, cities are places where intensive buying,

[1] for a detailed overview of OD flow data, see Section 4

selling or exchanging goods is taking place, where individuals commute to work or meet with other individuals. An obvious means to achieve all this is transportation. Here is where technology enters the picture via the average and maximum velocity of different transportation modes. This average velocity has increased considerably as technology evolved and modified the spatial organization of cities. For instance, as we can see in Figure 2-1b, the reachability of an individual depends on the transportation mode. For a pedestrian, the reachability horizon is typically isotropic and small, whereas the car permits a wider yet anisotropic exploration of space due to the existing infrastructures. The described correlation between the spatial organisation of a city and the available technology at the time has been demonstrated by [18] for American cities. The authors of the study show how many big cities, such as Denver, grew around rail stations which unleashed the development of central business districts. Later automobile-era cities such as Dallas, on the other hand, display a spatial structure primarily conditioned by the highway system.

In terms of mobility, the traditional city center can be regarded as the location that mimimises the average distance to all other locations in the city. As a natural consequence, it has thus historically attracted businesses and residences, leading to competition for the limited space among individuals or firms, which gave rise to the real-estate market. There exists also a well-studied relationship between land-use and accessibility, as was shown half a century ago in [120], and it can be expected that new datasets will certainly offer new possibilities to precisely portray the relation between these and other important factors.

It is certainly neither reasonable nor possible to make an all-encompassing review on all available studies on mobility and our focus in this thesis will be rather on certain specific points. We will firstly describe the general features of urban mobility considering the central quantity in these studies - the origin-destination (OD) matrix - and discussing how to extract useful information from it. Next, we will formulate hypotheses about how to approach the intricate relationships between urban mobility and urban spatial structure.

2.1.4 Why spatial structure?

Morphological aspects of the city, such as the quantitative description and comparison of cities according to their density landscape, spatial organisation, polycentricity, or the clustering variation of their activity centers, have already been studied for a long time in urban geography and spatial economy [18, 39, 261, 213, 233, 255, 116, 38, 159]. However, there seems to be a lack of precision when dealing with the fundamental object of our study: the city. Despite some efforts from urban geographers to build a common ground as to the definitions of a city [219], we still lack an univocal, theoretically sound definition of what a city is. And this is problematic, since statistical results stem from what is deemed as the most suitable definition of the city at the time and context of the study. This in turn influences the ability to generalise research results on cities. If we aim to obtain robust empirical results, compare the results obtained in different countries, we would need to begin thinking about the definition of the system of our study. With a definition of the study object more or less set up, a usual starting point towards our goal of understanding how cities are spatially structured is to study how objects are scattered within it. By objects it is meant buildings, roads, economic activities, but most importantly, people.

The way the distribution of objects in space is traditionally studied is via the study of densities. With a growing scale, however, density profiles become too complicated to comprehend and work with. Several attempts have been made towards approaching this problem in the field of spatial statistics as well as urban form [233, 261, 182, 264]. Authors in them attempt to resolve this issue by proposing simple measures that extract a single index from the density profile. A paramount example is the Moran's I [182] defined as:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}, \tag{2.1}$$

where N is the number of spatial units indexed by i and j; x is the variable of interest; $\bar{x}$ is the mean of x; w_{ij} is a matrix of spatial weights with zeroes on the diagonal (i.e., $w_{ii} = 0$); and W is the sum of all w_{ij}. The Moran's I measures the

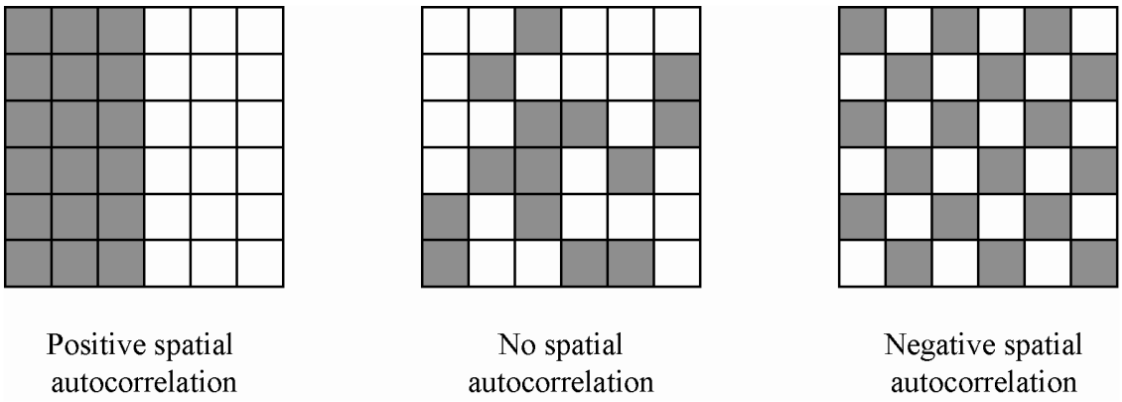


Figure 2-2: Measure of the spatial autocorrelation among spatial units

degree to which similar objects in space tend to cluster together. Its values range from -1 to 1, with -1 corresponding to perfect clustering of dissimilar values, with +1 to perfect clustering of similar values, and with 0 indicating no autocorrelation (perfect randomness, Figure 2-2). It is worth mentioning that Morans's I index is related to the "First Law of Geography" which states that "everything is related to everything else, but near things are more related than distant things." [257]

Another example of a single index, in this case measuring the heterogeneity of population densities in the city is the modified Gini coefficient [264]:

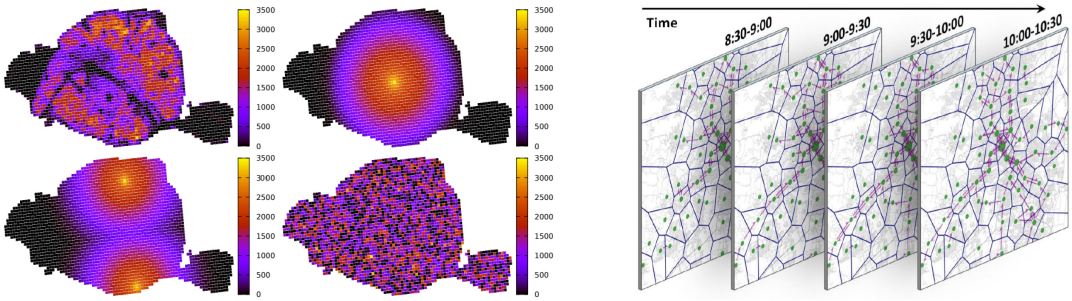
$$G_\alpha = \frac{\sum_{i,j\in\alpha} |P_i - P_j|}{2n_\alpha \sum_{i\in\alpha} P_i}, \tag{2.2}$$

where the sums run over all n_α cells covering the surface of the municipality α.

The Gini coefficient is predominantly used in Economics [110, 111], and was originally proposed to measure the inequality degrees in distributions of wealth and income but has been modified in [264] to capture the level of heterogeneity of population densities. It takes on the value of zero for a city in which the population is uniformly distributed in all grid cells, and is maximum for an extremely concentrated city, with the total population residing in a single grid cell.

A single index is nonetheless too simple to accurately capture complex spatial relationships. For example, as we can see in Figure 2-3, the spatial organisation of the population densities can be reshuffled to obtain different layouts with exactly the same Gini coefficient, demonstrating the inability of this index to capture how

values are organised in space. Although the authors go on to introduce another index compensating for this shortcoming, the example demonstrates the need for more sophisticated representations. Hence, what is needed is rather a meso-scale measure, somewhere between the micro-scale representation (the density profile itself) and the macro-scale representation (a single index summarising the density profile). We conjecture that since 'centers' are themselves a meso-scopic system, their working definition ought to emerge readily from such a representation.



(a) (b)

Figure 2-3: (a) The population distribution in Paris in 2013 and different distributions with exactly the same Gini [264]. (b) Illustration of a trajectory flow map, a dynamic graph of aggregated traffic flows constructed from trajectory data. The presented example is based on bus passenger trajectories obtained in Brisbane, Australia [143].

Until relatively recently these quantitative characterisations of urban form were primarily based on transport surveys, census data, and remote sensing data, all allowing for a fine spatially granular population density and land use estimation, but lacking the same granularity in the temporal dimension. It should be noted here that early studies in urban geography [96, 114] estimated population density at different hours of the day using transport surveys and could trace the morphological and socioeconomic evolution of urban areas during the day. In addition, various traffic surveys in cities around the world have provided an overall outline of the temporal dimension of urban mobility.

However, given their rather coarse temporal resolution and the absence of appropriate data, these studies were not able to study some crucial questions related to dynamical characteristics of the spatial organisation of cities: how does the city's population and/or activity density profile change throughout the day? What is the

spatial distribution of the city's hotspots at different times of the day? How are these hotspots or points of interest (POI) spatially organized? Is there any hierarchy in the spatial organization of hotspots, and if so, is it robust through time? Is there some kind of characteristic distance(s) characterizing the invariant core of a city?

Given the importance and challenges associated with the study of the spatial structure of cities and its non-trivial relationships with urban mobility, this thesis will attempt to approach the above-mentioned issues from a complex network theoretical perspective.

2.1.5 Why complex networks?

The science of networks has been witnessing a rapid development in recent years: the metaphor of the network, with all the power of its mathematical devices, has been applied to complex, self-organized systems as diverse as social, biological, technological and economic, leading to the achievement of several unexpected results in the seminal works of Barabási, Strogatz, Pastor-Satorras and others [15, 249, 209]. Our understanding of spatial networks that are omnipresent in biological, technological and infrastructural systems [181, 27] has seen an unprecedented progress in the recent years. However, notwithstanding a significant amount of research on these kinds of networks, in disciplines covering among others mathematics, physics, biology, and geography, their topological, structural and dynamical properties are not yet completely understood. These networks have proven to be relevant in urban systems [55, 40, 206, 29, 26] where the studies of their structure and topology have revealed particular characteristics of cities as well as shown remarkable statistical properties such as scale invariant patterns across various urban spaces [112, 41, 138]. The street network with its geometry is of particular importance, providing the residents functional connections for navigating various components of the urban area. Different street patterns allow for different levels of efficiency, accessibility, and utilisation of infrastructure [137, 72, 270, 140]. Thus, structural properties of street networks have been the object of several studies [152, 268, 173, 248].

Urban morphology and morphogenesis, activity residence and workplace spa-

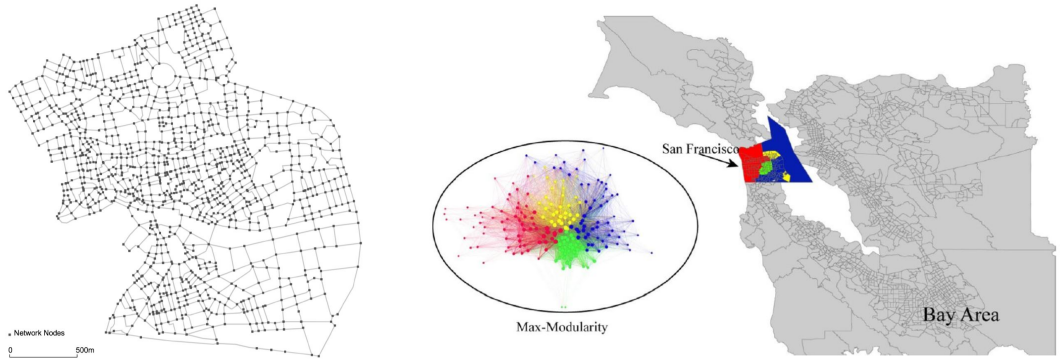
tial distributions, urban sprawl and the evolution of urban networks, are but a few of the important mechanisms that have been systematically studied but that we now hope to comprehend quantitatively. The various network models can be thought of as a simplified abstracted view of cities, which capture important parts of their structure and organization [243] and contain the possibilities to unlock the underlying universal processes behind their formation and development. Apart from modelling the street network as a graph, thus restricting oneself to its planar properties, other network approaches capturing various kinds of relationships, particularly mobility, between different parts of the urban area, have also been experimented with [143, 229, 308, 250, 163, 134, 302, 292] (Figure 7-2). Extracting similar patterns among cities is one of many ways towards the identification of these conjectured underlying processes. One important question, for example, boils down to the mechanisms behind bottom-up 'organic' patterns - which evolve under local constraints - and whether and how they are different from the top-down, planned patterns by a central authority which appear under large scale constraints. This direction of research is by no means new [283, 28], but the recent unprecedented increase of available data such as historical or contemporary digital maps [198, 247] and temporally granular mobility data allow to proceed with large scale cross-sectional models and their evolution in both short and longer period of time (Figure 2-3b).

Some types of networks, like street and road networks, are now more or less adequately described [136, 228, 217, 217, 77, 67, 73]. Because of spatial constraints, they show a peaked degree distribution, large assortativity and clustering coefficient, and the most revealing and valuable characteristic is the spatial distribution of a graph-theoretical measure called betweeenness centrality (see section 2.2.1)

A crucial aspect is that the main instrument for mathematically representing a network - the adjacency matrix - is not sufficient to capture all relevant information about the system. In particular, the spatial distribution of node geometry plays a critical role. A classification of cities according to their street network should then rely on both topology and geometry.

However, the particular relationships between the street network, differently constructed mobility networks along with their evolution over time, and the actual

physical patterns of the city are non-trivial and poorly understood. We believe that results in this direction would open up the possibility for devising methods to the challenging problem of urban mobility prediction and transfer learning to cities with scarce data. Although some results, primarily in traffic and travel demand forecasting [267, 295] and graph-based transfer learning have been achieved [123, 133, 160], we still lack systematic methods for mobility-informed prediction of urban spatio-temporal dynamics. This thesis will attempt to tackle this critical problem.



(a) (b)

Figure 2-4: (a) Map and network of the city of Murcia, Spain. [1] (b) The community structure of San Francisco urban regions. Different color represents different traffic community, the spatial partition among the four communities are quite obvious [250].

2.2 Network centrality measures

A crucial set of instruments indispensable to the study of most kinds of networks are network centrality measures. Centrality measures serve to quantify the idea that in a network some nodes are more important (central) than others.

As mentioned before, the science of networks has witnessed a dramatic increase in its applications in systems spanning social, economic, technological and other disciplines. In particular, the issue of centrality in networks has remained pivotal, since its introduction in a part of the studies of humanities named structural sociology [274]. The idea of centrality was first applied to human communication by Bavelas [32] who was interested in the characterization of the communication in small groups

of people and assumed a relation between structural centrality and influence and/or power in group processes. Since then, various measures of structural centrality have been proposed over the years to quantify the importance of an individual in a social network [32]; and the issue of centrality has found many applications also in biology and technology. Currently, centrality is a fundamental concept in network analysis though with a different purpose: while in the past the role and identity of central nodes were investigated, now the emphasis is more shifted to the distribution of centrality values through all nodes. Centrality, as such, is treated like a shared resource of the network community, like wealth in nations, with the focus being on the homogeneity and/or heterogeneity of distributions [15]. In urban planning and design, as well as in economic geography, centrality, though under different terms like accessibility, transport cost or effort, has entered the scene stressing the idea that some places are more important than others because they are more central [278]; all these approaches have been following a primal representation of spatial systems, where punctual geographic entities - street intersections, settlements - are turned into nodes and their linear connections - streets, infrastructures - into edges. A pioneering discussion of centrality as inherent to urban design in the analysis of spatial systems has been successfully operated after Hillier and Hanson seminal work on cities [125] since the late 1980s. Space Syntax, the proposed methodology of urban analysis, has been raising growing evidence of the correlation between the so-called integration of urban spaces, a closeness centrality in all respects, and phenomena as diverse as crime rates, pedestrian and vehicular flows, retail commerce vitality and human way-finding capacity [124]. The Space Syntax approach follows a dual representation of street networks where streets are turned into nodes and intersections into edges. An outcome of the dual nature of Space Syntax is that the node degree is not limited by physical constraints, since one street has a conceptually unlimited number of intersections; this property makes it possible to witness the emerging of power laws in degree distributions [136, 228] that have been found to be a distinct feature of other nongeographic systems [15, 249, 209, 23]. On the other hand, the dual character leads Space Syntax to the abandonment of metric distance: a street is one node no matter its real length. Metric distance, conversely, was the core of

most territorial studies [231] and is a key ingredient of spatial networks.

When dealing with urban street patterns, centrality has been investigated in relational (topological) networks only, neglecting a fundamental aspect of the system as the geography. In the majority of past approaches a city is transformed into a spatial graph by mapping the intersections into the graph nodes and the roads into links between nodes. By using a set of different centrality indices (multiple centrality assessment [77, 76]), extended or defined on purpose for spatial graphs, it is possible to spot the relevant places of a city. By relevant places it is meant places closer to other places (closeness centrality), places that are structurally made to be traversed (betweenness centrality), places whose route to other places deviates less from the virtual straight route (straightness centrality), and places whose deactivation affects the structural properties of the system (information centrality). Apart from the mentioned purely structural centrality measures aaplied to urban networks, attention has recently been drawn towards more sophisticated centrality concepts allowing for integration of valuable information about different places in the city in measuring their respective centralities. Such measures include, among others, the modified Google PageRank and Eigenvector centralities (see section 2.2.2). Moreover, by investigating how centrality is distributed among the nodes of the graph, how the different centrality indices are correlated, and how they evolve in the temporal dimension, it is possible to study urban dynamics and also characterise classes of cities [76].

2.2.1 Multiple centrality assessment

The multiple centrality assessment relies on three basic principles [77, 76] as follows:

label=(0) primal graphs, rather than dual;

lbbel=(0) metric distance, rather than topological;

lcbel=(0) many centrality indices, rather than mainly closeness

The following is a list of common centrality measures we. The definitions are given in terms of an undirected, weighted graph G, of N nodes and K edges. The

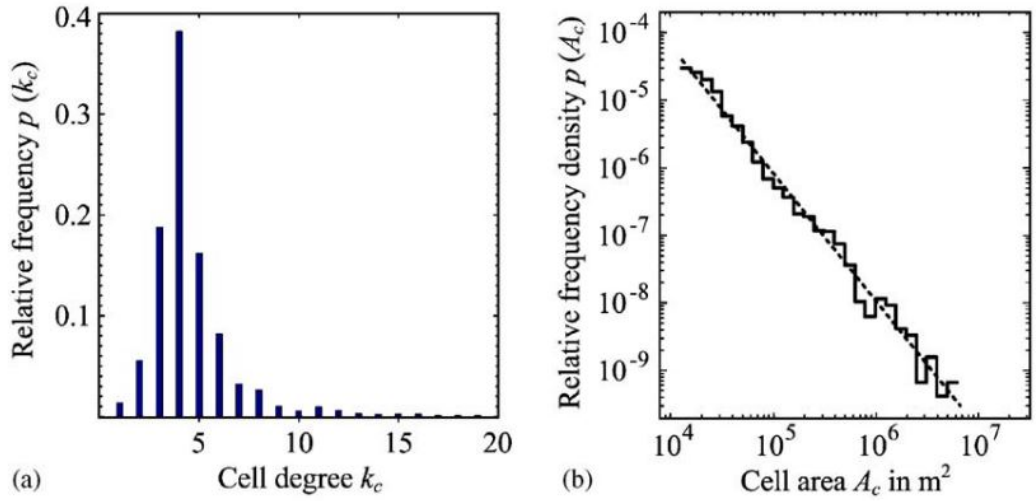


Figure 2-5: (a) Degree distribution of degrees for the road network of Dresden. (b) The frequency distribution of the cells surface areas A obeys a power law with exponent $\alpha \approx 1.9$ (for the road network of Dresden) [152].

graph is described by the adjacency $N \times N$ matrix A, whose entry a_{ij} is equal to 1 when there is an edge between i and j and 0 otherwise, and by a $N \times N$ matrix L, whose entry l_{ij} is the value associated to the edge: for planar street networks usually the metric length of the street connecting i and j; for mobility networks usually the amount of traffic flowing from i to j in a fixed amount of time.

Degree centrality

Degree centrality, C^D, is the simplest definition of node centrality. It is based on the idea that important nodes have the largest number of ties to other nodes in the graph. The degree centrality of i is defined as [274, 193, 105]:

$$C_i^D = \frac{\sum_{j=1}^{N} a_{ij}}{N-1} = \frac{k_i}{N-1}, \tag{2.3}$$

where k_i is the degree of node i, i.e., the number of nodes adjacent to i. Degree centrality is not particularly interesting in primal urban networks where node degrees are limited by geographic constraints and show a peaked distribution. For example, in a study of 20 German cities, Lämmer et al. [152] showed that most nodes have four neighbors (the full degree distribution is shown in Figure 2-5) and that the degree rarely exceeds 5 for various world cities [61].

Closeness centrality

Closeness centrality, C^C, measures to which extent a node i is near to all the other nodes along the shortest paths, and is defined as [274, 230]:

$$C_i^C = \frac{N-1}{\sum_{j \in G, j \neq i} d_{ij}}, \tag{2.4}$$

where d_{ij} is the shortest path length between i and j, defined, in a weighted graph, as the smallest sum of the edges length throughout all the possible paths in the graph between i and j.

Betweenness centrality

Betweenness centrality, CB, is based on the idea that a node is central if it lies between many other nodes, in the sense that it is traversed by many of the shortest paths connecting couples of nodes. The betweenness centrality of node i is [105, 146]:

$$C_i^B = \frac{1}{(N-1)(N-2)} \sum_{s \neq t \in V} \frac{\sigma_{st}(i)}{\sigma_{st}}, \tag{2.5}$$

where σ_{st} is the number of shortest paths going from nodes s to t and $\sigma_{st}(i)$ is the number of these paths that go through i [104].

Straightness centrality

Straightness centrality, C^S, originates from the idea that the efficiency in the communication between two nodes i and j is equal to the inverse of the shortest path length d_{ij} [156]. The straightness centrality of node i is defined as

$$C_i^S = \frac{1}{N-1} \sum_{j \in G, j \neq i} d_{ij}^{Eucl} / d_{ij}, \tag{2.6}$$

where d_{ij}^{Eucl} is the Euclidean distance between nodes i and j along a straight line, and there has been adopted a normalization proposed for geographic networks [265]. This measure captures the extent to which the connecting route between nodes i and j deviates from the virtual straight route.

Information centrality

Information centrality, C^I, is a measure introduced in [158], and relating a node importance to the ability of the network to respond to the deactivation of the node. The network performance, before and after a certain node is deactivated, is measured by the efficiency of the graph G [156, 157]. The information centrality of node i is defined as the relative drop in the network efficiency caused by the removal from G of the edges incident to i,

$$C_i^I = \frac{\delta E}{E} = \frac{E[G] - E[G']}{E[G]} \tag{2.7}$$

where the efficiency of a graph G is defined as

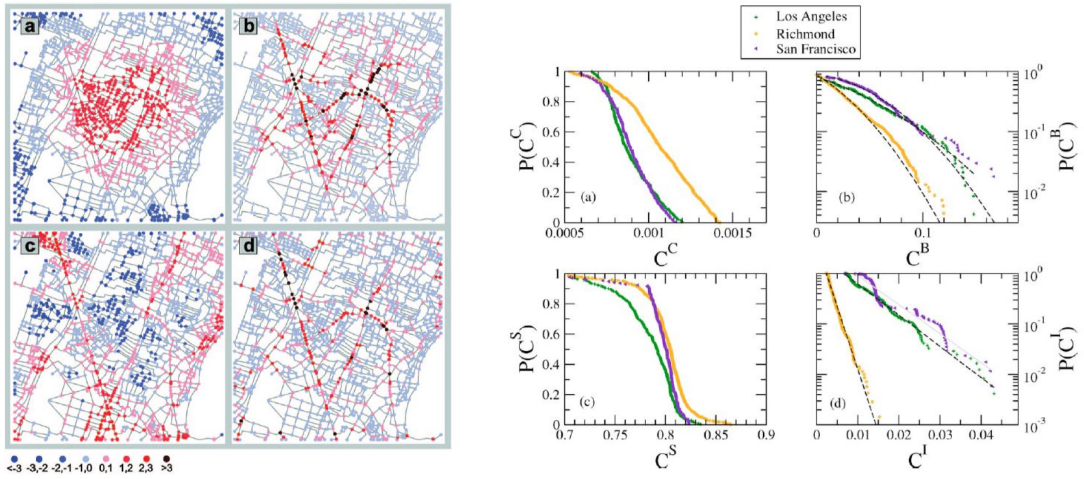
$$E[G] = \frac{1}{N(N-1)} \sum_{j \in G, j \neq i} d_{ij}^{Eucl} / d_{ij} \tag{2.8}$$

and where G' is the graph with N nodes and $K - k_i$ edges obtained by removing from the original graph G the edges adjacent to node i. An advantage of using the efficiency to measure the performance of a graph is that $E[G]$ is finite even for disconnected graphs [76].

As shown in [76], Closeness, straightness, and betweenness centrality distributions, where the cumulative distribution $P(C)$ is defined as

$$P(C) = \int_C^{+\infty} \frac{N(C')}{N} dC', \tag{2.9}$$

where $N(c)$ is the number of nodes with centrality equal to C, are quite similar in both self-organized and planned cities, despite the diversity of the two cases in socio-cultural and economic terms could not be deeper. On the other hand, the information centrality distributions notably differentiate self-organized cities from planned ones, being broad-scale (power law) in the first case, and single-scale (exponential) in the second case (Figure 2-6**b**).



(a) (b)

Figure 2-6: **a.** Thematic color map representing the spatial distributions of centrality in Cairo, an example of a largely self-organized city. The four indices of node centrality, (a) closeness C^C, (b) betweenness C^B, (c) straightness C^S, and (d) information C^I, used in the MCA, are visually compared over the primal graph. Different colors represent classes of nodes with different values of the centrality index. The classes are defined in terms of multiples of standard deviations from the average, as reported in the color legend. **b.** Cumulative distributions of (a) closeness C^C, (b) betweenness C^B, (c) straightness C^S, and (d) information C^I for three planned cities, Los Angeles, Richmond, and San Francisco. The dashed lines in panels (b) are Gaussian fits to the betweenness distributions, while the dashed lines in panel (d) are exponential fits to the information centrality. [76].

2.2.2 Ranking measures

As already mentioned, the classic centrality measures do not allow us, in a simple way, to work with the data associated with a network. Therefore, it becomes necessary to introduce centrality measures which account for two factors: first, the network topology and, moreover, the importance of existing data, allowing to differentiate places with external values other than those related to topology.

Eigenvector centrality

Eigenvector centrality, denoted by C^E, was proposed by Bonacich [50] to measure the influence of a node in a network from the importance of its connections. Degree centrality gives an idea about the number of connections a vector has. However, not all the connections or links are equally important. Therefore, somehow we should weight the importance of each node connection. If it is assumed that a node is more central if it is in relation with nodes that are themselves central, it can be argued that the centrality of the nodes of a graph does not only depend on the quantity of its adjacent nodes, but also on their value of centrality.

In [1], the authors denote the centrality of node n_i by x_i, allowing to take into account the importance of each node's links by making x_i proportional to the average of the centralities of i's network neighbours:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j, \tag{2.10}$$

where λ is a constant. Defining the vector of centralities $\mathbf{x} = (x_1, x_2, ...)$, they rewrite equation (9.4) in matrix form as

$$\boldsymbol{A} \cdot \mathbf{x} = \lambda \mathbf{x} \tag{2.11}$$

It is clear from the expression (10.9) that $\mathbf{x}$ is an eigenvector of the adjacency matrix $\boldsymbol{A}$ associated to the eigenvalue λ. As $\boldsymbol{A}$ is the adjacency matrix of an undirected graph and $\boldsymbol{A}$ is non-negative, it can be shown (using the Perron-Frobenius theorem) that there exists an eigenvector corresponding to the largest eigenvalue (the

authors denote it by λ_1) with only non-negative (positive) entries. This eigenvector constitutes a ranking of the nodes in the graph.

In [1], the authors go on to construct a data matrix $\boldsymbol{D}$ by collecting four types of different activity data for each node (number of bars, shops, offices, and malls):

$$\boldsymbol{D} = \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} & d_{1,4} \\ d_{2,1} & d_{2,2} & d_{2,3} & d_{2,4} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n,1} & d_{n,2} & d_{n,3} & d_{n,4} \end{pmatrix} \tag{2.12}$$

Thus, the matrix $\boldsymbol{D}$ is given by has n rows, corresponding to the n nodes of the urban network studied, and has 4 columns, each corresponding to the four different types of data that were collected. The authors then go on to construct a weight matrix and use it to calculate the vector of rankings for the nodes of the urban network (for details, see [1]).

Google PageRank centrality

Nowadays, it is essential to have a fast and reliable ranking system for the websites in the World Wide Web to bring order to the chaos of data. For a deeper discussion of the structure of the net see, for example, [91, 234]. For an excellent explanation of the different search engine models that have appeared in the recent decades, we refer the reader to [154].

The Web's hyperlink structure forms a massive directed graph, where the nodes in the graph represent Web pages and the directed arcs or edges represent the hyperlinks. The hyperlinks into a page are called *inlinks* (or incoming edges) and point into nodes. The hyperlinks that point from nodes are called *outlinks* (outgoing edges). If there are multiple links from one page to another, they are considered as a single link. Finally, links to the page itself are not considered.

The first search engines, back in the 90s, based management results pages on the number of times the search text appeared on each page, regardless of other factors. This system did not provide suitable results in many cases, since the fact that a page often repeats a word in its content does not guarantee its relevance within

their field. In a nutshell, PageRank's thesis is that a Web page is important if it is pointed to by other important pages [202].

The PageRank method was proposed to compute a ranking for every Web page based on the graph of the Web, that is, PageRank constitutes a global ranking of all Web pages, regardless of their content, based solely on their location in the Web's graph structure. The purpose of the method is obtaining a vector, called PageRank vector, which gives the relative importance of the pages. Since this vector is calculated based on the structure of the Web connections, it is said to be independent of the request of the person performing the search.

If we denote the web-graph as $G = (V, E)$, where V is the set of webpages on the internet, and E is the set of hyperlinks between them, then the classical Google PageRank is the solution $\pi(n) = (\pi_i(n))_{i=1,..,|V_n|}$ to the system

$$\pi_i(n) = c \sum_{j \to i} \frac{\pi_j(n)}{outdeg\ j} + \frac{1-c}{|V_n|}, \qquad i = 1, ..., |V_n|, \tag{2.13}$$

where c is a weighting parameter.

Some modifications of this method have been proposed in [122, 225]. An application of PageRank centrality for describing the urban network has relatively recently been proposed in [6].

2.3 Human mobility models

Planning and managing city and transportation infrastructures requires understanding the relationship between urban mobility flows and spatial, structural, and socio-economic features associated with them. There exists extensive literature addressing this problem ranging from the classical gravity model and its modifications [279, 101] to the more recent spatial econometric interaction models [166] and the non-parametric radiation models [235] that attempt to characterise cross-sectional origin-destination (OD) flow matrices. Furthermore, various neural network-based models have been proposed for predicting temporal OD flow matrices [66, 258].

In this section, we will briefly present the most widely used models for human

mobility flows, discuss their properties, underlying assumptions, advantages and disadvantages, and prepare ground for two novel frameworks proposed in this thesis, namely, a temporal network regression for *explaining* urban mobility flows from urban socio-economic attributes in Chapter 9, and several Graph Neural Network (GNN) architectures for *predicting* OD flows to a location of interest in Chapter 10.

2.3.1 Gravity models

The theoretical framework for estimating flows between locations in space has been put forth by Wilson (1971) through a family of spatial interaction models and extended and elaborated by (Fotheringham, O'Kelly 1989) Consider the basic gravity model[cit.]:

$$F_{ij} = k\frac{M_i^{\alpha} N_j^{\beta}}{d_{ij}^{\gamma}} \tag{2.14}$$

where F is an $m \times n$ matrix of origin-destination flows between m origins (subscripted by i) and n destinations (subscripted by j), M is an $m \times l$ matrix of l origin attributes describing the "emissivity" of i, N is an $n \times l$ vector of l destination attributes describing the attractiveness of j, d is an $m \times n$ matrix describing the cost of travelling from i to j (usually distance or time), k is a scaling factor, α is a $l \times 1$ vector of parameters denoting the effect of l origin attributes on the observed flows, β is a $l \times 1$ vector of parameters denoting the effect of l destination attributes on the observed flows, γ is a parameter denoting the effect of travel costs on the observed flows.

Given F, M , N, and d, the model parameters, which describe how each model component contributes to explaining the observed flows (F), can be estimated and used to predict unobserved flows.

Using an entropy-maximizing approach, Wilson extended the Gravity model by proposing a more elaborated family of spatial interaction models (Wilson 1971): *unconstrained*, *production-constrained*, *attraction-constrained*, and *doubly constrained*. The latter aim at assigning flows to origin-destination pairs by finding the most probable configuration of flows out of all possible configurations, without making

any additional assumptions.

The unconstrained model does not preserve the total in- nor out-flows during model calibration. The production-constrained and attraction-constrained models preserve the number of total in-flows or out-flows, respectively, and hence are used for assigning flows either to a set of origins or to a set of destinations. The doubly-constrained model, on the other hand, preserves both the in- and out-flows at each location during parameter estimation. The models' predictive power increases as more built-in information (i.e. total in or out-flows) are accounted for, leading to the doubly-constrained model usually performing best.

The models are obtained by using conventional optimization techniques imposing constraints on the total inflows and outflows at each location. For instance, the doubly-constrained model can be formulated as follows:

$$F_{ij} = U_i V_j O_i D_j f(d_{ij}) \tag{2.15}$$

subject to

$$U_i = 1/\sum_j V_j D_j f(d_{ij}), \quad \text{and} \quad V_j = 1/\sum_i U_i O_i f(d_{ij}), \tag{2.16}$$

where O_i is an $n \times 1$ vector of the total number of out-flows from origin i, D_j is an $m \times 1$ vector of the total number of in-flows to destination j, U_i and V_j are $n \times 1$ and $m \times 1$ vectors respectively, ensuring that the total out- and in-flows are preserved in the model predictions, $f(d_{ij})$ is usually referred to as the distance-decay function, most commonly given by power function $f(d_{ij}) = d_{ij}^{\gamma}$, where γ is expected to take on a negative value.

It is worth noting that the scaling factor in equation (2.14) is absent in all of the maximum entropy models because an imposed total trip constraint is implied in their derivation making such a necessity redundant (Fotheringham, O'Kelly 1989). Also, in the doubly-constrained maximum entropy model, the values for U_i and V_j depend on each other and hence need to be computed iteratively.

The parameters of spatial interaction models are often estimated via linear programming, non-linear optimization, or, more commonly, through linear regression.

By taking the natural logarithm of both sides of, for instance, the basic gravity model, the so-called log-linear or log-normal model is obtained:

$$\ln F_{ij} = k + \alpha \ln M_i + \beta \ln N_j - \gamma \ln d_{ij} + \epsilon, \tag{2.17}$$

where ϵ is a normally distributed error term with 0 mean. The constrained versions of the model can be obtained by incorporating fixed effects for the origins (production-constrained), destinations (attraction-constrained) or both (doubly-constrained).

However, the log-normal gravity model suffers from such serious drawbacks as failing to capture the discrete nature of flows of vehicles or individuals; the flows are usually not distributed normally; due to estimating the logarithm of flows the predictions are downward biased; the inability to handle zero flows due to the logarithmic framework. To partially mitigate the mentioned shortcomings, a Poisson log-linear regression for the family of spatial interaction models was formulated (Flowerdew, Aitkin 1982, Flowerdew, Lovett 1988). This model considers the number of flow units between i and j to be sampled from a Poisson distribution with mean, $\lambda_{ij} = T_{ij}$, where λ_{ij} is logarithmically linked to the linear combination of features, yielding the unconstrained Poisson log-linear gravity model,

$$F_{ij} = \exp\left(k + \alpha \ln M_i + \beta \ln N_j - \gamma \ln d_{ij}\right) \tag{2.18}$$

The constrained versions in the family of spatial interaction models can be obtained by including fixed effects for the balancing factors as in equation 2.16. In particular, for the *doubly constrained* model this will yield

$$F_{ij} = \exp\left(k + \alpha + \beta - \gamma \ln d_{ij}\right), \tag{2.19}$$

where α and β are origin and destination fixed effects, respectively, to the same effect as including balancing factors as constraints (Tiefelsdorf, Boots 1995). The introduction of the Poisson regression model partially mitigates the drawbacks mentioned above. It also allows the avoid the iterative computation of the balancing constraints in the previously available models (Fotheringham, O'Kelly 1989). Parameter estimation for the Poisson regression models is usually conducted within

the framework of generalized linear models (GLM) using iteratively weighted least squares (IWSL), which guarantees convergence to the parameter maximum likelihood estimates (Nelder, Wedderburn 1972).

2.3.2 Poisson regression

With the aim to extend the modelling to include a larger set of potential explanatory variables within the context of the wider Generalized Linear Model framework, we first consider the number of flow units from i to j to be a random variable drawn from a Poisson distribution, with a mean rate potentially influenced by the array of covariates. These explanatory variables include but are not limited to public transport connections, real estate price averages, travel time, velocity, traffic activity correlations, residential-business, residential-airport, residential-school, and other types of relations between pairs of locations in the city.

Within this framework, our model becomes

$$F_{ij} = F_{i_{out}} F_{j_{in}} \exp\left\{\beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{kij}\right\}, \tag{2.20}$$

where the Poisson incidence rate $\frac{F_{ij}}{F_{i_{out}} F_{j_{in}}}$ is determined by the set of k regressor covariates (the X's). Since the flow F_{ij} is *exposed* to the number of possible in- and out-flow combinations $F_{i_{out}} \times F_{j_{in}}$ between i and j, we include the latter as *exposure* in the model.

The most common approach to measuring the overall performance of the fitted model is by using the Pearson statistic chi-square test:

$$\chi^2 = \sum_{i,j} \frac{\left(F_{ij} - \exp\left\{\mathbf{X}_{ij}\hat{\beta}\right\}\right)^2}{\exp\left\{\mathbf{X}_{ij}\hat{\beta}\right\}}, \tag{2.21}$$

which is approximately chi-square distributed with $n - k$ degrees of freedom. If the Pearson statistic chi-square test is rejected at a chosen significance level, it implies a statistically significant lack of fit. If the test is not rejected, there is no evidence of lack-of-fit.

This same test is typically also used as a test for overdispersion. Since the

Poisson distribution has its mean equal to its variance, this enters as an assumption in our modelling of flows as count data drawn from a Poisson distribution. If the model is correct, the expected value of the statistic should be $n - k$, the degrees of freedom. However, in many real-world settings, this assumption does not hold, so the variance is assumed to be a factor of the mean.

This is often tested by looking at the Pearson residuals which correct the unequal variance in the raw residuals, $r_{ij} = F_{ij} - \exp\left\{\mathbf{X}_{ij}\hat{\beta}\right\}$ by dividing by the standard deviation:

$$p_{ij} = \frac{r_i}{\sqrt{\hat{\phi}\exp\left\{\mathbf{X}_{ij}\hat{\beta}\right\}}}, \tag{2.22}$$

where $\hat{\phi}$ is a dispersion parameter to help control overdispersion:

$$\hat{\phi} = \frac{1}{n-k}\sum_{i,j}\frac{\left(F_{ij} - \exp\left\{\mathbf{X}_{ij}\hat{\beta}\right\}\right)^2}{\exp\left\{\mathbf{X}_{ij}\hat{\beta}\right\}}. \tag{2.23}$$

The obtained Pearson residuals can then be inspected in diagnostic plots against the fitted means, which will show whether overdispersion should be addressed with other models. A common solution is offered by the Negative Binomial regression discussed below.

2.3.3 Negative Binomial regression

The Negative Binomial regression is a generalization of the Poisson regression in which the restrictive assumption about the equality of the mean and the variance is loosened. Although usually defined in terms of a sequence of Bernoulli trials, it is convenient to regard the Negative Binomial distribution as a mixture distribution with samples drawn from a Poisson distribution the mean of which is itself a Gamma distributed random variable. This allows the mean-variance relationship in the Negative Binomial distribution to be controlled through a continuous positive dispersion parameter, α:

$$\sigma^2 = \mu + \alpha g(\mu_i), \tag{2.24}$$

where $g(\cdot)$ is a known function, most commonly $g(\mu) = \mu^2$. We see that the Poisson behaviour is recovered as α tends to zero. The value of the parameter α enters as an input to the model, raising the question of how to select it correctly. With the aim of detecting and evaluating overdispersion more reliably than the Pearson statistic, a statistical test estimating α has been proposed by [60]. It simply uses the Poisson model's fitted values $\widehat{\mu_i}$ and performs an auxiliary OLS regression without intercept:

$$\frac{(y_i - \widehat{\mu}_i)^2 - y_i}{\widehat{\mu}_i} = \alpha \frac{g(\widehat{\mu}_i)}{\widehat{\mu}_i} + \epsilon_i, \tag{2.25}$$

where the left-hand side is treated as the response variable, α is the unknown parameter, and ϵ_i is an error term. The fitted α coefficient has a Student's t distribution which lends itself to constructing a confidence interval for selecting a suitable α for the Negative Binomial Regression.

2.3.4 Spatial autoregressive models

A major concern in the modelling scenarios discussed so far are the complex interactions often caused by spatial dependencies and non-stationarity. The former arises from spill-over effects from a location to its neighbourhoods, while the latter is caused by the influence of independent variables varying across space. These issues have been addressed in literature by spatial autocorrelation and geographically weighted modelling techniques [100, 166, 79, 297].

Although in all the discussed models spatial interdependence among observations has been latently accounted for by including the network distance among the covariates, this certainly does not capture how origin or destination cells might affect the flows to or from their geographical neighbours. This obvious shortcoming of the hitherto considered models is amenable to building the spatial interdependence structure by the approach proposed by [166]. Modifying this approach, we take a typical n by n first-order contiguity matrix $\mathbf{D}$, weight it by the observed OD matrix

$\mathbf{F}$, and row-standardize it to obtain $\mathbf{W} = \widetilde{\mathbf{DF}}$, reflecting relations among the n cells. *Destination − based spatial dependence*, reflecting the intuition that flows from an origin to a destination may affect flows to nearby destinations, can be captured by the the $n^2 \times n^2$ row-standardized spatial weight matrix $\mathbf{W}_d$ which can be obtained from $\mathbf{W}$ by the Kronecker product $\mathbf{W}_d = \mathbf{I}_n \otimes \mathbf{W}$:

$$\mathbf{W}_d = \begin{pmatrix} \mathbf{W} & \mathbf{0}_n & \dots & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{W} & \mathbf{0}_n & \vdots \\ \vdots & \mathbf{0}_n & \ddots & \mathbf{0}_n \\ \mathbf{0}_n & \dots & \mathbf{0}_n & \mathbf{W} \end{pmatrix}, \tag{2.26}$$

where $\mathbf{0}_n$ represents an $n \times n$ matrix of zeros. The $n^2 \times n^2$ spatial weight matrix $\mathbf{W}_d$ captures flow relations between an origin and the neighbors of the destination. Similarly, an *origin − based spatial dependence* can be modelled simply by $\mathbf{W}_o = \mathbf{W} \otimes \mathbf{I}_n$. The latter captures weighted average flows from neighbours of each origin to each of the destinations. Thus, the new spatially adjusted OD flow matrix becomes $F_{sp} = vec^{-1}(\mathbf{W}_d \mathbf{W}_o y)$, where $y = vec(\mathbf{F})$.

$\mathbf{W}_d \mathbf{W}_o y$) can be interpreted as a successive spatial filter. Indeed, the order of applying the spatial filters $\mathbf{W}_d$ and $\mathbf{W}_o$ does not matter due to the mixed-product rule for Kronecker products.

Modifying the spatial autoregressive model proposed in [166], this framework thus results in:

$$F_{sp_{ij}} = F_{sp_{i_{out}}} F_{sp_{j_{in}}} \exp\left\{\beta_1 X_{1ij} + \beta_2 X_{2ij} + \dots + \beta_k X_{kij}\right\}, \tag{2.27}$$

for which we can run the models described in 2.3.2 and 2.3.3 and which has been shown to capture the variability in spatial OD flows significantly better than all the models discussed earlier.

2.3.5 The Huff model

Another approach within the spatial interaction modelling paradigm is the Huff model and its extensions [129]. Originally developed mainly for retail location choice and turnover prediction, they represent a probabilistic formulation of the gravity model. The Huff model considers OD flows as proportional to the relative attractiveness and accessibility of the destination compared to other competing destinations. The probability P_{ij} of a consumer at location i of choosing to shop at a retail location j is framed as:

$$P_{ij} = \frac{A_j^{\alpha} D_{ij}^{-\beta}}{\sum_{j=1}^{n} A_j^{\alpha} D_{ij}^{-\beta}}, \tag{2.28}$$

where A_j is a measure of attractiveness of retail location j, such as area or a linear combination of different features, D_{ij} is the distance between locations i and j, α and β, estimated from empirical observations, are attractiveness and distance decay parameters, respectively.

Along with traditional gravity methods, the Huff model and its variations have found their way to numerous applications including location selection of movie theaters [80], a university campus [57], or the analysis of spatial access to health care [266].

However, these models suffer from too restrictive assumptions such as considering the ratio of the probabilities of an individual selecting two alternatives as being unaffected by the introduction of a third alternative. Although the competing destinations model [98] has overcome this, it has the disadvantage of considering either spatial agglomeration or competition effects, ignoring the fact that they can coexist in the same location. Even though a number of extensions to the Huff model and the gravity framework in general have been proposed to overcome spatial non-stationarity and to include a larger array of features affecting the flows [82, 167], this family of models, along with the non-parametric radiation and population-weighted opportunities model, have demonstrated to fall short of high predictive capacity particularly at the city scale [179, 168, 288].

More recently, machine learning, particularly a Random Forest approach, has

shown promising results in reconstructing inter-city OD flow matrices [244]. However, its performance on intra-urban flow data remains to be tested.

Despite the proven utility of the discussed models, when applied to real intra-urban mobility flow data, we shall see that the gravity, Poisson, Negative Binomial, Spatial autoregressive, and Huff models fail to capture enough of the relationships between urban mobility and socio-economic and spatial attributes in cities in order to warrant a convincing explanation. We will extend the discourse initiated by these models, extend it, and propose two new modelling frameworks for *explaining* and *predicting* urban mobility flows.

2.3.6 Machine learning models

In the previous sections of this Chapter, we discussed classical modelling and statistical learning frameworks for human mobility models. However, recent advances in machine learning have also contributed to the task of modelling OD flows in geographic areas. In particular, a notable method in this respect has been proposed in [244], in which the authors use a well-known random forest algorithm for reconstructing OD flows between cities, each of which is described by a set of attributes. The problem of estimating OD flows has also been addressed with neural network methods [171]. As flows are most naturally modelled by graphs, most work has focused on the use of graph neural networks for flow estimation. In this regard, it is worth mentioning a few words on this recently developed class of neural network models. An early neural network model for graph structured data has been suggested in [232]. Later work has specifically focused on generalising Convolutional Neural Networks from the domain of regular grids to the domain of irregular graphs [56, 86]. One of the most commonly used graph neural network models is the Graph Convolutional Neural Network (GCN) proposed in [145].

Graph neural networks have previously been applied to urban planning tasks. In [66], they have been used to predict the flow of bikes within a bike sharing system. In this approach, mobility flows are modelled as node-level features, which requires a particular neural network model treating graphs in their entirety and does not allow to predict flows between specific pairs of nodes. Although [271] uses graph

neural networks to predict flows between parts of a city, their model operates on spatio-temporal data and focuses on the temporal aspect of the data. Beyond flow prediction, in [307], a graph neural network model has been proposed for building site selection. A broader overview of machine learning methods applied to the task of urban flow prediction is given in [284].

"There's no sense in being precise when you don't even know what you're talking about."

John von Neumann, 1958

Chapter 3

Thesis Objectives

In this chapter we define the goals and derive the specific questions to be addressed in our research. In Chapter 1 we introduced the topic of research, outlined the overarching objective of the PhD thesis, and described the thesis structure. Let us now formulate the specific objectives and research questions we will tackle with the techniques proposed in the subsequent chapters.

1. The city as a network of relations

Are PageRank-based centrality measures in mobility flow network models - with the mathematically formulated self-referential principle that a web page is important if other important web pages point at it - capable of capturing meaningful part-to-whole relationships in the city?

2. Urban socio-economic attributes and network centrality

Can the PageRank-based network centrality measures give insight into the role socio-economic attributes describing city locations play in the spatial organisation of cities?

3. Multilayer network centrality

Can the mentioned PageRank-based centrality in urban mobility flow networks be extended to multiple types of relations between city locations, such as distance, time, speed, public transport connections, etc., forming a multilayer

network of relations in cities and their contribution be assessed?

4. Temporal network centrality

What can a temporal analysis of the behaviour of urban location centrality tell us about the spatio-temporal characteristics of various socio-economic factors in the city?

5. Explaining mobility flows from urban attributes

Classical models explaining why and how people and goods move in a city do not account for the network structure of mobility flows. Can a regression model be devised that inherently respects the network structure of urban flows, in which the latter can be *explained* by an array of socio-economic attributes describing city locations and the relations between them?

6. Centrality measures and urban mobility

Can the above-mentioned centrality measures enhanced with urban socio-economic attributes be informative for *explaining* urban mobility flows within the network regression framework?

7. Socio-economic attributes and urban flow prediction

Is it possible to construct a network-based city model with socio-economic attributes capable of predicting urban mobility flow with high predictive power?

8. Network centrality and urban flow prediction

Can the above-mentioned centrality measures enhanced with urban socio-economic attributes be informative for *predicting* urban mobility flows?

This thesis will address the listed research questions within the overarching objective of exploring the relationships between ***urban structure*** and ***urban mobility*** by proposing specific methods and techniques drawing heavily from previous work in network theory, probability theory, machine learning, and neural networks.

"The year 2000 was essentially the point at which it became cheaper to collect information than to understand it."

Freeman Dyson

Chapter 4

Data

4.1 What are OD flow data?

As the name suggests, origin-destination (OD) data, also known as flow data, represent transportation flows through geographic space, from an origin (O) to a destination (D). OD datasets represent information on trips between two geographic areas or, more commonly, zones, often represented by the geographical centroids of the areas in question. Typically encoded with a square symmetric matrix, OD flow data contain numerical data on the aggregate quantity of vehicles or individuals travelling from one geographic area to another [178] over a specific time period. Mostly used in transportation planning, OD flows are an invaluable source of data for understanding spatial and temporal patterns of urban mobility and dynamics [254].

As discussed in Chapters 1, 2, and 3, where we outlined the methodology and main objectives of this thesis, we aim to model the relationship between urban mobility and urban socio-economic characteristics by seeing the OD matrix as a network graph of OD flows between city locations, augmenting it with quantitative socio-economic attributes describing these city locations and the various kinds of relations between them.

In this Chapter, we outline and describe the construction process of the node- and edge- attributed urban mobility OD flow network datasets in Rome and London which we will use throughout this work.

These flows can be modelled as attributed graphs with both node and edge

attributes characterising locations in a city and the various types of relationships between them.

To make our work reproducible and to contribute to the scientific communitiy, we publicly release[1] a custom dataset of aggregate origin-destination (OD) flows of private cars in London augmented with attribute data describing city locations and dyadic relations between them.

The principal data source for building such OD networks consists of GPS trajectories of private car in Rome and in London, provided within the scope of the EU Horizon 2020 "Track & Know" programme. Other open data sources have been used to construct or augment information intrinsic to the nodes and edges of the obtained OD networks. Such data sources include OpenStreetMap [198], Airbnb [185], London transport data [260], and London housing density [92].

In what follows, we first describe the detailed workflow of the attributed OD network dataset construction, and then provide an exploratory data analysis of some of the most important node and/or edge attributes of the OD networks.

4.2 Building the data set

The workflow of building the data sets for both Rome and London is as follows:

1. The urban territory has been subdivided into n Cartesian grid cells of different resolutions (250×250 m, 500×500 m, 1000×1000 m, 1500×1500 m, and 2000×2000 m), and each such quadratic cell is considered a node in the graph of the given resolution.

2. The raw GPS trajectories of around 10000 private cars spanning two years have been obtained from proprietary car insurance data for research purposes within the EU H2020 "Track & Know" programme. The data have been cleaned, processed, and superimposed on the grid. Then, trip origin and destination GPS positions have been identified by interpreting the engine ignition on/off interval for each vehicle. Time intervals between 10 and 35 minutes showed

[1] Dataset will be released at https://trackandknowproject.eu/file-repository/.

robust outcomes, and 20 minutes were chosen for identifying car trips. The extracted origin-destination points were then mapped to the respective grid cells.

3. The *OD* networks have been built from the extracted origin-destination pairs described by the weighted adjacency matrix, $\mathbf{W}^A \in \mathbb{R}^{n \times n}$, the element $\mathbf{W}^A_{ij}$ of which represents the number of car trips starting at node (cell) i and ending at node j.

4. The node features have been built by engineering 35 features from various open sources [198, 185, 260] and from the GPS data. These features include population density, average Airbnb prices, parking areas, areas covered by residential buildings, number of restaurants, bars, banks, museums, road network density, average radius of gyration, etc. per cell. Examples of node features and their spatial distribution are visualised in Figures 4-5 and 4-6.

5. Similarly, the edge features encode information on 12 dyadic relations such as network distance, average time, average speed, temporal correlation between car incidence in cells, public transport connections, etc.

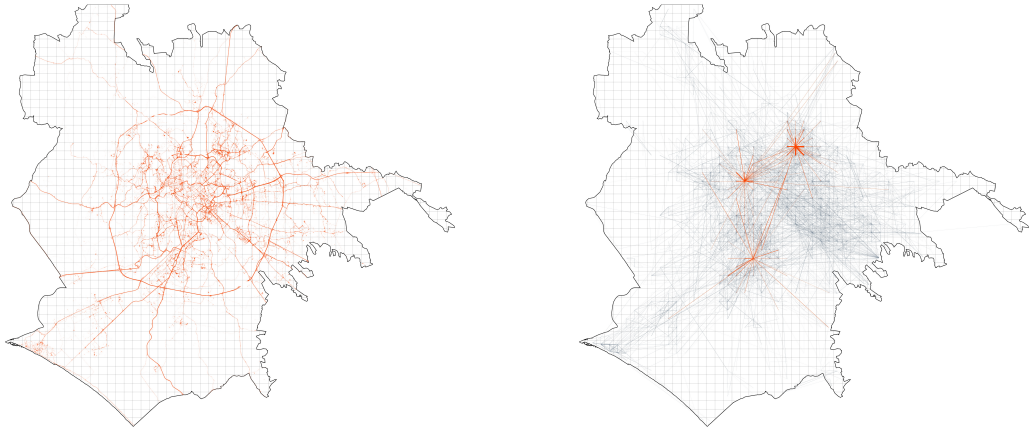
The reason we construct OD networks for different grid resolutions as stated above, is for testing the results obtained in this study concerning both aforementioned thesis objectives for robustness to the spatial scale of the grid.

Next, the above-described procedure for obtaining OD networks is repeated on hourly and daily temporal scales to obtain temporal OD networks for studying their behaviour across different hours of the day and different days of the week.

4.3 Exploratory Data Analysis

4.3.1 From individual mobility to OD networks

The concept of urban mobility is typically understood as a derivative of individual human mobility [24]. In this respect, building urban OD flow networks requires spatio-temporal aggregation of individual human mobility patterns. This necessarily



(a) (b)

Figure 4-1: (a) Car GPS trajectories over 1×1 km grid cells in Rome. (b) Origin-Destination (OD) flow network in Rome with some popular travel locations highlighted.

implies loss of granularity and information about how individual mobility operates in an urban environment.

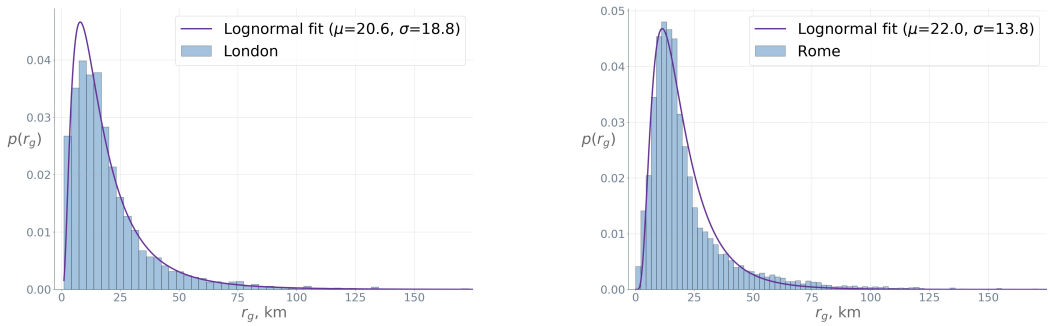
Within our representation framework of node- and edge- attributed urban mobility networks, we incorporate information on individual mobility as a node attribute in the OD network. In particular, we focus on a widely used measure describing the characteristic distance travelled by an individual [207]:

$$r_g = \sqrt{\frac{1}{N} \sum_{i \in L} n_i \left(\mathbf{r}_i - \mathbf{r}_{cm}\right)^2}, \tag{4.1}$$

where L is the set containing the locations visited by the individual, $\mathbf{r}_i$ is a two-dimensional vector describing location i in geographical space; n_i is the frequency with which location i is visited, $N = \sum_{i \in L} n_i$ is the total number of visits made by the individual, and $\mathbf{r}_{cm}$ is the center of mass of the individual, defined as the mean weighted point of all locations visited by the individual.

In Figures 4-2a and 4-2b, we see the empirical distributions of the individual radii of gyration in London and Rome fitted to Lognormal distributions with mean 20.6 and 22.0 km, respectively.

These individual radii are then aggregated into average values per cell and a cell attribute $\overline{r_g}$ is thus presented as a node-specific attribute in the OD flow network. The empirical distributions for this value in London and Rome are presented in



(a) (b)

Figure 4-2: Empirical distributions of the average radius of gyration per cell in (a) London (b) Rome

Figure 4-3, where you can see the London values tightly packed around the mean value of around 23 km, the Rome values being more spread out around a much higher mean value of around 34km.

4.3.2 Urban socio-economic attributes

As mentioned in section 4.2, the spatial variables describing specific locations in a city can be encoded in a vector of attributes intrinsic to each node in the urban OD network. We construct a total of 36 node attributes Examples of such node attributes are presented in Figures 4-5 and 4-6.

A particular node (cell) attribute of interest is the average betweenness centrality of the street junctions contained in each grid cell (Figure 4-4). The arterial hub and spike structure of the London street network can be clearly distinguished in the aggregated 500×500m grid resolution.

For a complete description and summary of node and edge attributes of the urban mobility OD network, we refer the reader to Appendix A.

4.3.3 OD network flows

Urban economics has consistently showed the emergence of spillover effects and agglomeration economies in cities, resulting in rich-get-richer effects through preferential attachment mechanisms [200]. A similar phenomenon is also observed in the

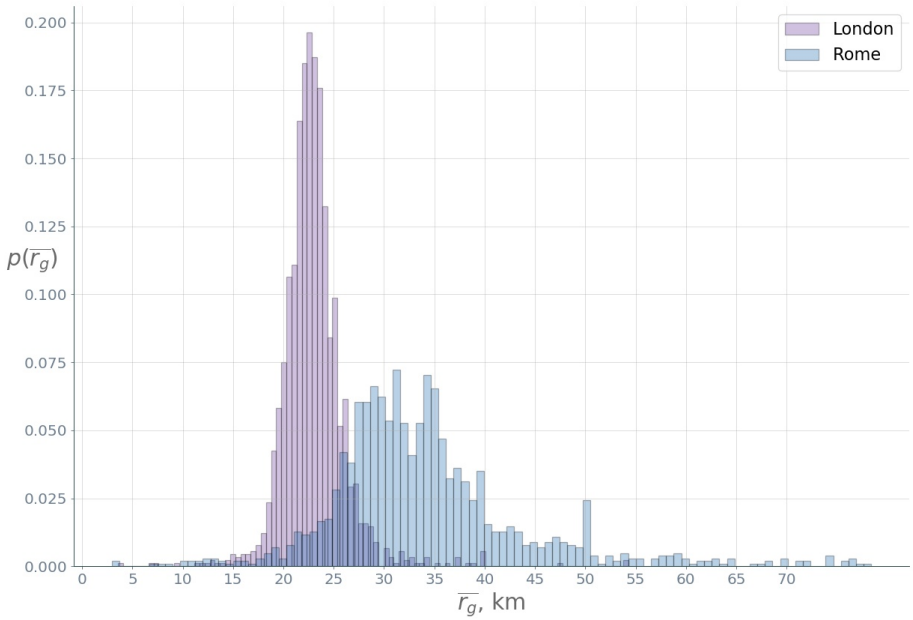


Figure 4-3: Empirical distributions of the average radius of gyration per cell in London and Rome.

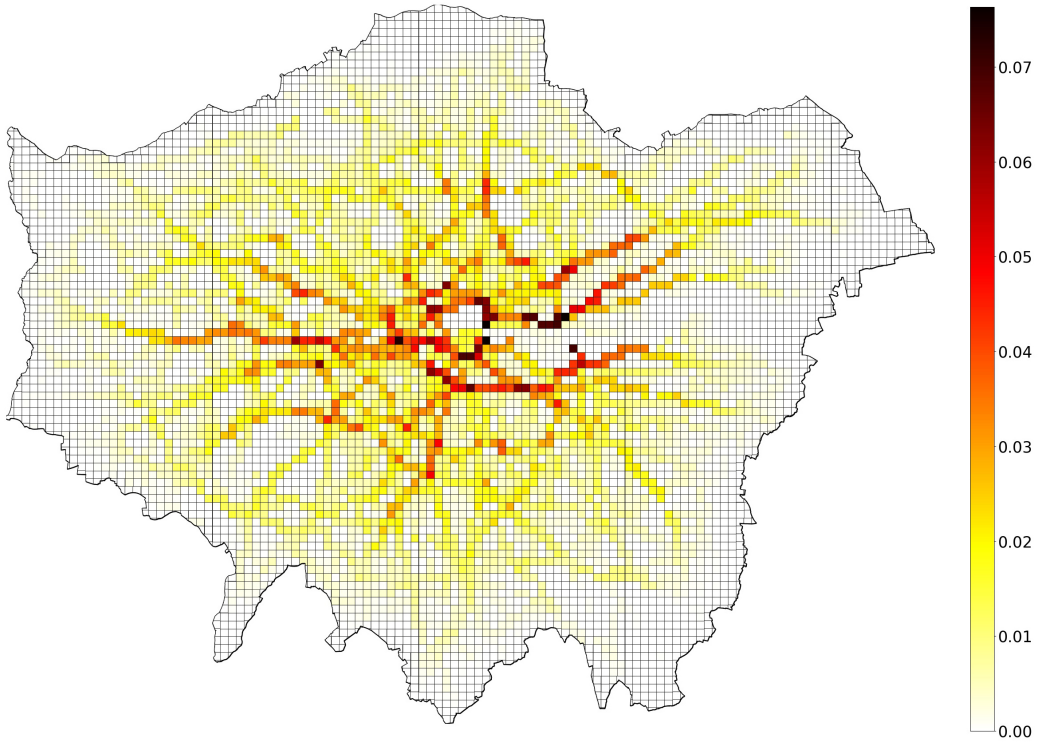
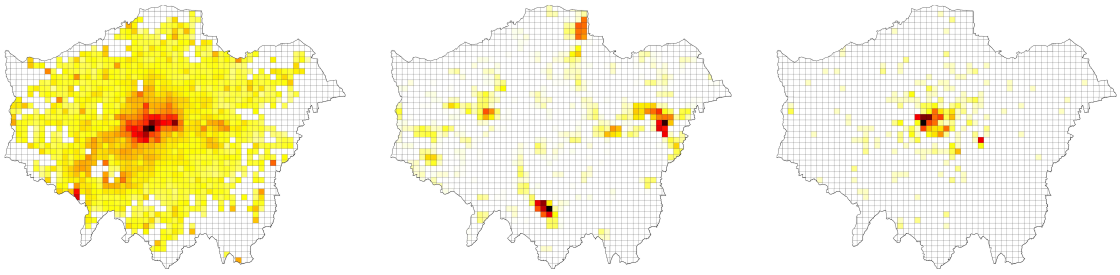
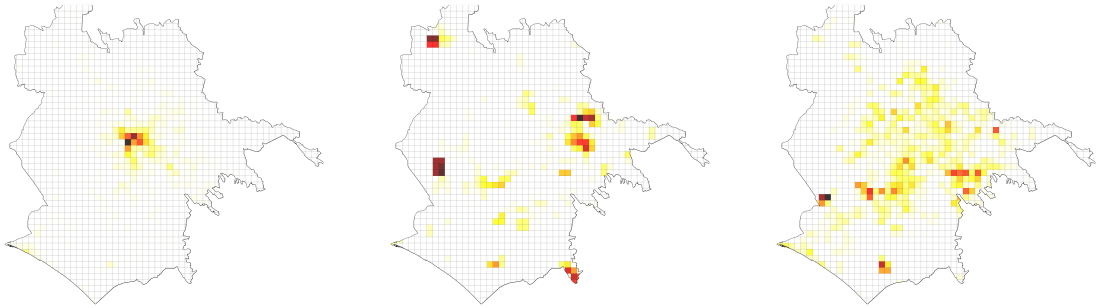


Figure 4-4: Average street junction betweenness centrality in each 500 × 500m grid cell in London.



(a) (b) (c)

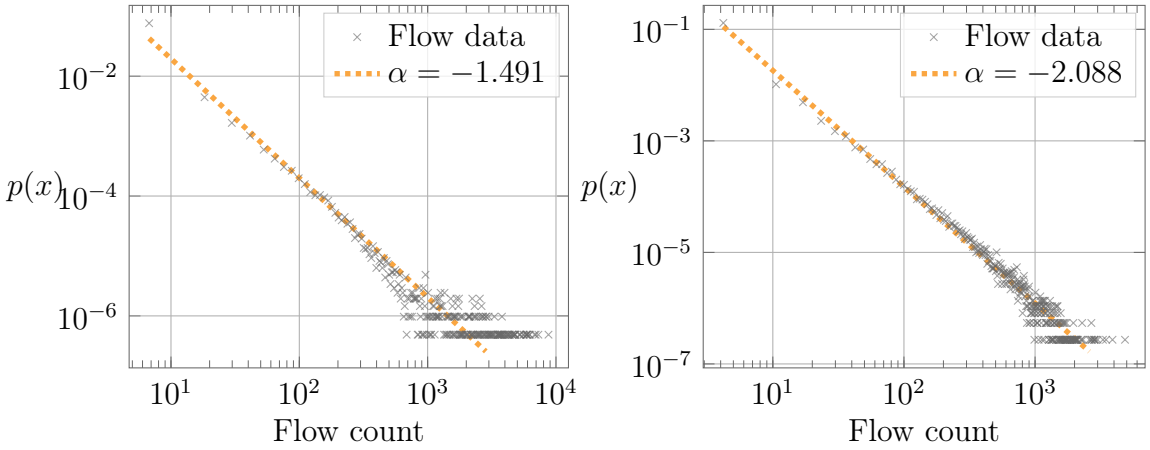
Figure 4-5: Examples of node (cell) features in London (a) Average Airbnb listing prices (b) Proportion of grid cell area allotted to industrial activity (c) Number of museums and galleries per grid cell. Darker colours indicate higher values.



(a) (b) (c)

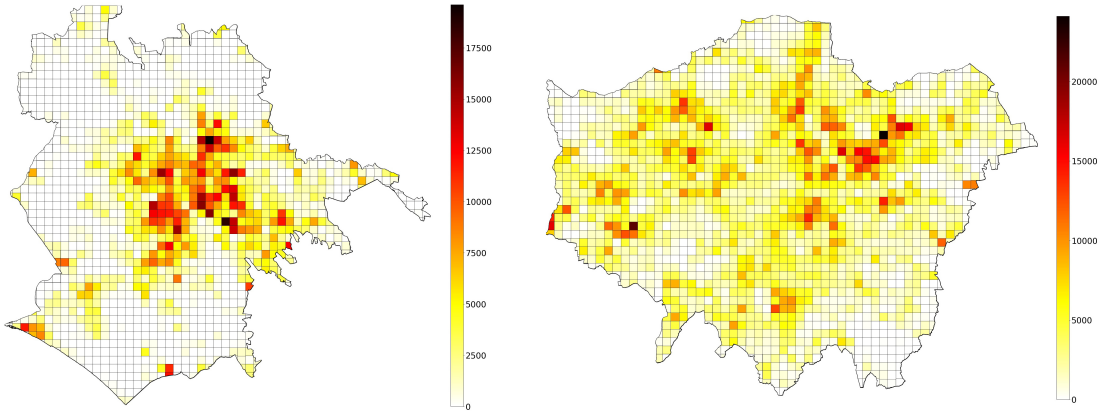
Figure 4-6: Examples of node (cell) features in Rome (a) Number of restaurants (b) Proportion of grid cell area allotted to industrial activity (c) Cell area alotted to parking. Darker colours indicate higher values.

degree and flow distributions in urban OD networks, where these distributions seem to exhibit power law properties [229].



(a) (b)

Figure 4-7: Log-log plots of the probability distributions of the OD flows fitted with a power-law distribution $p(x) \propto x^{-\alpha}$ with exponents of (a) $\alpha = -1.491$ in Rome. (b) $\alpha = -2.088$ in London.



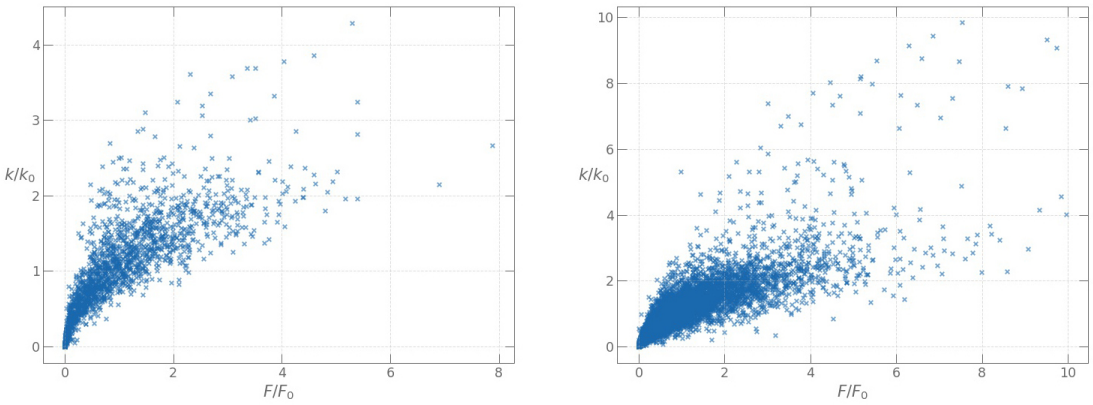
(a) (b)

Figure 4-8: Total mobility in-flows in (a) Rome (b) London

Indeed, statistical fitting to our data with Kolmogorov-Smirnov tests shows that the OD flow distributions in Rome and London display power law behaviour with exponents -1.491 and -2.088, respectively (Figure 4-7).

The lower power law exponent signifying a heavier right tail hints at a more unequal concentration of flows in Rome. Indeed, a visual inspection of the spatial distribution of total in-flows in the grid cells in both cities suggests a monocentric

concentration of flows in Rome, while London flows display a clear polycentric spatial structure (Figure 4-8). We will take a closer look and elaborate on this observation of the differences in the spatial organisation of mobility flows in Rome and London in Chapter 8.



(a) (b)

Figure 4-9: Correlation between node degree and node total in-flow in the London OD flow network of grid resolution (a) 1000×1000 m (b) 500×500 m

Enquiring into the possible mechanisms behind the generation of the power laws in urban OD networks observed above is a challenging research question in itself. As a first caveat, we should be aware of the so-called Modifiable Aerial Unit Problem [197] - a statistical bias in which a spatial unit-based variable describing a spatial phenomenon is influenced by both the shape and scale of the aggregation unit. In our case, the different grid resolutions - 250×250 m, 500×500 m, 1000×1000 m, 1500×1500 m, and 2000×2000 m - can have an impact on the empirical distributions of OD network flows. To illustrate this, we plot the node degrees, k, against the total node in-flows, F, normalised by the average degree k_0 and the average total in-flow F_0, respectively (Figure 4-9). As one might have expected, we see a positive relationship between node degree and total in-flow. However, this relationship is by no means trivial and differs significantly across different grid resolutions.

This calls for a closer study on the relationship between spatial scale and network properties of OD flows. However, it is beyond the scope of the present thesis. As such, it will be formulated as future work.

"The origins of graph theory are humble, even frivolous."

N. Biggs, E. K. Lloyd, and R. J. Wilson

Chapter 5

PageRank & Eigenvector Centrality

In this Chapter, we study and compare several measures of centrality specifically applied to urban networks. We show how these centralities are based on the calculation of the eigenvectors of a matrix and are very suitable for attributed urban networks described in Chapter 4. With the aim of expanding the range covered by these measures, we present a new centrality measure based on the Google PageRank algorithm. We then compare the performance of three discussed centrality measures by applying them to the attributed urban OD network in Rome.

This Chapter is a modified version of our paper Manuel Curado, Leandro Tortosa, Jose F Vicent, and Gevorg Yeghikyan. Analysis and comparison of centrality measures applied to urban networks with data. *Journal of Computational Science*, page 101127, 2020

5.1 Introduction

5.1.1 Motivation

The number and diversity of relationships that occur between space, information and social processes endow the city with characteristics of a complex system. One way of dealing with the complexity of the city is through networks, since they capture the relations (edges) between objects (nodes) [22].

How to effectively identify influential nodes (or edges) in urban networks is a

question that has been paid attention to because of the great impact on the lives of millions of people. The idea behind the importance of a node in urban networks is related with the mathematical concept of centrality. There is an extensive bibliography regarding the design and implementation of centrality measures in complex networks but a high percentage of them are based on the topology of the network as the main element.

It is a fact that the city is a complex network where a great deal of information are generated. It can be said that the sources of this data are diverse as for example social networks existing databases.

On the one hand, complex networks have emerged as a model to understand, analyse and visualize characteristics of complex systems, such as cities. On the other hand, the city is a source of data, both physical and virtual, which constitute an essential part of it and must not be omitted.

Hence, it is necessary to study measures of centrality that consider both characteristics, the connectivity of the nodes (topology) and the information associated with them (data). Because of that, the comparison of existing measures of centrality in urban networks that take into account the influence of both factors, topology and data, is essential. This significant fact constitutes the main motivation that has led us to develop this Chapter.

5.1.2 Literature Review

Nowadays, centrality measures have become an essential tool in network analysis, and are extensively used for classifying the influence in such networks as: social networks [102], Internet web-page popularity [203], computer networks [59], spread of epidemic diseases [208], ranking reputation of scientists [305], urban networks [75, 74], etc.

In the literature, there exist studies that focus on defining centrality measures based on graphs. One of the most used centrality measures is the degree centrality [49], in which the most important nodes are those with higher degrees in the network graph. Other centrality measures, widely used, are: eigenvector centrality [49], Katz centrality [141], closeness centrality [54], PageRank [36, 121, 139, 153, 203,

210], betweenness centrality [17, 20, 106, 188], percolation centrality [214], Freeman centralities [102], and others centralities based on topology [222]. For a deeper study, in [35] a comprehensive review of centrality measures is presented.

In spite of the existence of many investigations on the concept of centrality, the presence of data in the study of centralities in networks is very recent [7, 8, 9, 11]. In [7], the authors propose a new centrality measure (called Adapted PageRank Algorithm -APA-) which the main contribution of which is to establish a node classification taking into account the topology and the data associated with the nodes. This algorithm, based on the PageRank concept, considers the connectivity between nodes and the data used for each specific problem.

Important centrality measures are those obtained as a solution of the eigenvalue problem [42]. In these cases, the classifications of the nodes are given by the values of the dominant eigenvector. Important examples are the eigenvector centrality [49] or PageRank [203, 139, 155]. Over the years, some modifications of the PageRank model have been proposed. For instance, in [121] the authors develop precise search results calculating a set of PageRank vectors. A link-based algorithm built on a random surfer model reflecting back steps is presented in [252]. In [42], the authors show the fundamental properties related with: the complexity of the computational scheme of the PageRank, the stability of the algorithm and the role of parameters in the computation of the PageRank. To compute the PageRank an iterative method, named *the Power method* is used. The objective of this method is to converge to the principal eigenvector of the Markov chain representing the Web graph. In [180], the authors show an algorithm based on the *Power method* that accelerates the convergence.

5.1.3 Main contribution

In a city, the importance of a place depends not only on the topology of the network but also, among other factors, on geolocated information. Consequently, the algorithms studied in this Chapter incorporate, in the calculation process, the geolocated data. This variable is reflected, in the different algorithms compared, using a data matrix that is used for the construction of the primitive transition matrix.

From this point of view, the resulting transition matrix incorporates both the jump probabilities derived from the topology of the network and those derived from the data assigned to the nodes

Two main characteristics of the compared measures of centrality can be highlighted: they are based on the calculation of the eigenvector of a matrix and they are suitable for urban networks with data. These centralities are closely related to each other because they all depend, largely, on the degree of the nodes (topology) and the information assigned to the nodes. Therefore, a comparative analysis that determines the relationships, differences and the meaning of all of them is required. Before this comparison, a new measurement is introduced, similar to APAM1 but which introduces a greater range for a better visualization of the measurement.

Therefore, the objective is twofold, on the one hand introduce a new measure to increase the range and on the other hand, analysis and study of the centrality measures applied to urban networks that are based on the computation of the eigenvector.

The remainder of the Chapter is organized as follows: Section 2 outlines the basic characteristics of the algorithms based on eigenvector. Section 3 shows the methodology used in the comparison. In section 4 some numerical results, based in different networks and data, are illustrated. Finally, a simple conclusions are presented in Section 5.

5.2 Algorithms based on eigenvector.

The following subsections show a reminder of the centrality measures based on the calculation of the eigenvector: Adapted PageRank Algorithm (APA), Adapted PageRank Algorithm Modified (APAM1) and Eigenvector Centrality Modified (CVP). Moreover, a new measure, called APAM2, is presented as a result of a modification of the existing APAM1.

5.2.1 The Adapted PageRank Algorithm, APA

The *PageRank* algorithm [203] was proposed to calculate a classification for each Web page, based on the Web link graph, regardless of their content It was founded, solely, on their location in the Web's graph structure.

The method aims at obtaining a *PageRank vector*, which gives the importance of the web pages. An important characteristic is that the *PageRank vector* does not depend on the request of the person performing the search. In [210], a detailed description of the PageRank algorithm is shown.

A modification of the PageRank model with the aim of establishing a classification of nodes taking into account the information present in the network, is proposed in [7]. This algorithm is called by the authors as *Adapted Pagerank Algorithm* (APA algorithm). A remarkable characteristic is that it can be applied to different types of networks by assigning additional numerical information to the nodes on the network.

In the APA algorithm the construction of a data matrix D is crucial because it summarizes the numerical value of the data assigned to the nodes. This matrix represents, numerically, the analysed information placed in columns. Each column represents a specific type of information that is evaluated or analysed.

Furthermore, the weight vector $\mathbf{v}_0$ establishes the importance assigned to any of

the type of data measured in the network.

Input: Let $G = (V, E)$ be a primary graph representing a network with n nodes.

Output: $\mathbf{x}$ representing the network centrality

begin

Obtain the transition matrix P

Obtain the data matrix D considering different characteristics associated with the nodes

Select the vector $\mathbf{v_0}$, depending on the importance of the characteristics studied

Obtain the vector $\mathbf{v} = D \cdot \mathbf{v_0}$

Normalize $\mathbf{v}$ vector, $\mathbf{v} \longrightarrow \mathbf{v}^*$

Construct V matrix as $V = \mathbf{v}^* e^T$

Obtain M_{APA} matrix as $M_{APA} = (1 - \alpha)P + \alpha V$

The eigenvector $\mathbf{x}$ associated with the dominant eigenvalue $\lambda_1 = 1$ of M_{APA} is the ranking

end

Algorithm 1: APA Algorithm

The numerical classification obtained by the APA algorithm has important characteristics. With this objective, a reformulation of the meaning of some of the matrices in probabilistic terms is necessary:

- On the one hand, the probability of moving from a node to any of its neighbouring nodes is represented by the P transition matrix.
- On the other hand, V matrix can be interpreted as the probability of moving from one node to any other, considering the quantity of data assigned to each node.
- It may be said that in M_{APA} matrix the part αV is related to the probability of moving from a node to any other without there being a link between them. The jump is not influenced by the topological distance between them but the

data present in the network.

5.2.2 Adapted PageRank Algorithm Modified, APAM1

Urban networks have characteristics that make them very particular when applying traditional algorithms of classical network theory. For instance, the equiprobability of moving from one node to another node is at least debatable. The influence that neighbouring nodes exercise on the node itself must be taken into account. These considerations lead the authors to propose a new centrality measure considering the particular characteristics of the urban networks [9].

Input: Let $G = (V(G), E(G))$ be a graph of an urban network.

Output: $\mathbf{x}$ representing the network ranking

begin

- Obtain the adjacency and the transition matrices A and P
- Construct the data matrix D considering different characteristics associated with the nodes
- Select the vector $\mathbf{v_0}$, depending on the importance of features studied
- Obtain the $\mathbf{v}$ vector $\mathbf{v} = D \cdot \mathbf{v_0}$
- Normalize the vector $\mathbf{v}$ using the standard method, $\mathbf{v} \longrightarrow \mathbf{v}^*$
- Compute the V matrix $V = \mathbf{v}^* e^T$
- Obtain the matrix K according to $K = A \otimes V + \epsilon F$
- Normalize the matrix K transforming it in a stochastic matrix, $K \longrightarrow K_N = K^*$
- Construct the matrix $M_{APAM1} = (1 - \alpha)P + \alpha K^*$
- The eigenvector $\mathbf{x}$ associated with the dominant eigenvalue $\lambda = 1$ of M_{APAM1} is the ranking

end

Algorithm 2: APAM1 algorithm

The jump probabilities matrix K^* is stochastic by columns and it constitutes the main difference between the APA and APAM1 algorithm. Because of this matrix, APAM1 (Algorithm 2) is influenced by the data, the topology of the network and

the topological distances between the nodes that contain data.

5.2.3 A new Adapted PageRank Algorithm Modified, APAM2

The APAM1 centrality represents an alternative centrality to that of the APA algorithm but it covers a very small range of values when working with networks and large volumes of data. We describe a modification of the APAM1 algorithm in order to extend the range of the resulting values.

The way to increase the resulting values is related to the normalization of the data vector $\mathbf{v}$. In the construction of the V matrix of APAM1 algorithm, the normalized vector $\mathbf{v}^*$ is used.

$$V = \mathbf{v}^*\mathbf{e}^T = \frac{\mathbf{v}}{\mathbf{v}}\mathbf{e}^T. \tag{5.1}$$

In the construction of the K matrix, the term $A \otimes V$ represents the probability of jumping from a node to its adjacent nodes based on the quantities of data of the adjacent nodes.

The algebraic operation of dividing by the norm of the vector $\mathbf{v}$ implies a very important meaning regarding the data that appears in the V matrix. The normalization of the vector $\mathbf{v}$, before constructing the V matrix, implies that the quantity of data of V represents the portion of data with respect to the totality of the data assigned to the network nodes. This is fundamental, since with the normalization of the K matrix, its terms are very small values in networks with a large volume of data. As a consequence there are a very low range values of the ranking.

Thus, to increase the range implies improving the values of the term $A \otimes V$, and this requires to construct the vector $\mathbf{v}$ without normalizing. That is to say, we build the matrix V as

$$V = \mathbf{v}\mathbf{e}^T. \tag{5.2}$$

In this way, the term $A \otimes V$ of the matrix K represents the data of the adjacent nodes without normalizing. When we make K stochastic by columns, we will obtain the matrix K^* representing the jump probability directly proportional to the data

quantities of the adjacent nodes and not to the proportions of normalized data, as was the case with the APAM1 algorithm. As a consequence of this new approach, the resulting ranking covers a greater numerical range.

The main advantage of extending the numerical range is reflected in the clarity of the visualization and greater numerical stability.

Input: Let $G = (V(G), E(G))$ be a graph of an urban network.
Output: $\mathbf{x}$ representing the network ranking
begin
- Obtain the adjacency and the transition matrices A and P
- Compute the data matrix D considering different characteristics associated with the nodes
- Select the vector $\mathbf{v_0}$, depending on the importance of features studied
- Obtain a vector $\mathbf{v} = D \cdot \mathbf{v_0}$
- Construct the matrix $V = \mathbf{v}e^T$
- Compute K matrix according to $K = A \otimes V + \epsilon F$
- Normalize the matrix K transforming it in a stochastic matrix, $K \longrightarrow K_N = K^*$
- Obtain the matrix $M_{APAM2} = (1 - \alpha)P + \alpha K^*$
- The eigenvector $\mathbf{x}$ associated with the dominant eigenvalue $\lambda = 1$ for the matrix M_{APAM2} is the ranking

end

Algorithm 3: APAM2 algorithm

The interpretation of this result is similar to that of the APAM1 algorithm. The matrix M_{APAM2} has two different terms: the term related with the topology of the network $(1-\alpha)P$ and the term related to the data αK^*. In addition, the α parameter allows to determine the importance associated to both parts. Taken $\alpha = 0.5$ equally importance of both aspects is considered. Hence, we can modulate the importance of the quantity of data present in each node and its nearest neighbours by means of the α parameter.

5.2.4 Eigenvector Centrality applied to urban networks, CVP

The direct application of measures of centrality based on the eigenvector in urban networks does not provide relevant information about the characteristics related to the importance of the nodes in the network.

There are factors from the urban context that significantly affect the importance of spaces in the city. Just think of the commercial streets, places with built heritage or other characteristic spaces of the city that constitute their identity. Because of this, in [11], the authors consider incorporating data in the process of calculating the eigenvector centrality for its application in urban networks. The new proposed measure, called eigenvector centrality modified (CVP), is based on the fact that the topological distribution of the data determines the importance of the network areas. The most important places are those that have greater quantities of data or that are connected to other places with considerable amounts of them.

As the other analysed measures, the algorithm starts from a data matrix D and a weighting vector $\mathbf{v}_0 \in R^{n\times 1}$. The normalized resulting vector $\mathbf{v}^*$ is used to calculate the matrix $W = (w_{ij}) \in R^{n\times n}$ with

$$w_{ij} = \begin{cases} v_i^* + v_j^* & \text{if } i \text{ has a link with } j \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{5.3}$$

This matrix is symmetric and the element w_{ij} represents the quantity of data associated with the edge between the nodes i and j. With this approach the importance of a node depends both on the data itself and on the data of adjacent nodes.

In a city there are areas without data. This implies that some entries of the W matrix are zeros and, consequently, there is loss of the topological information of the network. To solve this problem, the authors introduce a parameter β as a minimum amount of data associated with the edges.

$$\beta = \min(w_{ij}), \quad \text{for } w_{ij} > 0. \tag{5.4}$$

The following step is the construction of matrix $M_{CVP} \in R^{n\times n}$

$$M_{CVP} = A \otimes (W + \beta J) + \epsilon J, \tag{5.5}$$

where $J \in R^{n\times n}$ is a matrix of ones and $\otimes$ is the Hadamard product. The term ϵJ serves to avoid localized solutions and represents a small portion of the parameter

β. An appropriate value of ϵ, which is

$$\epsilon < \frac{1}{10}\beta. \tag{5.6}$$

Finally the dominant eigenvalue λ_1 and its corresponding vector $\mathbf{x}_1$ are calculated. Then the centrality $\mathbf{c}$ is calculated as

$$\mathbf{c} = \frac{1}{\lambda_1}\left(A\mathbf{x}_1 + \mathbf{x}_1\right). \tag{5.7}$$

The term $\lambda_1^{-1}A\mathbf{x}_1$ comes from the centrality proposed by Bonacich and represents the centrality of the adjacent nodes. With this, the local centrality of the node and the centrality of neighbouring nodes are considered.

The calculation process of the CVP centrality is summarized in the following algorithm:

Input: Let $G = (V(G), E(G))$ be a graph of an urban network.
Output: $\mathbf{c}$ representing the network ranking
begin
Obtain the data matrix D
Select the weighting vector $\mathbf{v}_0$
Calculate the data vector $\mathbf{v} = D\mathbf{v}_0$
Normalize the data vector $\mathbf{v}^* = \max(v_i)^{-1}\mathbf{v}$
Compute the matrix W
Calculate $M_{CVP} = A \circ (W + \beta A) + \epsilon J$, where $\beta = \min(w_{ij}), w_{ij} > 0$, and $\epsilon < \frac{1}{10}\beta$
Calculate the vector λ_1 and the eigenvector $\mathbf{x}_1$ de M_{AVP}
The centrality measure CVP is

$$\mathbf{c} = \frac{1}{\lambda_1}\left(A\mathbf{x}_1 + \mathbf{x}_1\right).$$

end

Algorithm 4: CVP Centrality

5.3 The methodology of the comparison

The centrality measures adapted to urban networks previously exposed, are based on the calculation of the dominant eigenvector. These measures require a comparative analysis to determine the relationships and differences between them and to better understand the meaning and applications of each of them.

The APA algorithms are the adaptation of the PageRank model for its application in the urban context. In these algorithms, the nodes classification does not depend exclusively on the urban network topology but it also depends (among other factors) on the geolocated data. Consequently, the APA models incorporate the geolocalized data in the calculation process by means of a data matrix. Taking into account this perspective, the resulting transition matrix incorporates both the jump probabilities derived from the topology of the network and those probabilities derived from the amounts of data available at the nodes.

On the other hand, the CVP centrality constitutes the adapted version of the Bonacich's measure known as eigenvector centrality. It represents the importance of a node according to the importance of adjacent neighbouring nodes. The centrality defined in this way depends on two factors: the number of adjacent nodes and the centrality of each of the adjacent nodes. The main contribution of the CVP centrality is the incorporation of data from the urban context in the process of computing centrality. In this way, the importance of a node in the network depends not only on the degree and the centralities of the adjacent nodes, but also on the data itself and the data associated with the neighbouring nodes.

It can be affirmed that the final result of the APA centralities and the CVP centrality is directly related to the calculation of the dominant eigenvector. However, the meaning of the components of this vector is different for each case. These numerical differences in the final results are determined by the characteristics of the matrix used to calculate the eigenvector.

To carry out the comparison of the APA centralities based on the eigenvector, we have chosen the most representative centralities of this group: APA and APAM2. This choice is due to the fact that the APAM1 and APAM2 centralities are very

similar.

The expressions for APAM1 and APAM2 models are

$$M_{APAM1} = (1 - \alpha)\, P + \alpha V, \tag{5.8}$$

$$M_{APAM2} = (1 - \alpha)\, P + \alpha K^*. \tag{5.9}$$

As we can show in expressions 5.8 and 5.9, the matrix that represents the topology in both models (matrix P) is the same. By contrast, the main difference between both algorithms is the matrix that represents the data (V and K^* respectively). The V matrix represents the jump probabilities from a source node to any other destination node. This probability depends only on the amount of data associated with each node. On the other hand, the K^* matrix of the APAM2 model also represents the jump probabilities between the nodes based on the amounts of data associated with them. However, this matrix only considers the probabilities of jumping from a node to the adjacent nodes. This matrix is more consistent with the idea of pedestrian traffic in urban contexts.

The APA and APAM2 centralities are obtained through the eigenvector associated with the dominant eigenvalue $\lambda = 1$. This eigenvector can be interpreted, in probabilistic terms, as the stationary vector of a Markov chain that has as transition matrices M_{APA} and M_{APAM2}, respectively. Mathematically, we can express the stationary vector as

$$\mathbf{x}_{APA} = \lim_{n\to\infty} M_{APA}^n \mathbf{x}_0 \tag{5.10}$$

and

$$\mathbf{x}_{APAM2} = \lim_{n\to\infty} M_{APAM2}^n \mathbf{x}_0, \tag{5.11}$$

where $\mathbf{x}_0$ is the initial vector.

The stationary vector resulting from the APA and APAM2 centralities matches with the eigenvector associated with the dominant eigenvalue $\lambda_1 = 1$ of the transition characteristics matrices of the two centralities. The stationary vector or the dominant eigenvector has a different interpretation in the urban context. Vector

$\mathbf{x}_{APA}$ represents the probabilities of locating a person in a certain place (node of the network), taking into account the topological configuration and the amount of data available in this place. We consider both the continuous and discontinuous displacements. By continuous displacements, a sequential transition between the nodes is understood.

Following a similar scheme to the case of matrices M_{APA} and M_{APAM2}, theM_{CVP} matrix includes the topological variable and the data variable of the network. This matrix is not a stochastic transition matrix as the APA type. It is given by the expression

$$M_{CVP} = A \otimes (W + \beta J) + \epsilon J. \tag{5.12}$$

However, for a better interpretation of the M_{CVP} matrix, we need to rewrite it as

$$M_{CVP} = W + \beta A + \epsilon J. \tag{5.13}$$

In 5.13 the W matrix is the adjacency matrix with weights. It represents the amounts of data associated with pairs of nodes joined by an edge. The βA matrix ensures a minimum link between the nodes when there is no data associated with them. Matrix ϵJ avoids localized solutions and speeds up the calculation process. We can say that the matrix W is the most important component of the expression 5.13, since it combines topological and data variable of the network and it represents the topological distribution of data in the network.

The addition of W, βA and ϵJ matrices gives a non-negative symmetric matrix. This property ensures that all eigenvalues associated with the resulting matrix are real.

The CVP centrality is also related to the dominant eigenvector $vecx_1$ associated with its characteristic matrix M_{CVP}. That is to say,

$$\mathbf{c}_{CVP} = \frac{1}{\lambda_1}(M_{CVP}\mathbf{x}_1 + \mathbf{x}_1). \tag{5.14}$$

However, the meaning of the components of the vector $\mathbf{x}_1$ is different from the APA

type centralities. To better understand the meaning of this centrality we briefly study the Katz centrality, given that CVP is a particular case of this centrality.

The katz centrality can be expressed as

$$\mathbf{c}_{CVP} = \beta A(I - \beta A)^{-1}\mathbf{e} = (\beta A + \beta^2 A^2 + \beta^3 A^3 + \ldots)\mathbf{e} = (\sum_{k=1}^{\infty} \beta^k A^k)\mathbf{e}, \qquad (5.15)$$

where $0 \leq \beta \leq 1$ is the damping factor, $\mathbf{e}$ is a vector of ones and A is the adjacency matrix.

It should be noted that the convergence of the expression 5.15 is guaranteed only if $\beta < 1/\lambda_1$. In contrast, when β tends to $1/\lambda_1$, the Katz's centrality $\mathbf{c}_{CVP}$ tends to the eigenvector of the matrix A. Therefore, the Katz centrality with $\beta \to \infty$ and the eigenvector of a node is the sum of all the links of this node with the remaining nodes of the network. From the expression 5.15 we can see that the indirect links (links with $k > 1$) counts less in the centrality, since the parameter β penalizes links by increasing the values of k .

5.4 Numerical results

In order to analyse the differences in the centrality measures studied, some numerical results are presented. The reasons to calculate the centralities are: on the one hand, to determine the coherence in the results of the classifications provided by the measures and, on the other hand, to determine the magnitude of differences in the results of the centralities. All the numerical tests have been carried out by implementing the different algorithms (APA - 1, APAM2 - 3 and CVP - 4) in R, a Free Software under the terms of GNU project. It constitutes a language and environment especially efficient for computing and graphics.

5.4.1 Network and dataset

The three centralities have been calculated independently, for the particular network shown in Figure 5-1.

The network that has been used is based on a partition of a specific city. That

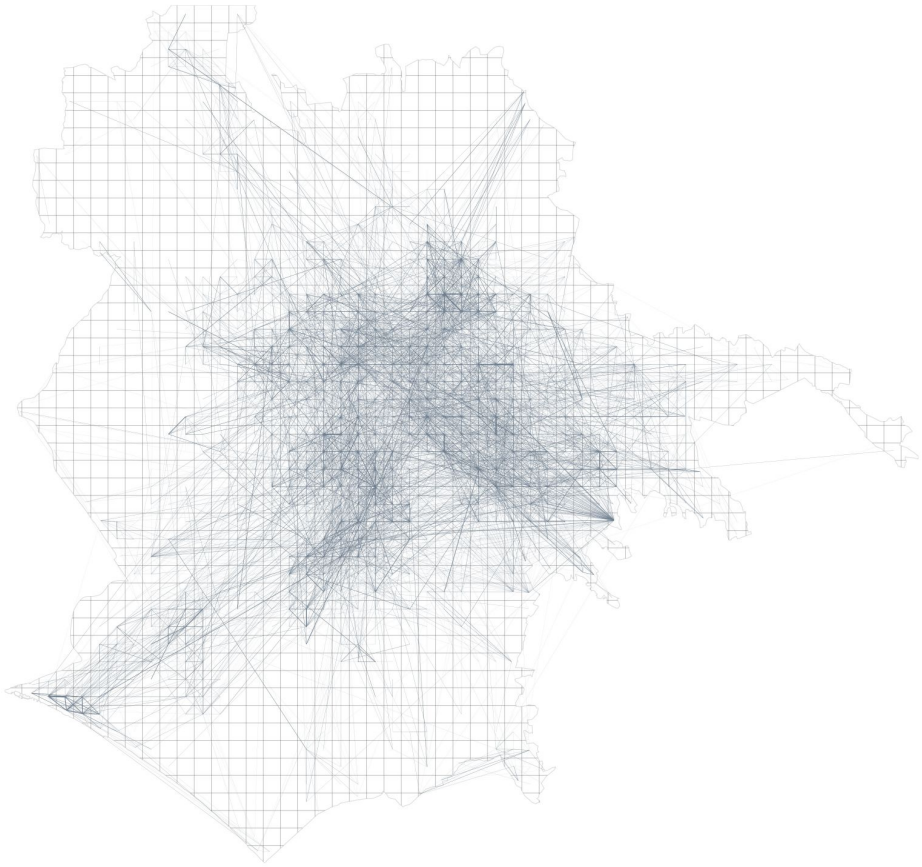


Figure 5-1: The urban network of the city of Rome.

is to say, the city of Rome has been subdivided into a grid with cells of size 1×1 km. Each quadratic cell is a node of the graph.

On the other hand, each day, the raw GPS trajectories of around 10000 cars have been imposed on the grid, and trip origins and destinations have been extracted. If there exists a trip between two cells (nodes) then an edge is constructed between them. With this method, a network of 1359 nodes and 178844 edges has been built from the extracted origin-destination pairs described.

The node attribute matrix (data) has been built using features corresponding to tourist extracted from geo-referenced data from OpenStreetMap. More specifically,

we have used data from monuments, theatres, museums and airports.

Summarizing, the adjacency matrix size is $A = 1359 \times 1359$, the matrix v_0 has size $\mathbf{v_0} = 4 \times 1$ and the data matrix $D = 1359 \times 4$.

Matrix D represents the dataset evaluated in the network. The rows correspond to the nodes while the columns correspond to the nodes' attributes that must be evaluated. In our example there are 4 columns related to the monuments, theatres, museums and airports associated to each node, respectively. We have chosen the weighted vector as $\mathbf{v_0} = [1, 1, 1, 1]^T$, which means that we give the same value to all the attributes within the general quantification process of the data.

Note that the product $D \cdot \mathbf{v}$, from which a vector of data associated with the nodes is obtained, supposes in a certain way the construction of a weighted network, with the characteristic that the weights are assigned to the nodes and not to the edges, since the numerical value incorporated into each node is directly related to the data that exists in its proximity.

With these matrices, we can construct the three algorithms used in the comparison.

5.4.2 Discussion

We have applied three centrality measures (APA, APAM2 and CVP) to the network of Rome using data about tourism.

The numerical study has been developed for different values of the parameter α of the centralities APA and APAM2. This parameter controls the importance that we assign to the calculation of centralities to the topology and data. Remark that the greater the parameter is, the greater the importance for data is assumed.

Table 5.1 shows the value of the APA of fifty nodes in decreasing order when $\alpha = 0.15$. The first column is the node identifier, the second column is the degree on the node, third column represents the amount of data and finally, the rest of the columns summarize the results obtained for different values of the α parameter ($\alpha = 0.15, 0.3, 0.6$ and 0.85). As we can see in this table, the first place in the classification is the node number 1349 that has high degree and large amount of data.

n	dg	data	$\alpha = 0.15$	$\alpha = 0.3$	$\alpha = 0.6$	$\alpha = 0.85$	n	dg	data	$\alpha = 0.15$	$\alpha = 0.3$	$\alpha = 0.6$	$\alpha = 0.85$
1349	82	75	0.01841	0.03467	0.06714	0.09421	682	81	6	0.00314	0.00413	0.00613	0.00781
1348	80	43	0.01122	0.02041	0.03878	0.05411	1331	101	3	0.00291	0.00311	0.00359	0.00408
1341	91	40	0.01096	0.01943	0.03633	0.05043	649	91	4	0.00284	0.00335	0.00441	0.00532
1347	67	41	0.01054	0.01934	0.03693	0.05158	685	94	4	0.00283	0.00337	0.00444	0.00534
1346	60	32	0.00869	0.01549	0.02906	0.04035	629	90	3	0.00282	0.00305	0.00357	0.00408
324	24	32	0.00772	0.01469	0.02861	0.04019	630	82	4	0.00280	0.00337	0.00447	0.00536
1356	81	21	0.00654	0.01085	0.01945	0.02662	222	119	0	0.00272	0.00214	0.00111	0.00038
1355	80	19	0.00606	0.00994	0.01767	0.02411	707	93	4	0.00271	0.00330	0.00444	0.00535
1354	76	18	0.00578	0.00943	0.01673	0.02283	251	106	1	0.00269	0.00241	0.00193	0.00162
1340	95	16	0.00570	0.00884	0.01511	0.02038	1339	96	2	0.00265	0.00267	0.00274	0.00285
1350	82	17	0.00565	0.00908	0.01592	0.02162	1333	60	5	0.00264	0.00349	0.00516	0.00654
1357	76	16	0.00543	0.00865	0.01606	0.02037	305	109	0	0.00261	0.00206	0.00108	0.00038
684	78	14	0.00502	0.00774	0.01322	0.01783	651	89	3	0.00257	0.00289	0.00352	0.00407
1342	95	11	0.00448	0.06499	0.01059	0.01407	286	111	0	0.00254	0.00201	0.00106	0.00037
653	82	12	0.00447	0.00679	0.01142	0.01532	687	97	2	0.00254	0.00257	0.00269	0.00283
1337	83	12	0.00447	0.00678	0.01142	0.01532	654	88	3	0.00253	0.00286	0.00351	0.00407
554	93	10	0.00446	0.00622	0.00981	0.01287	652	83	2	0.00252	0.00259	0.00273	0.00286
1358	76	12	0.00438	0.00672	0.01141	0.01532	1330	78	3	0.00251	0.00284	0.00349	0.00406
647	84	9	0.00405	0.00567	0.00891	0.01162	33	109	0	0.00250	0.00198	0.00103	0.00035
683	79	9	0.00386	0.00549	0.00879	0.01157	1334	59	4	0.00248	0.00312	0.00434	0.00532
686	102	8	0.00386	0.00522	0.00798	0.01034	553	71	2	0.00248	0.00259	0.00272	0.00286
650	89	8	0.00385	0.00524	0.00803	0.01037	1338	77	3	0.00243	0.00277	0.00345	0.00405
1345	74	9	0.00362	0.00533	0.00872	0.01155	648	47	5	0.00241	0.00336	0.00516	0.00656
1332	88	6	0.00337	0.00429	0.00619	0.00782	555	47	4	0.00241	0.00307	0.00433	0.00532
646	85	5	0.00321	0.00393	0.00538	0.00661	123	94	0	0.00240	0.00192	0.00103	0.00036

Table 5.1: Fifty first values of the APA centrality for different α values

In table 5.2 we can see the numerical results for the second measure studied APAM2 and the similarity regarding to APA is clearly seen. The position of the nodes in both measures show a high relationship between the two measures.

Finally table 5.3 summarizes the numerical results in the case of CVP centrality. It is necessary to highlight that there is not parameter α in this measure and, because of this, it always gives the same value of centrality.

Except for minor non-significant variations, no node with a high APA centrality has a low APAM2 centrality and vice-versa. Comparing tables 5.1 and 5.2, it is concluded that the APA and APAM2 values are closely related. The nodes are located in very similar positions in both measures.

Since, the data assigned to the nodes are referred to tourism, we can interpret that nodes without data are nodes not exploited tourism. Consequently, we can see as some nodes without data (222 or 33) appears between 50-top highest values in APA and APAM2. These nodes, whit high degree, have the highest CVP values, being this measure an indicator of potential of what places could be better exploited in tourism.

n	dg	data	$\alpha = 0.15$	$\alpha = 0.3$	$\alpha = 0.6$	$\alpha = 0.85$	n	dg	data	$\alpha = 0.15$	$\alpha = 0.3$	$\alpha = 0.6$	$\alpha = 0.85$
1349	82	75	0.01839	0.03494	0.06835	0.09588	682	81	6	0.00322	0.00432	0.00651	0.00836
1341	91	40	0.01252	0.02230	0.0410	0.05559	1331	101	3	0.00316	0.00354	0.00416	0.00460
1348	80	43	0.01097	0.02009	0.03859	0.05385	629	90	3	0.00302	0.00340	0.00407	0.00456
1347	67	41	0.00914	0.01714	0.03422	0.04926	649	91	4	0.00301	0.00365	0.00486	0.00583
1346	60	32	0.00806	0.01462	0.02859	0.04081	685	94	4	0.00297	0.00362	0.00489	0.00593
1356	81	21	0.00688	0.01161	0.02117	0.02917	630	82	4	0.00291	0.00359	0.00491	0.00605
1340	95	16	0.00654	0.01036	0.01756	0.02314	251	106	1	0.00282	0.00261	0.00219	0.00189
1355	80	19	0.00631	0.01050	0.01902	0.02618	1339	96	2	0.00279	0.00290	0.00309	0.00325
1350	82	17	0.00603	0.00985	0.01754	0.02401	707	93	4	0.00276	0.00342	0.00479	0.00600
1354	76	18	0.00591	0.00974	0.01746	0.02383	222	119	0	0.00269	0.00210	0.00107	0.00036
1357	76	16	0.00572	0.00929	0.01656	0.02273	651	89	3	0.00267	0.00306	0.00382	0.00446
684	78	14	0.00547	0.00855	0.01446	0.01916	687	97	2	0.00266	0.00278	0.00299	0.00315
554	93	10	0.00531	0.00767	0.01182	0.01489	1333	60	5	0.00260	0.00343	0.00517	0.00672
1342	95	11	0.00502	0.00743	0.01191	0.01531	654	88	3	0.00259	0.00299	0.00375	0.00440
1337	83	12	0.00466	0.00716	0.01211	0.01617	652	83	2	0.00259	0.00273	0.00301	0.00324
653	82	12	0.00463	0.00712	0.01211	0.01623	305	109	0	0.00258	0.00202	0.00104	0.00035
1358	76	12	0.00448	0.00694	0.01193	0.01621	553	71	2	0.00258	0.00273	0.00300	0.00323
647	84	9	0.00444	0.00640	0.01021	0.01329	1330	60	3	0.00257	0.00296	0.00372	0.00436
686	102	8	0.00429	0.00598	0.00911	0.01151	286	111	0	0.00251	0.00197	0.00101	0.00035
650	89	8	0.00423	0.00595	0.00925	0.01192	33	109	0	0.00249	0.00196	0.00101	0.00035
683	79	9	0.00409	0.00594	0.00953	0.01245	1338	77	3	0.00249	0.00287	0.00363	0.00429
1332	88	6	0.00367	0.00482	0.00697	0.00865	1023	106	1	0.00246	0.00230	0.00198	0.00174
1345	74	9	0.00353	0.00522	0.00879	0.01192	612	75	2	0.00243	0.00258	0.00284	0.00306
646	85	5	0.00349	0.00444	0.00619	0.00759	1334	49	4	0.00243	0.00306	0.00440	0.00563
324	24	32	0.00345	0.00652	0.01356	0.02034	518	77	1	0.00242	0.00228	0.00199	0.00178

Table 5.2: Fifty first values of the APAM2 centrality for different α values

An essential point, that must be addressed in the comparison between the algorithms, is that the results they offer are coherent when they are applied to the same network. When talking about consistency in the results, it should not be understood that they must be the same. The central issue is that the measures offer a ranking of the network nodes, that is, a classification of the nodes in the network according to its importance within it.

The coherence of the values can be demonstrated by checking that the measures present a high correlation, in the sense that high values of one centrality correspond to high values of the other. If we refer to a mutual relationship or association between the centrality vectors studied, we must analyse the consistency of the results by means of correlation coefficients. In statistical terms, correlation is a method of assessing a possible two-way linear association between two variables. The correlation coefficient, which represents the strength of the putative linear association between the variables in question. It is a dimensionless quantity that takes a value in the range 1 to +1. A correlation coefficient of zero indicates that no linear relationship exists between two variables, and a correlation coefficient of 1 or +1 indicates

n	dg	data	CVP value	n	dg	data	CVP value
222	119	0	1.0746	305	109	0	0.9553
33	109	0	1.0541	286	111	0	0.9529
506	99	0	1.0367	123	94	0	0.9519
251	106	1	1.0353	122	98	0	0.9438
171	104	0	1.0197	719	105	0	0.9409
1331	101	3	1.0155	121	86	0	0.9356
1339	96	2	1.0140	1341	91	40	0.9321
1157	106	0	1.0138	1340	95	16	0.9304
53	92	0	1.0016	223	84	0	0.9302
629	90	3	0.9991	686	102	8	0.9296
1349	82	75	0.9970	652	83	2	0.9280
554	93	10	0.9966	650	89	8	0.9279
1332	88	6	0.9942	687	97	2	0.9253
647	84	9	0.9862	1342	95	11	0.9237
646	85	5	0.9776	1023	106	1	0.9226
627	86	0	0.9775	250	85	1	0.9222
1336	92	1	0.9774	304	86	0	0.9221
518	77	1	0.9739	143	98	0	0.9208
1357	76	16	0.9715	628	84	1	0.9178
684	78	14	0.9641	649	91	4	0.9166
1355	80	19	0.9633	1356	81	21	0.9156
1350	82	17	0.9614	651	89	3	0.9146
685	94	4	0.9593	593	81	0	0.9136
919	97	0	0.9585	1021	91	0	0.9128
683	79	9	0.9570	932	95	0	0.9122

Table 5.3: Fifty first values of the CVP centrality.

a perfect linear relationship. If the coefficient is a positive number, the variables are directly related and, in the other hand, if the coefficient is a negative number, the variables are inversely related. There are many different types of correlation coefficients that reflect somewhat different aspects of a monotone association and are interpreted differently in statistical analysis. In this Chapter, we focus on three popular indices that are often provided next to each other by standard software packages, namely Spearman, Pearson and Kendall coefficient.

Some tests calculating the Spearman, Person and Kendall correlation coefficients between APA-APAM2, APA-CVP and AMAP2-CVP have been developed, taking different values of the α parameter (see Table 5.4).

The results can be summarized in these key points:

- In the comparison between APA and APAM2 centralities, the values of the Spearman, Pearson and Kendall coefficients remain practically constant between 97.8% and 99.9%. This correlation values means that the the relation between APA and APAM2 is strong and although they are different measures, both centralities take into account the topology of the network and the data

Correlation between APA and APAM2

	$\alpha = 0.15$	$\alpha = 0.30$	$\alpha = 0.60$	$\alpha = 0.85$
Spearman	0.9989	0.9978	0.9959	0.9971
Pearson	0.9914	0.9876	0.9882	0.9898
Kendall	0.9893	0.9826	0.9782	0.9785

Correlation between APA and CVP

	$\alpha = 0.15$	$\alpha = 0.30$	$\alpha = 0.60$	$\alpha = 0.85$
Spearman	0.9940	0.9895	0.9783	0.9727
Pearson	0.7445	0.5247	0.3431	0.2793
Kendall	0.9429	0.9378	0.9074	0.8791

Correlation between APAM2 and CVP

	$\alpha = 0.15$	$\alpha = 0.30$	$\alpha = 0.60$	$\alpha = 0.85$
Spearman	0.9962	0.9947	0.9866	0.9777
Pearson	0.7527	0.5366	0.3555	0.2920
Kendall	0.9494	0.9498	0.9192	0.8868

Table 5.4: Pearson, Spearman and Kendall correlation coefficients.

present in it, being specially adapted for urban networks.

- Comparing APA and CVP, the Spearman coefficient remains constant with values close to 98%, the Kendall coefficient remains between 87.9% and 94.2%. This gives an idea of the degree of correlation between the two measures.

- A different situation occurs if we pay attention to the Pearson coefficient. If $\alpha = 0.15$ the measure has high relationship but, when the parameter α increases, the correlation decreases much until reaching 27.9%, which means almost absence of correlation. A possible explanation is that the Pearson coefficient measures the degree of covariation between different *linearly* related variables. This means that there may be strongly related variables, but not linearly, in which case Pearson's correlation does not provide sufficient information.

  In addition, the distortion caused by outliers in the behaviour of the correlation coefficient can be fairly large in some cases, especially when outliers are present in both variables at the same time.

- The comparison between APAM2 and CVP is very similar to the preceding

type.

Let us note, through the tables, the notable differences between CVP centrality and the APA centralities type. The CVP centrality of a node can be interpreted as the sum of values related to the amounts of data located along all the possible paths from a node. From this perspective, a high value of the CVP centrality of a node means that this node has a high amount of available data.

The studied algorithms establish a classification of the nodes of a network that allow us to develop some applications. From the urban planning point of view these models become an effective tool to rigorously evaluate the urban fabric because of we can monitor areas in the city which are most and least relevant, in terms of the effects correlated with this activity. Similarly, it is possible to evaluate the pre-urban development projects, as may be partial plans, master plans and interior renovation projects. We can also assess the value of the land based on the centrality of their closest node. In a wider field of use, these algorithms can be a tool to perform simulations in the whole network performing actions on discrete parts.

5.5 Conclusion

In this Chapter, we studied, analysed and compared centrality measures applied to an urban network. Two characteristics of these measures can be highlighted: they are based on the calculation of the eigenvector of a matrix and they are suitable for urban networks with data. One of these measures has the characteristic of covering a small range of values. This is inconvenient when working with networks with large amount of data and because of this, a new measure is presented. Once the centrality measures applied to urban networks have been analysed and studied, a network of the city of Rome is used together with data related to tourism, to apply the centrality measures described. Subsequently, a comparison of three measurements is made using the most usual correlation coefficients (Spearman, Pearson and Kendall). When comparing the APA and APAM2 centralities, a correlation is clearly seen. The three correlation coefficients studied show values close to one indicating similar measures.

"The only way to know how a complex system will behave-after you modify it-is to modify it and see how it behaves."

George E. P. Box

Chapter 6

APA for Biplex urban networks

In this Chapter, we propose a new algorithm for attributed multiplex networks with the main objective to compute the centrality of the nodes based on the original PageRank model used to establish a ranking in the Web pages network. Taking as a basis the Adapted PageRank Algorithm for monoplex networks with data and the two-layer PageRank approach, an algorithm for biplex networks is designed with two main characteristics. First, it solves the drawback of the existence of isolated nodes in any of the layers. Second, the algorithm allows to choose the value of the parameter controlling the importance assigned to the network topology and the data associated to the nodes in the Adapted PageRank Algorithm, respectively. The proposed algorithm inherits this ability to determine the importance of node attribute data in the calculation of the centrality; yet, going further, it allows to choose different parameter values for each of the two layers. The biplex algorithm is then generalised to the case of multiple layers, that is, for multiplex networks. Its possibilities and characteristics are demonstrated using the dataset of aggregate OD flows of private cars in Rome described in Chapter 4. Further, a biplex network is constructed by taking the data about car mobility for layer 1. Layer 2 is generated from data describing the local bus transport system. The algorithm then establishes the most central locations in the city when these layers are intertwined with the location attributes in the biplex network. Four cases are evaluated and compared for different values of the parameter that modulates the importance of data in the network.

This Chapter is a modified version of our paper Leandro Tortosa, Jose F Vicent, and Gevorg Yeghikyan. A centrality measure based on the Adapted PageRank Algorithm for multiplex networks with data. *Applied Mathematics and Computation*, 2020 (under review).

6.1 Introduction

Identifying influential vertices in a network can be useful in many practical fields. Examples of this are risk identification in infrastructure [246], determining influential nodes in social networks [81], ensuring the security and reliability of the network [300], collaborating with the most influential media for advertising [170, 142], evaluating the influence of junctions with the aim of avoiding overloading roads [132, 276], or defining the influence maximization problem as an algorithmic problem [224].

Various centrality measures including closeness, degree, and betweenness centrality [103] are widely used to this end, with the choice of the measure depending on the specific application. Further, PageRank and related algorithms have been proposed, extending the concept of network centrality and the range of applications [2, 4]. The main idea of the PageRank algorithm proposed by Page et al. [201] is that a network node is relevant if other important nodes have a link to it. If a node with a high PR value is linked to another node, the value of the page being linked increases. It plays a crucial role in the ranking of nodes in complex networks [194, 303]. Many scientists have used the PageRank and its modifications to address various problems. In [253], Wu and Chen introduce a hierarchical hybrid ranking algorithm to study entrepreneurship and innovation activities. The relative importance of scientific articles based on PageRank is presented in [273]. A social activity ranking method based on the PageRank algorithm is introduced in [190]. Ma et al. [259] create a novel ImageRank algorithm for image retrieval and relevance feedback. In [174], the authors design an algorithm similar to PageRank to identify important news events.

One of the most important characteristics of complex systems is that the collective behaviour of the system cannot be predicted from the properties of its components. The many types of inter-dependencies call for new ways to represent networks

in which nodes have more than one type of interaction between them. These systems –called multiplex networks– are characterized by different layers representing different interaction types between nodes. An overview of research on multiplex networks can be found in [46, 148]. Modelled by a set of networks with interacting layers, these multilayer networks have been used to describe a many real-world complex systems, such as financial [58], ecological [215], information [130], urban [3], and transportation networks [108]. The recent advances in Big Data technologies allow to capture more and more types of relations in observed systems. In this context, it may be advantageous to represent and study these systems by representing them by multiplex networks [47, 44, 113].

Multiplex networks allow to connect pairs of nodes with multiple links in multiple layers. In addition, the ability to capture relations *between* layers is also important in modelling and explaining empirical multilayer networks. Ranking the nodes of these multiplex networks requires to highlight the importance of nodes in each of the interdependent layers [239]. In [118], Halu et al. proposed a PageRank algorithm for measuring node centralities in multiplex networks by introducing a bias exerted by a network layer on the jumps of the random walk in another layer. However, in many real world networks, attributes described by data intrinsic to the nodes play an important role and require further modelling. In order to take into account the data intrinsic to nodes, in this Chapter, we present a network centrality measure for biplex networks. The proposed algorithm is based on the Adapted PageRank Algorithm (APA) centrality [2, 4] and is further is extended in a natural way to multiplex networks. A key aspect of the proposed algorithm is that, built upon the APA centrality, the random walk jumps in the algorithm are modelled by the network node attribute data. This allows us to study a range of relationships between among nodes modelled by different layers, as well as to measure the influence of the node attribute data in each of the network layers. In addition, the proposed algorithm can be applied in any multilayer setting, since it avoids the problem of isolated nodes in any of the layers.

To achieve these objectives, the structure of this Chapter is as follows. A method to construct a multiplex centrality based is the APA model is presented in Section

6.2 . In Section 6.3, the biplex centrality algorithm is modified with the aim to avoid the existence of dangling nodes. Then, some characteristics about the meaning of the parameter *alpha* are discussed (Section 6.4). Section 6.5 aims to show an extension of the centrality to multiplex networks. Numerical results after analysing a real urban network in Rome are presented in Section 6.6. Finally, the conclusions of the work are exposed.

6.2 Building Multiplex centrality from APA

Some classical notation for graphs will be used. So, a graph is represented by $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with $\mathcal{N}$ a set of n vertices or nodes and $\mathcal{E}$ a set of links between the nodes. The links are represented by the adjacency matrix $A = (a_{ij})$ square of size $n \times n$, where

$$a_{ij} = \begin{cases} 1 & \text{if it exists a link between the nodes } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases}$$

6.2.1 Previous work

This section briefly describes the steps that take us from the original system to the model proposed to measure the centrality of the nodes of a multiplex complex network.

PageRank is an algorithm, based on the webgraph, that produces a classification of the web pages according to their importance.

The core of PageRank is the construction of the called *Google matrix*

$$G_{ij} = \alpha S_{ij} + (1 - \alpha)\frac{1}{N}, \tag{6.1}$$

whereS_{ij} is, by columns, an stochastic matrix obtained from the adjacency matrix of the graph and the number of outgoing links from a node to the rest. The existence of null columns in S is a consequence of isolated nodes, which is solved by introducing a constant value $1/N$ –with N the number of nodes–. The parameter labeled as α is known as the *damping factor*. As it is well-known, the spectral properties

of G, defined by (6.1) causes Perron-Frobenius's theorem to be satisfied, so for $0 < \alpha < 1$, there exists a unique and non-negative eigenvector associated to the maximal eigenvalue $\lambda = 1$. This vector –called PageRank vector– constitutes the rank of the nodes.

In the APA model described by the Adapted PageRank Algorithm (see [2], page 2190), a similar reasoning is used to determine a ranking of the nodes according to their importance.

In this case, the core of the model is the construction of a matrix M_{APA} given by

$$M_{APA}(ij) = (1 - \alpha)P_{ij} + \alpha V. \tag{6.2}$$

The *transition matrix* P_{ij} is defined as

$$p_{ij} = \begin{cases} \frac{1}{c_j} & \text{if } a_{ij} \neq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad 1 \leq i,\ j \leq n, \tag{6.3}$$

with c_k representing the sum of the k-th column of the adjacency matrix. The matrix V in the second term of equation (6.2) is constructed from a data matrix associated to the network. For more details, see [2].

It is important to highlight that the M_{APA} matrix, due to the way it is defined by (6.2) and (6.3), inherits the spectral characteristics of the Google matrix, with the property of being a stochastic matrix by columns. This fact assures us the existence of the dominant eigenvalue $\lambda = 1$ and the consequent right-side eigenvector that constitutes the expected classification of the nodes.

The second idea in which it is based the proposed algorithm for multiplex networks is the PageRank approach described by Pedroche et al. [212], known as *two-layer approach*. They state that Google matrix (6.1) may be divided into two terms and associated to two different layers representing the network. On the one hand, the *physical layer*, given by the term αS, and, on the other side, a *teleportation layer*, given by the term $1/N$.

In mathematical terms, Pedroche et al. [212] construct the 2×2 block matrix

$$M_A = \left(\begin{array}{c|c} \alpha P_A & (1-\alpha)I \\ \hline 2\alpha I & (1-\alpha)\mathbf{e}\mathbf{v}^T \end{array} \right) \quad \in^{2n \times 2n} . \tag{6.4}$$

where M_A represents a Markov chain.

The algebraic characteristics of M_A –irreducible and primitive– allows us to affirm that

$$\hat{\pi}_A = \pi_u + \pi_d \ \in^n,$$

where $\left[\pi_u^T \ \pi_d^T\right]^T \in^{2n}$, is the only positive and normalised eigenvector. The structure of this matrix will be generalised for the case of multiplex networks as it will be described in the following section.

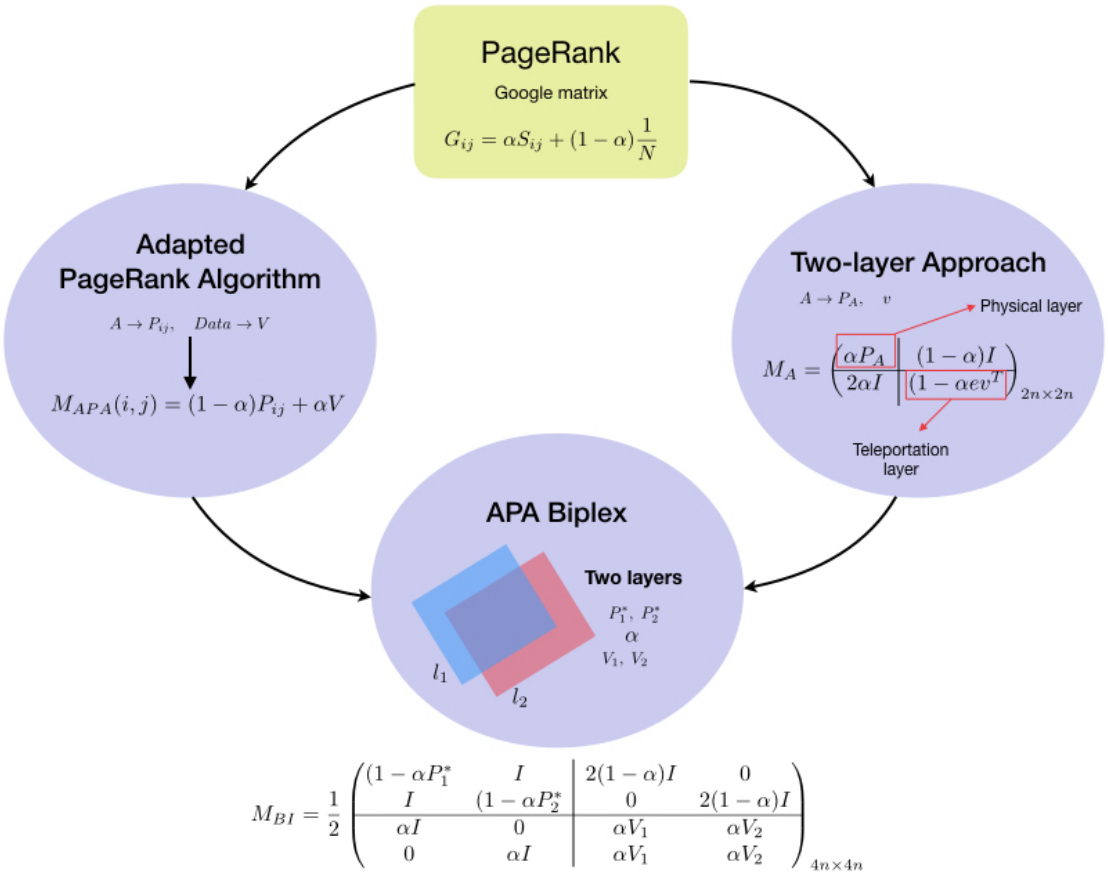


Figure 6-1: Schematic representation of the models used to design the APA biplex centrality algorithm

In Figure 6-1 an schematic graphic of the models used to design and implement

the APA biplex algorithm for calculating the nodes' centrality in biplex networks is shown. For both detailed description –the APA algorithm and the two-layers PageRank approach–, see [2, 3, 212].

6.2.2 Constructing biplex centrality from APA and the two-layer approach

Taking into account that a multiplex networks is a case in which there are different relationships in each layer but the same nodes in all of them, we can extend the two layer approach to the case of multiplex netwoks. Behind this process is the idea of applying the two-layer model to every layer of the multiplex network.

Let us follow the classical notation for a multiplex network $\mathcal{M} = (\mathcal{N}, \mathcal{E}, \mathcal{S})$ with $\mathcal{S} = (l_1, l_2, \ldots, l_k)$ a set of layers.

Considering the simplest case of a biplex networks $\mathcal{M} = (\mathcal{N}, \mathcal{E}, \mathcal{S})$, with two layers $\mathcal{S} = (l_1, l_2)$ with adjacency matrices $A_1, A_2 \in R^{n\times n}$.

A matrix M_2 ([212]) is constructed as

$$M_2 = \frac{1}{2}\left(\begin{array}{cc|cc} \alpha P_1 & I & (1-\alpha)I & 0 \\ I & \alpha P_2 & 0 & (1-\alpha)I \\ \hline 2\alpha I & 0 & (1-\alpha)\mathbf{ev}_1^T & (1-\alpha)\mathbf{ev}_2^T \\ 0 & 2\alpha I & (1-\alpha)\mathbf{ev}_1^T & (1-\alpha)\mathbf{ev}_2^T \end{array}\right) \tag{6.5}$$

with P_i, for $i = 1, 2$ a stochastic matrices and $\mathbf{v}_i$, for $i = 1, 2$ personalized vectors.

Thinking in terms of the APA algorithm, and taking the matrix M_{APA} (6.2) as the reference, a 2×2 block matrix is built

$$M_{APA2} = \left(\begin{array}{c|c} \alpha P_A & (1-\alpha)I \\ \hline \alpha I & \alpha V \end{array}\right) \quad \in^{2n\times 2n} . \tag{6.6}$$

An algebraic study of the spectral characteristics of M_{APA2} allows us to affirm that this matrix is a class of Perron-Frobenius operators, ensuring the existence and uniqueness of the eigenvector associated with the dominant eigenvalue.

The following step consists on extending the two-layers APA model given by

(6.6). Reordering the blocks so that the physical layer appear in the position $(1,1)$ and the data layer in the block $(2,2)$, we have

$$M_{BI} = \frac{1}{2}\left(\begin{array}{cc|cc} (1-\alpha)P_1 & I & 2(1-\alpha)I & 0 \\ I & (1-\alpha)P_2 & 0 & 2(1-\alpha)I \\ \hline \alpha I & 0 & \alpha V_1 & \alpha V_2 \\ 0 & \alpha I & \alpha V_1 & \alpha V_2 \end{array}\right) \tag{6.7}$$

with P_1 and P_2 stochastic matrices by columns and V_i, for $i = 1, 2$, data matrices of the two layers respectively.

As a consequencece, there exists an eigenvector

$$\hat{\pi}_{\mathbf{BI}} = (\pi_{\mathbf{u_1}}, \pi_{\mathbf{u_2}}, \pi_{\mathbf{d_1}}, \pi_{\mathbf{d_2}}) \ \in^{4n} \tag{6.8}$$

associated to $\lambda = 1$ (largest eigenvalue).

This vector is essential and represents the node's centralities. Therefore, a unique vector can be obtained.

$$\mathbf{x} = \frac{1}{2}\left(\pi_{\mathbf{u_1}} + \pi_{\mathbf{u_2}} + \pi_{\mathbf{d_1}} + \pi_{\mathbf{d_2}}\right) \ \in^{n}, \tag{6.9}$$

with $\mathbf{x}$ being a normalized vector.

6.2.3 Problems with dangling nodes in multiplex networks

Let us consider a biplex network with n nodes $\mathcal{N} = \{1, 2, \ldots, n\}$, and two layers, l_1 and l_2 corresponding to two different relationships between nodes that give rise to the adjacency matrices A_1 and A_2.

Because the nodes in the two layers are the same and the relationships are different, it is possible to for dangling nodes to appear in each of the layers; that is, nodes with no link to other nodes. In this way, we have nodes with degree greater than zero (*real nodes*) and nodes with degree equal to zero (*virtual nodes*). This has an undesirable effect on the spectral properties of the transition matrix P that is designed to be irreducible and stochastic by columns. The appearance of

rows and columns of zeros corresponding to the virtual nodes causes P not to be stochastic, making the numerical resolution of the system based on the calculation of the PageRank vector impossible.

Let us discuss this instability with a simple case. Considering a biplex network with 10 nodes and the following two adjacency matrices

$$\mathrm{A}_1 = \begin{pmatrix} 0&0&0&1&0&0&0&0&0&1 \\ 0&0&1&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&1&0 \\ 1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&0&1&0&0&0 \\ 0&0&0&0&0&1&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&1&0&0&0 \\ 1&0&0&0&0&0&0&0&0&0 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 0&0&1&1&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&1&0 \\ 1&0&0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&1&0&0&0 \\ 0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \end{pmatrix}. \tag{6.10}$$

We can observe from the adjacency matrices that in layer l_1 there is a dangling node (node 8) while in layer 2 there are two dangling nodes (nodes 2 and 10).

If we take A_1 from the first layer and calculate the transition matrix P_1, we notice that the sum vector of the columns is $\mathbf{sum} = [2, 1, 2, 1, 1, 2, 2, 0, 2, 1]$. This leads us to a transition matrix

$$\mathrm{P}_1 = \begin{pmatrix} 0&0&0&1&0&0&0&NaN&0&1 \\ 0&0&1/2&0&0&0&0&NaN&0&0 \\ 0&1&0&0&0&0&0&NaN&1/2&0 \\ 1/2&0&0&0&0&0&0&NaN&0&0 \\ 0&0&0&0&0&1/2&0&NaN&0&0 \\ 0&0&0&0&1&0&1/2&NaN&0&0 \\ 0&0&0&0&0&1/2&0&NaN&1/2&0 \\ 0&0&0&0&0&0&0&NaN&0&0 \\ 0&0&1/2&0&0&0&1/2&NaN&0&0 \\ 1/2&0&0&0&0&0&0&NaN&0&0 \end{pmatrix},$$

where *NaN* is the acronym for *Not a Number*, the result of dividing by zero. As a consequence of this, the matrix

$$M_{APA} = (1-\alpha)P_1 + \alpha V_1$$

will not be stochastic by columns and it would be impossible to obtain a classification vector for the nodes.

If you look at layer l_2, the adjacency matrix A_2 has two rows of zeros, at nodes 2 and 10, which makes them dangling nodes. We deal with the same problem as in layer l_1 although now errors occur in the columns 2 and 10. The solution to this problem is addressed in the next section.

6.3 Adapting biplex centrality for dangling nodes

To solve the problem of isolated nodes, we must reformulate the basic principles of the PageRank model and the definition of the Google matrix for the most generic case. Let us consider the graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with dangling nodes, that is, nodes with zero degree. From the adjacency matrix $A = (a_{ij})$ of the graph G, the matrix $P = (p_{ij})$ is redefined as:

$$p_{ij} = \begin{cases} 0 & \text{if } i \text{ is a dangling node, for all} \quad j = 1, 2, \ldots, n. \\ \frac{1}{c_j} & \text{if} \quad a_{ij} \neq 0. \end{cases} \tag{6.11}$$

In this case, with the aim to construct a column stochastic matrix, we substitute the matrix P in the calculation of the matrix M_{APA} with a new matrix P^* given by

$$P^* = P + \mathbf{d} \cdot \mathbf{u}^T, \tag{6.12}$$

where $\mathbf{d}$ is the vector characterizing the dangling nodes defined as

$$\mathbf{d}_i = \begin{cases} 1, & \text{if } i \text{ is a dangling node,} \\ 0, & \text{otherwise,} \end{cases}$$

and $\mathbf{u} \in \Re^n$ characterizes the distribution of dangling nodes such that $\mathbf{u} > 0$ and $\mathbf{u}^T\mathbf{e} = 1$, with $\mathbf{e} = (1, 1, \ldots, 1)$.

Rewriting P as (8.4) we make sure that we continue working with a stochastic matrix by columns so that we have the proper spectral features to obtain the

classification vector.

Another aspect that must be highlighted is related to the construction of the V matrix based on the individual data associated with each node of the graph. For those *virtual nodes* that have no connection with the rest of the nodes, the data associated to them is zero. That means that the data vector $\mathbf{v}$ has a zero component in all the positions corresponding to dangling nodes.

It may be the case of having a large number of virtual nodes, which would lead to a high number of null values in V. On the other hand, it may be convenient, not only from the point of view of numerical stability, but also from the idea of the influence of the whole data in the network, to introduce a small coefficient in places that have a null value of data. This small coefficient, that will be denoted as *coef*, may be defined as

$$coef = \frac{\min(D) > 0}{k - n}, \tag{6.13}$$

where D is the data matrix, k is the quantity of dangling nodes and n the number of nodes.

From this coefficient, it is possible to redefine the data vector $\mathbf{v}$ adding these small values at the positions of dangling nodes. This new data vector $\mathbf{v}^{up}$ is then given by

$$\mathbf{v}^{up}(i) = \begin{cases} \mathbf{v}(i), & \text{if } i \text{ has nonzero degree,} \\ coef, & \text{otherwise,} \end{cases}$$

As can be seen, the introduced coefficient represents a small value obtained by dividing the minimum data value associated to the nodes by the total number of dangling nodes.

Taking into account the modifications proposed regarding the matrix P^* and vector $\mathbf{v}^{up}$, the matrix from which the centrality measure will be obtained may be rewritten as

$$M_{BI} = \frac{1}{2} \left(\begin{array}{cc|cc} (1-\alpha)P_1^* & I & 2(1-\alpha)I & 0 \\ I & (1-\alpha)P_2^* & 0 & 2(1-\alpha)I \\ \hline \alpha I & 0 & \alpha V_1 & \alpha V_2 \\ 0 & \alpha I & \alpha V_1 & \alpha V_2 \end{array} \right) \tag{6.14}$$

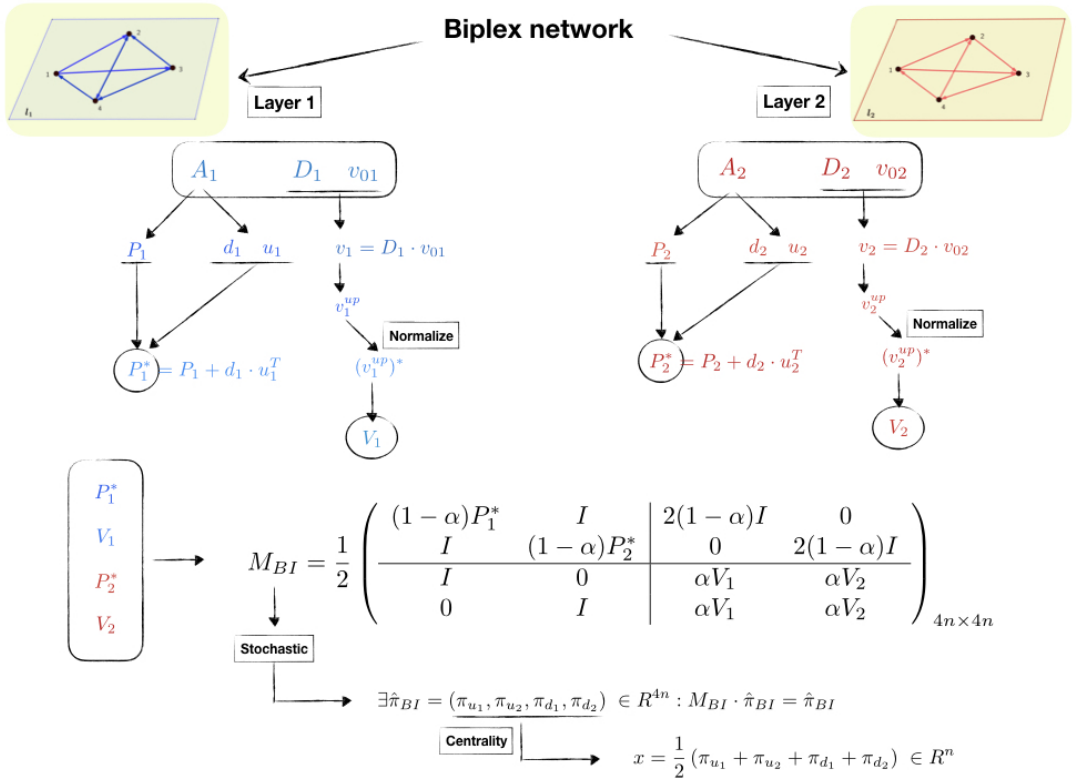


Figure 6-2: Schematic representation of the APA biplex centrality model

Figure 6-2 details a graphic scheme of the centrality model for biplex networks –based on the APA algorithm– with dangling nodes. Note how the most notable changes with respect to the previous model occur in the construction of the matrices P_1^* and P_2^* as well as in the inclusion of a residual value in the data vector representing nodes with zero connectivity in any of the layers.

The scheme shown in Figure 6-2 can be summarized in Algorithm 5.

Following the simple example studied above with 10 nodes, where there were two layers with the adjacency matrices given by (6.10). Let us assume that the data matrices D_1, D_2 are

$$D_1 = [2, 2, 5, 2, 1, 3, 2, 0, 7, 2]^T, \quad D_2 = [4, 0, 5, 6, 1, 5, 2, 4, 3, 0]^T.$$

From the adjacency matrices, we detect the dangling nodes and, using the defi-

Input: Let $\mathcal{M} = (\mathcal{N}, \mathcal{E}, \mathcal{S})$ be a biplex network with $\mathcal{N} = \{1, 2, \ldots, n\}$ the set of nodes, $\mathcal{S} = (l_1, l_2)$ two layers and A_1 and A_2 the respective adjacency matrices.
Let D_1 and D_2 the data matrices associated to nodes in layers l_1 and l_2, respectively, and weighted vectors $\mathbf{v_{01}}$ and $\mathbf{v_{02}}$, respectively.
Output: $\mathbf{x}$ representing the graph centrality
begin
For the layers l_i, for $i = 1, 2$, construct the vectors and matrices:
– P_i, the probability matrices from (6.3)
– vectors $\mathbf{d}_i, \mathbf{u}_i$ from the adjacency matrices A_i
Compute P_i^*, for $i = 1, 2$, from (8.4)
Compute the data vectors $\mathbf{v_i}$, for $i = 1, 2$, as $\mathbf{v_i} = D_i \cdot \mathbf{v_{0i}}$
Compute the coefficients $coef_i$, for $i = 1, 2$, from (6.8)
From $\mathbf{v}_i$ and $coef_i$, for $i = 1, 2$, compute $\mathbf{v}_i^{up}$
Normalize $\mathbf{v}_i^{up}$, for $i = 1, 2$ and denote it as $\{\mathbf{v}_i^{up}\}^*$
Construct V_i, for $i = 1, 2$
Construct the matrix M_{BI} from (6.14)
Compute the dominant eigenvector $\hat{\pi}_{\mathbf{BI}}$ of M_{BI}
Compute the centrality $\mathbf{x}$
end

Algorithm 5: APA biplex networks algorithm for computing the node's centrality.

nitions of the vectors $\mathbf{d}_i$ and $\mathbf{u}_i$, for $i = 1, 2$, we have that

$$\mathbf{d}_1 = [0,0,0,0,0,0,0,1,0,0]^T, \quad \mathbf{d}_2 = [0,1,0,0,0,0,0,0,0,1]^T.$$

$$\mathbf{u}_1 = [0,0,0,0,0,0,0,1,0,0]^T, \quad \mathbf{u}_2 = [0,1/2,0,0,0,0,0,0,0,1/2]^T.$$

Once $P_i, \mathbf{d}_i, \mathbf{u}_i$ have been computed, the matrices P_i^* are obtained as

$$P_1^* = P_1 + \mathbf{d}_1 \cdot \mathbf{u}_1^T = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{6.15}$$

and

$$P_2^* = P_2 + \mathbf{d}_2 \cdot \mathbf{u}_2^T = \begin{pmatrix} 0 & 0 & 1/2 & 1 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \end{pmatrix}. \tag{6.16}$$

Note that both P_i^* given by (6.15–6.16) are now column stochastic matrices.

Regarding the data vectors, we update vectors $\mathbf{v}_i$ to obtain vectors $\mathbf{v}_i^{up}$, for $i = 1, 2$, by computing the following parameters involved in the updating process.

$$\min(D_1 > 0) = 1, \quad \min(D_2 > 0) = 1, \quad k_1 = 1, \quad k_2 = 2.$$

Using the expression (6.13), the respective coefficients are

$$coef_1 = 1, \quad coef_2 = 1/2.$$

In this way, the updated data vectors are

$$\mathbf{v}_1^{up} = [2, 2, 5, 2, 1, 3, 2, 1, 7, 2]^T, \quad \mathbf{v}_2^{up} = [4, 1/2, 5, 6, 1, 5, 2, 4, 3, 1/2]^T.$$

6.4 Some considerations about the α parameter

.

In the PageRank model, the parameter α represented the probability that a "surfer" follows the links of a web page uniformly and randomly. In this context, –of random surfer– the choice of α is not well justified, although in the original proposal of Page and Brin [204] they used $\alpha = 0.85$. In the extensive literature for web surfer, two choices are highlighted: $\alpha = 0.85$ and $\alpha = 0.5$ (see, for instance, [68]).

In the APA centrality algorithm, the α parameter has a different meaning, since it represents the importance we attach to the data associated with each node, while the value $(1 - \alpha)$ represents the importance that we assign to the topology of the network we are studying.

In the design and assessment process of the different algorithms, both for single-layer networks and multi-layer networks, no mention about the alpha parameter at any time is done, since we assume that it is a fixed value that is initially chosen and used in both layers of the biplex network. This means that an equal importance to the data and the network connectivity in both layers is given. However, depending on the application context, it may happen that the node attribute data is more important in one of the layers and not in the other. If the same value of the

parameter in the two layers is chosen, these differences will not be considered. The objective is to be able to set different parameter values for the different layers. We could consider the node attribute data to be more important than the topology in one layer but not the other. To contemplate these multiple possibilities, we need to introduce α_i, for layers l_i, with $i = 1, 2$.

Therefore, the matrix M_{BI} may be adapted as:

$$M_{BI} = \frac{1}{2}\left(\begin{array}{cc|cc} (1-\alpha_1)P_1^* & I & 2(1-\alpha_1)I & 0 \\ I & (1-\alpha_2)P_2^* & 0 & 2(1-\alpha_2)I \\ \hline \alpha_1 I & 0 & \alpha_1 V_1 & \alpha_2 V_2 \\ 0 & \alpha_2 I & \alpha_1 V_1 & \alpha_2 V_2 \end{array}\right) \tag{6.17}$$

and Algorithm 5 modified as:

Input: Let $\mathcal{M} = (\mathcal{N}, \mathcal{E}, \mathcal{S})$ be a biplex network with $\mathcal{N} = \{1, 2, \ldots, n\}$ the set of nodes, $\mathcal{S} = (l_1, l_2)$ two layers and A_1 and A_2 the respective adjacency matrices.
Let α_1 and α_2 be the values of the parameter α for layers l_1 and l_2, respectively. Let D_1 and D_2 be the data matrices associated to nodes in layers l_1 and l_2, respectively, and weighted vectors $\mathbf{v_{01}}$ and $\mathbf{v_{02}}$, respectively. **Output:** $\mathbf{x}$ representing the graph centrality
begin
For the layers l_i, for $i = 1, 2$, construct the vectors and matrices:
– P_i, the probability matrices from (6.3)
– vectors $\mathbf{d}_i, \mathbf{u}_i$ from the adjacency matrices A_i
Compute P_i^*, for $i = 1, 2$, from (8.4)
Compute the data vectors $\mathbf{v_i}$, for $i = 1, 2$, as $\mathbf{v_i} = D_i \cdot \mathbf{v_{0i}}$
Compute the coefficients $coef_i$, for $i = 1, 2$, from (6.8)
From $\mathbf{v}_i$ and $coef_i$, for $i = 1, 2$, compute $\mathbf{v}_i^{up}$
Normalize $\mathbf{v}_i^{up}$, for $i = 1, 2$ and denote it as $\{\mathbf{v}_i^{up}\}^*$
Construct V_i, for $i = 1, 2$
Construct the matrix M_{BI} from (6.17)
Compute the dominant eigenvector $\hat{\pi}_{\mathbf{BI}}$ of the matrix M_{BI}
Compute the centrality $\mathbf{x}$
end

Algorithm 6: APA biplex networks algorithm to compute the node's centrality using different values of the α parameter.

Although the changes in the algorithms are minimal, the possibilities offered by

being able to choose different alpha values in each layer are very important, since it allows to establish the importance that we associate to the node attribute data in each of the layers.

6.5 Extending centrality to multiplex networks

.

Matrix M_{BI}, given by the expression 6.17, is the key in the entire process of the algorithm construction leading to the computation of the centrality. The spectral properties of this matrix and its stochastic characteristic allow us to calculate its dominant eigenvector that represents of the ranking of the nodes. A closer look at the structure of the matrix reveals that its extension to the case of multiple layers is easy. The block structure of this matrix favors its natural extension as we see below.

We have to remark that M_{BI} is constructed for biplex networks. Let us assume that we have a multiplex network with k layers $\{l_1, l_2, \ldots, l_k\}$, a set of adjacency matrices $\{A_1, A_2, \ldots A_k\}$ and k data matrices $\{D_1, D_2, \ldots, D_k\}$. Let $\alpha_1, \alpha_2, \ldots, \alpha_k$ be the parameter values for the layers $l_1, l_2, \ldots, l_k$, respectively. Then, M_{BI} may be extended to multiplex networks as

$$M_{multi} = \frac{1}{k}\left(\begin{array}{c|c} M_{1,1} & M_{1,2} \\ \hline M_{2,1} & M_{2,2} \end{array}\right) \tag{6.18}$$

with

$$\mathrm{M}_{1,1} = \begin{pmatrix} (1-\alpha_1)P_1^* & I & \cdots & I \\ I & (1-\alpha_2)P_2^* & \cdots & I \\ \cdots & \cdots & \cdots & \cdots \\ I & I & \cdots & (1-\alpha_k)P_k^* \end{pmatrix}, \; M_{2,2} = \begin{pmatrix} \alpha_1 V_1 & \alpha_2 V_2 & \cdots & \alpha_k V_k \\ \alpha_1 V_1 & \alpha_2 V_2 & \cdots & \alpha_k V_k \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_1 V_1 & \alpha_2 V_2 & \cdots & \alpha_k V_k \end{pmatrix}.$$

Both $M_{1,2}$ and $M_{2,1}$ are block diagonal matrices and are given by

$$\mathrm{M}_{1,2} = \begin{pmatrix} 2(1-\alpha_1)I & 0 & \cdots & 0 \\ 0 & 2(1-\alpha_2)I & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 2(1-\alpha_k)I \end{pmatrix}, \; M_{2,1} = \begin{pmatrix} \alpha_1 I & 0 & \cdots & 0 \\ 0 & \alpha_2 I & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \alpha_k I \end{pmatrix}.$$

The matrix M_{multi} constructed as shown in equation (7.1) inherits the spectral properties of M_{BI}; therefore, the dominant eigenvector allows us to compute the centrality $\mathbf{x}$. We must also comment that the size of the matrix M_{multi} grows remarkably when we are adding layers so it will be necessary to optimize the calculation of the dominant eigenvector when the number of layers is high.

6.6 Numerical results

Nowadays, we show a real example of biplex network related to the urban network of the city of Rome, Italy. Data about car flows and the public bus transport system will be used to analyse and determine the most central areas of the city when both data are studied. To perform this, let us first briefly describe the dataset used for the numerical example.

6.6.1 Rome dataset

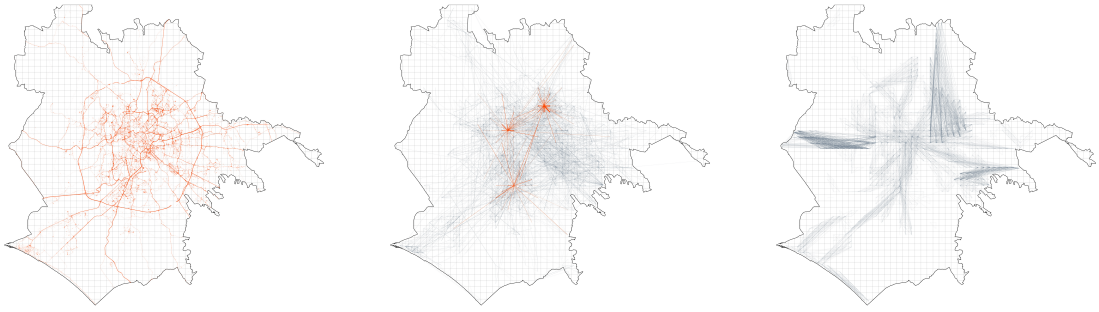


Figure 6-3: (*left*) Private car GPS trajectories superimposed on the grid in Rome (*middle*) Layer 1 of biplex network: Rome OD network with some popular locations highlighted (*right*) Layer 2 of biplex network: bus connection network.

6.6.2 The numerical results

As described in Section 6.6.1, we build a biplex network for the city of Rome with the nodes being the centroids of the grid representing the urban streets network. The two layers of the biplex network will represent the car flows between nodes and the bus transport system, respectively. As the biplex model allows us to establish different relationships between the nodes using different data, the idea in this case study is to analyse and visualise the relationship between a public transport system such as the urban bus connectivity and the car OD flows between different city locations.

Consequently, two layers may be defined with these characteristics:

- **Layer 1**: the graph is composed of the nodes of the urban network and an edge is drawn between two nodes if there is at least one car unit flow between these nodes. The attribute data associated to every node is the total quantity of in- and out-flows from the node.

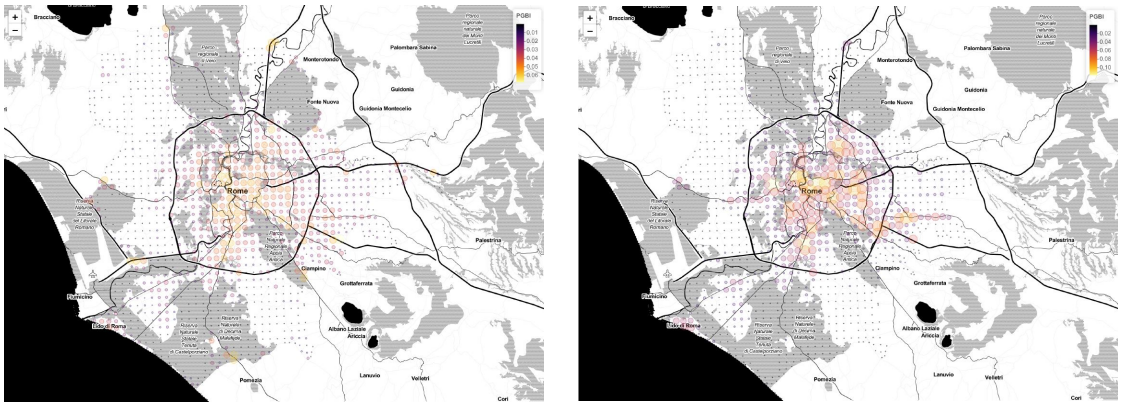
- **Layer 2**: this layer graph has the same nodes, and two nodes are linked by an edge if there exists non-zero car flow between them and there also exists at least a bus line connecting them. The data associated to the nodes is the total number of bus lines connecting a node with the remaining ones.

  For instance, node number 9 is linked by at least one OD car flow with the nodes 1, 298, 416, 633, 713, 715, 730, 775, 999, 1083, 1087, and 1486. However, in layer 2 node 9 is linked only with the nodes 1 and 416 since there exists at least one bus line connecting the nodes. More precisely, between nodes 9 and 1 there are 6 lines connecting them and there are 2 bus lines connecting the nodes 9 and 416. Therefore, the data associated to the node 9 is $6 + 2 = 8$.

In this example we are quantifying and analysing urban mobility as well as the bus transport system. The advantage to work with two or more layers is that it is possible to measure several relationships with different datasets between nodes, which is not possible in networks with only one layer.

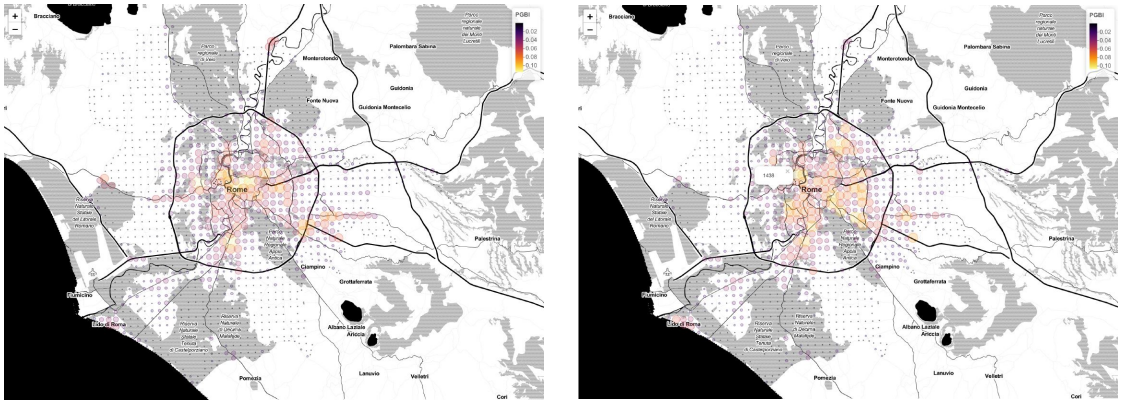
Another characteristic of this model is the possibility of differentiating the importance assigned to the data in each of the layers. In the definition of matrix M_{BI}

given by the expression 6.17 each of the blocks has its own parameter α_i. This allows us to consider giving more importance to the data associated to the nodes in the first layer or, on the contrary, giving more value to the data in the second layer.



(a) (b)

Figure 6-4: Biplex centrality PGBI for (a) $\alpha_1 = \alpha_2 = 0.2$ and (b) $\alpha_1 = \alpha_2 = 0.8$



(a) (b)

Figure 6-5: Biplex centrality PGBI for (a) $\alpha_1 = 0.3$, $\alpha_2 = 0.8$ and (b) $\alpha_1 = 0.8$ $\alpha_2 = 0.3$

Case 1: $\alpha_1 = 0.2, \alpha_2 = 0.2$ In the first analysed case, we choose the same value of the parameter for both layers, that is, $\alpha_i = 0.2$, for $i = 1, 2$. This choice of parameters means that we are giving the same importance to data in both layers. Note that, following the original APA algorithm, the parameter varies in the interval $(0, 1)$; therefore, a low value of α_i also means that we give much more importance to the network topology than to the node attribute data. On the contrary, if we give a high value for the parameter, we are giving

much more importance to the data associated to the nodes than to the graph topology. We executed Algorithm 6 to obtain the biplex centrality of all the nodes according to the model described in this Chapter. In Figure 6-4a we show the map of Rome with the values of the biplex centrality, that we will denote in the following as PGBI, plotting the size of the nodes proportional to their centrality value.

Case 2: $\alpha_1 = 0.8, \alpha_2 = 0.8$ In this case, we also choose the same value of the α parameter for both layers, that is, $\alpha_i = 0.8$, for $i = 1, 2$. However, as opposed to case 1, the value of the parameter is $\alpha = 0.8$, which means that we are giving much more importance to the data than to the connectivity in the individual layers. Now, we are more interested in measuring the influence of data than the influence of the links between the nodes. The values of the biplex centrality for this choice of α can be seen in Figure 6-4b. We can see the differences with respect to the first case.

Case 3: $\alpha_1 = 0.3, \alpha_2 = 0.8$ This case under study introduces a new variant since now the values taken by the alpha parameter are different for each layer. Thus, the α_1 value for layer 1 is 0.3, which means that in layer 1 we give less importance to the data and greater to the connectivity within the network. However, we take the value $\alpha_2 = 0.8$ for layer 2, which means that we measure the importance of the nodes in this layer giving more importance to the data and less importance to the connectivity. We can see a visualization of the biplex centrality for this case in Figure 6-5a.

Case 4: $\alpha_1 = 0.8, \alpha_2 = 0.3$ This case is similar to case 3 with the difference that now $\alpha_1 = 0.8$ and $\alpha_2 = 0.3$. This means that we give more importance to the node attribute data in the layer 1 and to the connectivity in layer 2. The results are displayed in Figure 6-5b.

Table 6.1 summarizes the results of the 25 most central nodes for the four studied cases, based on the choice of α_1 and α_2 in the two layers of the network. The centrality is shown in the column labeled PGBI.

	$\alpha_1 = 0.2$, $\alpha_2 = 0.2$		$\alpha_1 = 0.3$, $\alpha_2 = 0.8$		$\alpha_1 = 0.8$, $\alpha_2 = 0.3$		$\alpha_1 = 0.8$, $\alpha_2 = 0.8$	
ranking	node	PGBI	node	PGBI	node	PGBI	node	PGBI
1	270	0.068118	1437	0.114870	790	0.107058	1437	0.117651
2	367	0.060640	1485	0.114636	1460	0.099745	634	0.115931
3	634	0.060223	634	0.110817	776	0.096553	1485	0.112266
4	1460	0.059685	1478	0.107259	1122	0.095818	1460	0.106568
5	809	0.059439	740	0.104394	634	0.092842	790	0.104531
6	340	0.058988	1469	0.092490	673	0.092495	776	0.100668
7	44	0.058983	1460	0.091879	777	0.091264	740	0.099942
8	149	0.057718	1023	0.091119	732	0.090415	1001	0.099353
9	64	0.057618	794	0.090761	1120	0.089788	839	0.098693
10	1468	0.057595	1468	0.089183	1468	0.088741	1478	0.097365
11	150	0.057505	839	0.087562	1001	0.085205	1468	0.096991
12	301	0.057440	1013	0.086017	1437	0.083270	1469	0.093373
13	586	0.056891	776	0.083814	791	0.083016	673	0.090841
14	1469	0.055625	1486	0.082705	858	0.082981	794	0.088529
15	1122	0.055414	1014	0.082375	839	0.082880	732	0.084063
16	776	0.055333	1001	0.082046	861	0.082831	1023	0.083701
17	148	0.054751	1470	0.079746	1016	0.082021	1014	0.083121
18	1283	0.054659	1002	0.079721	860	0.080953	791	0.083013
19	740	0.054000	616	0.078429	270	0.080392	1120	0.082989
20	712	0.053724	817	0.076460	711	0.079724	858	0.082569
21	673	0.053722	738	0.076299	1469	0.078576	1013	0.082471
22	777	0.053695	1479	0.075673	712	0.078563	861	0.081994
23	714	0.053618	1438	0.074511	773	0.078317	862	0.081933
24	271	0.053441	64	0.074433	769	0.078037	777	0.080259
25	204	0.053046	739	0.073911	811	0.077536	1002	0.080107

Table 6.1: The first 25 most central nodes for the studied numerical cases.

If we look, for instance, at the results obtained for case 1 and case 4, where the network and data are assigned a greater importance, respectively, we observe that of the first ten most central nodes only 2 appear in both listings: nodes 634 and 1460.

Figure 6-4a offers useful information for the case $\alpha_1 = \alpha_2 = 0.2$. This choice of parameters means that we consider the node attribute data as less important than the network topology in both layers. In Figure 6-4b, the opposite case is shown, with $\alpha_1 = \alpha_2 = 0.8$. Now, we give much more importance to the total in- and out-flow as well as total bus connections associated to each node. We clearly see the difference in the maps. Specifically, in the lower figure the main roads of the city are clearly perceived, which is precisely where more public transport exists. These main roads, as well as train stations, contain nodes with a high centrality value.

In order to find whether there is a correlation between the ranking results in the four discussed cases, we compute the Spearman coefficient ρ, measuring the

statistical correlation between the rankings of the nodes. A positive value of ρ near +1 means a high association of ranks, while a value near 0 means no association between ranks.

This demonstrates the importance of choosing the α_i parameters in the model, giving more or less importance to the data than to the network itself.

6.7 Conclusion

In this Chapter, a measure of centrality for multiplex networks has been designed and evaluated with a real numerical example with the fundamental characteristic that both the connectivity of the graphs and a set of data present in each layer associated to the nodes are taken into account. The starting point is the original idea of the APA algorithm that introduces the influence of a set of data present in a network to the computation of the centrality of the nodes. The model solves the problem of the existence of isolated nodes in any of the layers by introducing a residual value for all nodes and representing the influence of the presence of data in the overall network. In addition, the proposed method introduces a variant with respect to the original alpha parameter related to the PageRank vector consisting of the choice of a different parameter for each of the layers. This difference in the value of α allows to take into account the importance assigned to the topology or to the data associated with the nodes in each of the layers. This allows for a flexibility that is demonstrated in the case study of the urban mobility OD and the urban bus network in the city of Rome. In that case study, a network with two layers is evaluated, where in the first layer a graph represents the OD car flows in the city, while the second layer represents local urban bus connectivity between city locations. The model solves the problem of the isolated nodes of the second layer and it allows to choose the importance of the node attribute data in each layer. Four different cases corresponding to different meaningful combinations of the α parameter are evaluated and visualised. The differences among the cases as visible from the most central city locations in each case show the advantage and utility of the proposed algorithm.

"We are losing the ability to understand anything that's even vaguely complex."

Chuck Klosterman, 2003

Chapter 7

APA for Multiplex urban networks

As discussed in Chapters 1 and 2, complex networks provide a framework for modelling real-world complex systems. Based on a set of data on mobility by car between different urban areas of the city of Rome described in Chapter 4, in this Chapter we represent and analyze these mobility data extended by urban public transport networks as additional network layers, augmenting the network nodes with data on commercial, economic, service and tourist activity in the city. In order to unravel the complex interdependencies of all these data, we propose a multiplex network consisting of four urban layers. Network centrality measures are then used to determine the most influential nodes or prominent areas of the city. In particular, we propose an adaptation of the APA centrality Algorithm for multiplex networks. This adaptation of the algorithm for multiplex networks offers the possibility to assign the importance given to node data relative to the network topology in each layer when computing the centrality. This allows a wider control in studying the mobility network, particularly generating different centrality maps according to the choice of this control parameter in each layer. We carry out experiments and present the results of a study of the network centralities considering different choices of the parameter.

This Chapter is a modified version of our paper Manuel Curado, Leandro Tortosa, Jose F Vicent, and Gevorg Yeghikyan. Understanding mobility in Rome by means of a multiplex network with data. *Applied Mathematics and Computation*, 2020 (under review)

7.1 Introduction

Detecting influential elements in complex systems is a crucially important task in several applied fields such as the identification of *influencer* people [147] in social networks, the spreading of a virus or fake news for detecting the original signal [95], [205], the detection of road traffic network dynamics [88], [87], [117], city growth [28], [281], global maritime flow [90], the study of the metropolitan rail transport network [65], [89] or urban congestion [242].

To quantify this influential information, many centrality measures have been proposed in literature, depending on the specific application. Such measures include degree, closeness or betweenness centrality. The concept of network centrality is exploited in the PageRank algorithm and its modifications [202][282][191][175] where the relevance of a node in a network is measured by how important the nodes linked to it are. This idea is essential in the ranking of nodes in a complex network [84][304][195].

A multidisciplinary part of complexity science is the complex network theory. It could be defined as the modelling of real systems through a graph (network) with non-trivial topological features [187]. A relevant characteristic of these networks is the impossibility of predicting the behaviour of the whole from the properties of its components. Intricate inter-dependencies and different kinds of interactions between the nodes of the complex network call for a more broader representation. A particular type of such a representation in the interconnected multilayer network is called multiplex, in which each network layer represents a different kind of relation among the shared nodes. Multilayer networks provide more flexible descriptions of nodes, edges and their interactions, generalizing single-layer networks. Multilayer data networks arise in a natural way as we observe complex systems in detail.

Different authors approach the representation of data into a multilayer network in different ways. Such a multiplex representation has proven useful in such real-world complex systems as urban systems, ecological information, financial data or mobility networks [48][149][62][126][176][183].

Much of the existing studies on multilayer networks focus on the basic question of

how to mathematically analyze a multilayer network effectively? On the one hand, we can see each layer as an isolated network, while on the other, we can add all the layers to form a single-layer network. Both methods are useful in some cases but are not suitable for all multilayer applications.

Several works have used a structure based on multiplex networks, but the main difference is that the topology adopted here relies on the connectivity between nodes representing the same entity in the different layers. There are many multiplex network works developing measures that allow to make comparisons between multiplex networks, and their single layer equivalents. Some of these studies propose different metrics to evaluate the importance of the nodes in the network with centrality rankings [240] such as eigenvector centrality [238] or random walk centrality [241]. In many applications, ranking nodes in multiplex networks requires highlighting the importance of nodes of each layer [240]. For instance, a version of the PageRank algorithm [119] measures the node centralities in these networks through an included bias in the jump between layers in a random walk [83]. However, a better model is required to correctly incorporate data intrinsic to the nodes. As mentioned in Chapter 5, the authors in [7], [10] proposed an Adapted PageRank Algorithm (APA) centrality for biplex networks for a better understanding of the relationships between nodes in different layers, and for measuring the importance of the nodes in each of these layers.

However, a biplex network is not enough in some complex problems as transport or mobility, where it is necessary to represent all information with more details. For this reason, in this Chapter, a multiplex network representing a complex model of the mobility in a metropolitan area (e.g., Rome) is proposed and analyzed. The analysis is based on the calculation of the importance of nodes given by the *APA Centrality Algorithm for multiplex networks*.

7.1.1 Motivation

In complex networks, one important line of research is how to integrate network science with time-series analysis [126]. That is to say, the study of the evolution and behaviour of networks over time. In this regard, the most central nodes of

the *OD* network as identified by the Adapted PageRank Algorithm (APA) [**?**] are an important point of focus. Some recent work has focused on analysing the spatial patterns of urban features [165, 220], studying urban networks with centrality measures [**?** **?**], as well as modelling the evolution of urban interaction networks over time [291]. However, there still seems to be a poor understanding of the interplay between urban location characteristics and the networks of interactions between these locations. All the more so, the temporal evolution of this interplay remains an unexplored area of research.

The study of origin-destination (OD) flow data is an important part for urban transport network management and strategic planning. More specifically, OD traffic matrices provide an estimate of the number of vehicles travelling between points in the city network over a given period of time [59]. That said, the objectives of this Chapter are to analyse the distribution of important nodes over time in urban OD networks. More precisely, we aim at studying the most central nodes of the urban multilayer network, as identified by the Adapted PageRank Algorithm (APA) ([7]).

The model presented in this Chapter allows to study and analyze several relationships of a set of nodes represented by different layers. In addition, it measures the influence of the data intrinsic to nodes in the different layers of the network. This is a crucial difference with respect to the classical multilayer approach. In our model, we adapt and generalize the APA algorithm to multilayers, and this leads to two main advantages: on the one hand, the advantages of the monoplex are exploited by associating a set of node data to each layer. On the other hand, we can play with the importance given to the data in each layer through a parameter. This allows to take into account both the the network topology is and the node data present in each layer. This feature is supported by the numerical results, showing the flexibility and versatility of the proposed model, allowing to work with different types of data - real and synthetic - and evaluate its importance within the network.

In Section 7.2 we review the centrality model based on the APA algorithm. Then, we discuss the case study of the urban OD network in Rome, augmented with commercial, tourist, economic and transport attributes in Section 7.3. Finally, we conclude the Chapter in Section 7.4

7.2 The centrality model

In [5], the authors propose a model to compute the centrality for attributed multiplex networks with the primary objective to classify the nodes in order of importance following the original PageRank vector concept used to establish a ranking in the Web graph.

They propose an algorithm for biplex networks that may be generalized to multiplex networks considering different relationships between nodes, but with the same set of nodes in each of the layers.

Let $\mathcal{M} = (\mathcal{N}, \mathcal{E}, \mathcal{S})$ be a multiplex network with $\mathcal{N} = \{1, 2, \ldots, n\}$ the set of nodes, $\mathcal{S} = (l_1, l_2, \ldots, l_k)$ layers and $A_1, A_2, \ldots, A_k$, the respective adjacency matrices. Further, let $\alpha_1, \alpha_2, \ldots, \alpha_k$ be the values of the parameter α for the k layers, respectively, and let $D_1, D_2, \ldots, D_k$ be the data matrices associated to nodes in the k layers. Finally, $\mathbf{v_{01}}, \mathbf{v_{02}}, \ldots, \mathbf{v_{0k}}$ represent the weighted vectors.

From all these parameters, the algorithm 7 computes the APA multiplex centrality for a multiplex network of k layers.

A detailed diagram of the steps of Algorithm 7 are shown in Figure 7-1 when considering a multiplex network with 4 layers.

The core of the process described in Figure 7-1 is the construction of the M_{MP} matrix, whose dominant eigenvector is the key to our calculation process, since it is the one that provides us with the classification vector of the nodes according to their importance. This matrix has a size of kn, where k is the number of layers in the network. It should be noted that when the number of layers is greater than 5 or 6, the matrix has a very large size, depending on the number of nodes. Undoubtedly, optimized numerical processes are necessary to reduce the calculation times of the centrality vector. The matrix M_{MP} is stochastic by columns, which ensures convergence and numerical stability essential in the calculation of the dominant eigenvector.

The equations that describe the construction of the 2×2 block matrix M_{MP} are the following:

APA Multiplex Centrality Algorithm. Input: A_i, α_i, D_i, $\mathbf{v_{0i}}$, for $i = 1, 2, \ldots, k$. Then,

1 For the layers l_i, for $i = 1, 2, \ldots, k$, construct the vectors and matrices:

(a) P_i, the probability matrices.

(b) vectors $\mathbf{d}_i, \mathbf{u}_i$ from the adjacency matrices A_i.

2 Compute P_i^*, for $i = 1, 2, \ldots, k$.

3 Compute the data vectors $\mathbf{v_i}$, for $i = 1, 2, \ldots, k$, as $\mathbf{v_i} = D_i \cdot \mathbf{v_{0i}}$.

4 Compute the coefficients $coef_i$, for $i = 1, 2, \ldots, k$.

5 From $\mathbf{v}_i$ and $coef_i$, for $i = 1, 2 \ldots, k$, compute $\mathbf{v}_i^{up}$.

6 Normalize $\mathbf{v}_i^{up}$, for $i = 1, 2, \ldots, k$ and denote it as $\{\mathbf{v}_i^{up}\}^*$.

7 Construct V_i, for $i = 1, 2, \ldots, k$.

8 Construct the matrix M_{MP}.

9 Compute the dominant eigenvector $\hat{\pi}_{\mathbf{MP}}$ of the matrix M_{MP}.

10 Compute the centrality $\mathbf{x}$.

Algorithm 7: APA Multiplex algorithm for multiplex networks.

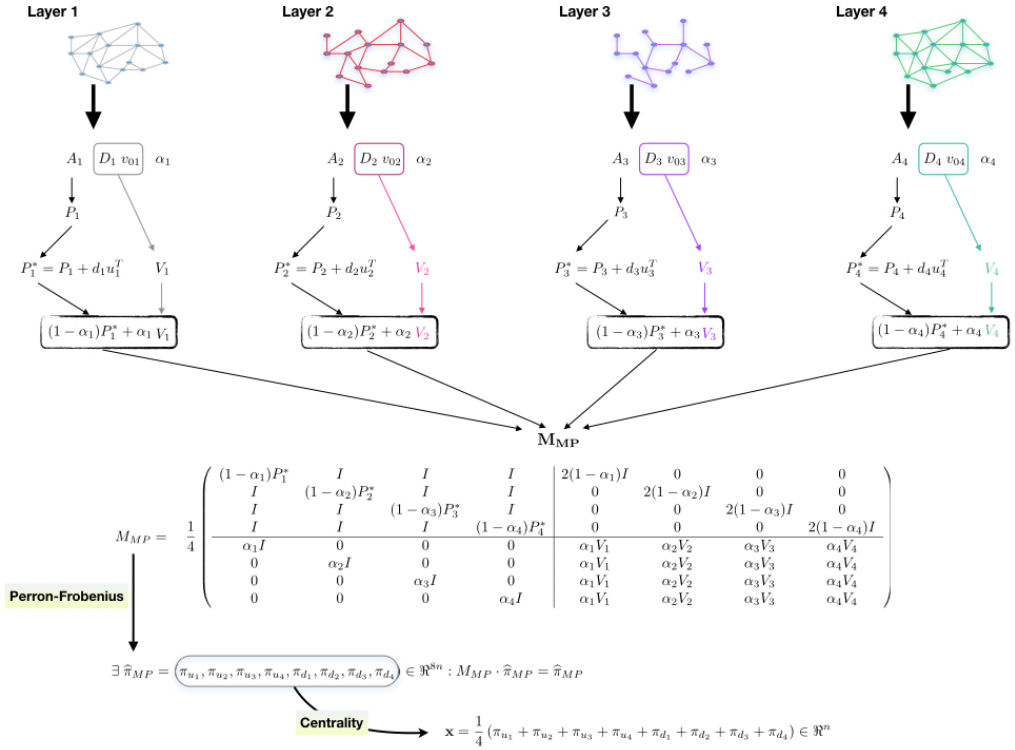


Figure 7-1: The APA centrality algorithm for a multiplex network with 4 layers.

$$M_{MP} = \frac{1}{k}\left(\begin{array}{c|c} M_{1,1} & M_{1,2} \\ \hline M_{2,1} & M_{2,2} \end{array}\right) \tag{7.1}$$

with

$$\mathrm{M}_{1,1} = \begin{pmatrix} (1-\alpha_1)P_1^* & I & \cdots & I \\ I & (1-\alpha_2)P_2^* & \cdots & I \\ \cdots & \cdots & \cdots & \cdots \\ I & I & \cdots & (1-\alpha_k)P_k^* \end{pmatrix}, \; M_{2,2} = \begin{pmatrix} \alpha_1 V_1 & \alpha_2 V_2 & \cdots & \alpha_k V_k \\ \alpha_1 V_1 & \alpha_2 V_2 & \cdots & \alpha_k V_k \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_1 V_1 & \alpha_2 V_2 & \cdots & \alpha_k V_k \end{pmatrix}.$$

Both $M_{1,2}$ and $M_{2,1}$ are block diagonal matrices and are given by

$$\mathrm{M}_{1,2} = \begin{pmatrix} 2(1-\alpha_1)I & 0 & \cdots & 0 \\ 0 & 2(1-\alpha_2)I & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 2(1-\alpha_k)I \end{pmatrix}, \; M_{2,1} = \begin{pmatrix} \alpha_1 I & 0 & \cdots & 0 \\ 0 & \alpha_2 I & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \alpha_k I \end{pmatrix}.$$

As we can see in Figure 7-1, each layer contains its own dataset D_i associated with

the nodes. In addition, the α parameter may be chosen in each of the layers; to recall, α is the parameter controlling the importance of data and network connectivity in the computation of the centrality. This implies that it is possible to assign different importance to the analyzed data in each layer.

On the other hand, we construct, for each layer, the matrices P_i^* which replace the probability matrices of the original algorithm. They are also defined from the probability matrices given by the node degrees. Its objective is that the isolated nodes in any of the graphs of the layers do not produce rows or columns of zeros in the M_{MP} matrix, which would lead to problems of numerical stability.

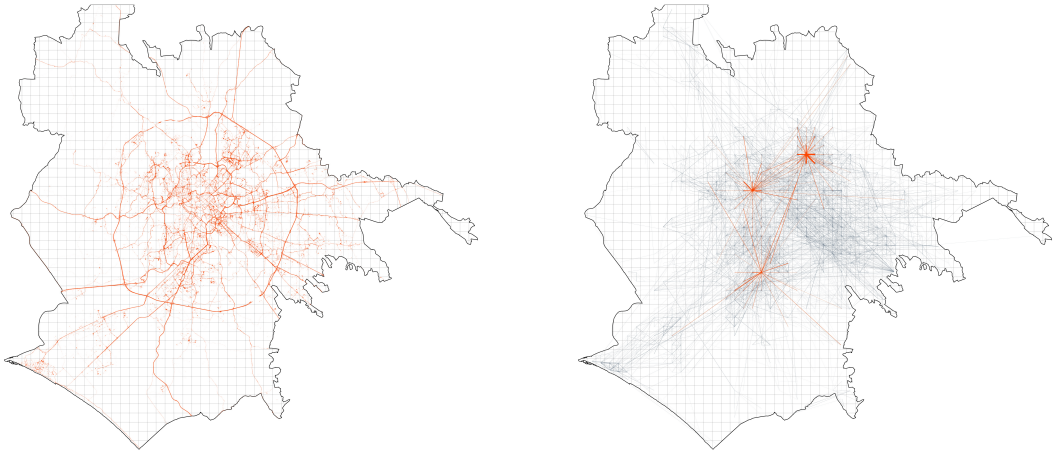
7.3 Multiplex Rome mobility network

In this section we present the case study of the city of Rome, where its mobility is analyzed together with a set of commercial, tourist, economic and cultural attributes associated with it, as described in detail in Chapter 4.

Multilayer networks offer the possibility of studying different relationships between the nodes of a graph, also including a set of data per layer associated with the nodes themselves. In addition, we have the opportunity to weight the data according to the importance that we give them within the general calculation of centrality. Taking advantage of this possibility, we present a numerical study on mobility in the city of Rome through the design and implementation of a four-layer multiplex network.

7.3.1 The dataset

Data about car flows and the public bus transport system will be used to analyse and determine the most central areas of the city when both data are studied (Figure 7-2).



(a) (b)

Figure 7-2: (a) Private car GPS trajectories superimposed on the grid in Rome (b) Rome OD network with some popular locations highlighted.

7.3.2 The construction of the multiplex network

Let us proceed with the construction of the4-layer multiplex network in the city of Rome. To better understand all the elements involved in the construction of the four-layer multiplex network that will be analyzed, we present Figure 7-3.

Below we describe in more detail each of the layers that make up the network. Since it is a multiplex network, the nodes are the same in each of the layers. What is different is the relationship between nodes. In all the layers the nodes are the geometric centre of the grid cells which the territory of Rome has been subdivided into using a Cartesian grid.

Layer 1 This layer represents the car flows. In other words, we represent the car displacements from a node to another node creating the OD network. Thus, an edge in the graph between nodes i and j means that there has been at least 1 car displacement between those nodes. The node data evaluated in this layer is the total number of displacements from each node. For instance, considering the node i, we associate to it the total number of displacements between itself and other nodes.

Layer 2 To construct the graph in this layer we introduce the public bus transport system. We also consider both car mobility and bus transport connections between nodes. Thus, an edge in the graph between nodes i and j means that

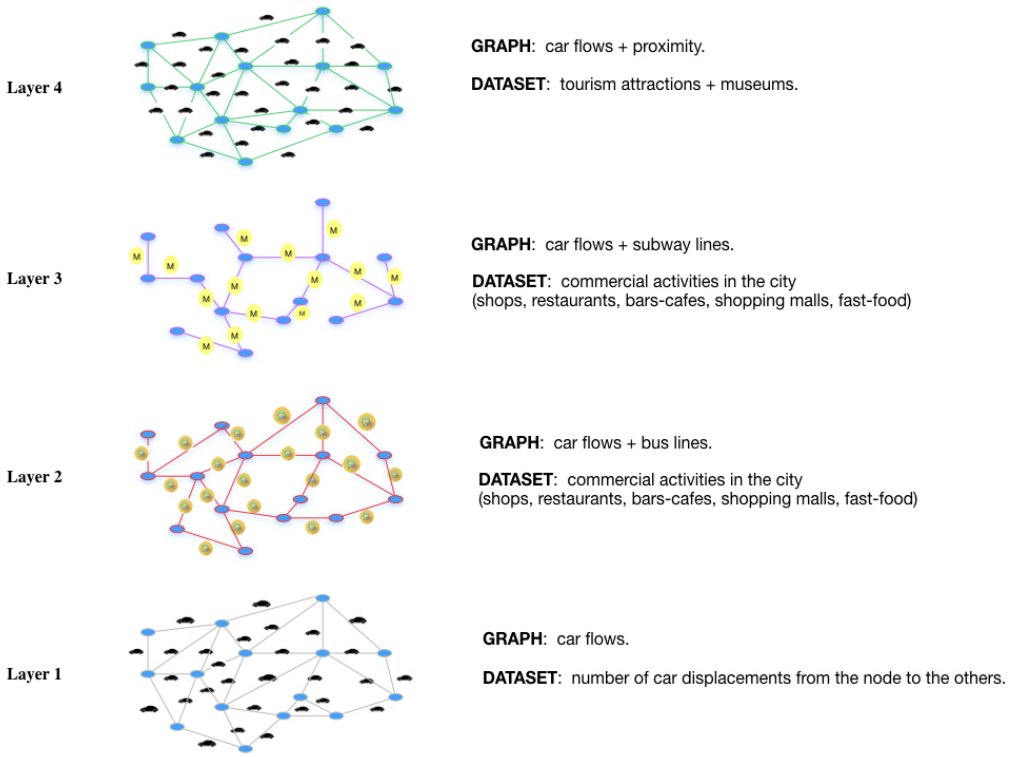


Figure 7-3: Rome multiplex mobility network with 4 layers.

there has been at least 1 car displacement between those nodes and, and, in addition, there is at least one bus line connecting the two nodes. The data associated with the nodes in this layer is related to the commercial activities in the city. More precisely, the number of shops, restaurants, bars-cafes, shopping malls, and fast-food has been counted in the area (cell) of each node and the aggregate count has been taken.

Layer 3 This layer takes into account, on the one hand, the trips by car between nodes and, on the other hand, the city's subway transport network. Thus, an edge in the graph between nodes i and j means that there has been at least 1 car displacement between those nodes and, also, there is at least one subway line connecting the two nodes. The data is similar to the one described in layer 2, that is, the commercial activities.

Layer 4 In this layer we construct an OD network taking into account the car displacements but with the restriction that the displacements must be *short*:

not exceeding a topological distance of 3 units. We now measure short mobility or even in the cell itself. The dataset chosen for this layer is the information on touristic attractions. Therefore, every node is associated with the number of museums and touristic attractions in the surrounding.

7.3.3 Numerical results

In this case study we are quantifying and analyzing the interconnected different variants of mobility, taking as a basis the mobility offered by car displacements in the different areas of the city. However, the richness of this approach is the possibility of mixing other variants of public transport such as the bus or the subway, considering different relationships between nodes.

In addition, we have the ability to evaluate various types of data in each layer. In our case, we have focused on information related to commercial and tourist activity in the city. This adds a framework of complexity to the analysis that is impossible in a single layer network approach. We must not forget the final objective, which is the classification of the nodes in order of importance within the network, taking into account the evaluated parameters.

Another characteristic of this model is the possibility of differentiating the importance assigned to the data in each of the layers. In the definition of the matrix M_{MP} given by the expression 7.1 each of the blocks has its own parameter α_i. This allows to consider giving more importance to the data associated to the nodes in all or any of the layers or, on the contrary, giving more value to the network topology in all or any of the layers.

Different numerical experiments have been conducted taking different values of the parameter α_i for the different layers. Some example cases are presented below.

Case 1: $\alpha_1 = 0.1, \alpha_2 = 0.8, \alpha_3 = 0.8, \alpha_4 = 0.8$ In the first analyzed case, the highest importance is assigned to the network and its connectivity in layer 1, while in the other layers the highest importance is assigned to the evaluated data associated with each node. This means that we are mainly establishing relationships between car mobility with commercial and tourist activity that

takes place in the city. While considering mobility, we do not give excessive importance to public transport networks, but rather to the private car flows.

We run Algorithm 7 to obtain the multiplex centrality of all the nodes according to the model described in this Chapter.

Case 2: $\alpha_1 = 0.2, \alpha_2 = 0.2, \alpha_3 = 0.2, \alpha_4 = 0.2$ In this case, we choose the same value of the parameter for all layers, that is, $\alpha_i = 0.2$, for $i = 1, 2, 3, 4$. This choice of α_i in each layer means that we give the same value to the data in all layers. The value chosen is small, only 0.2, which indicates that we give the network topology much more importance than the weight of the data at the nodes.

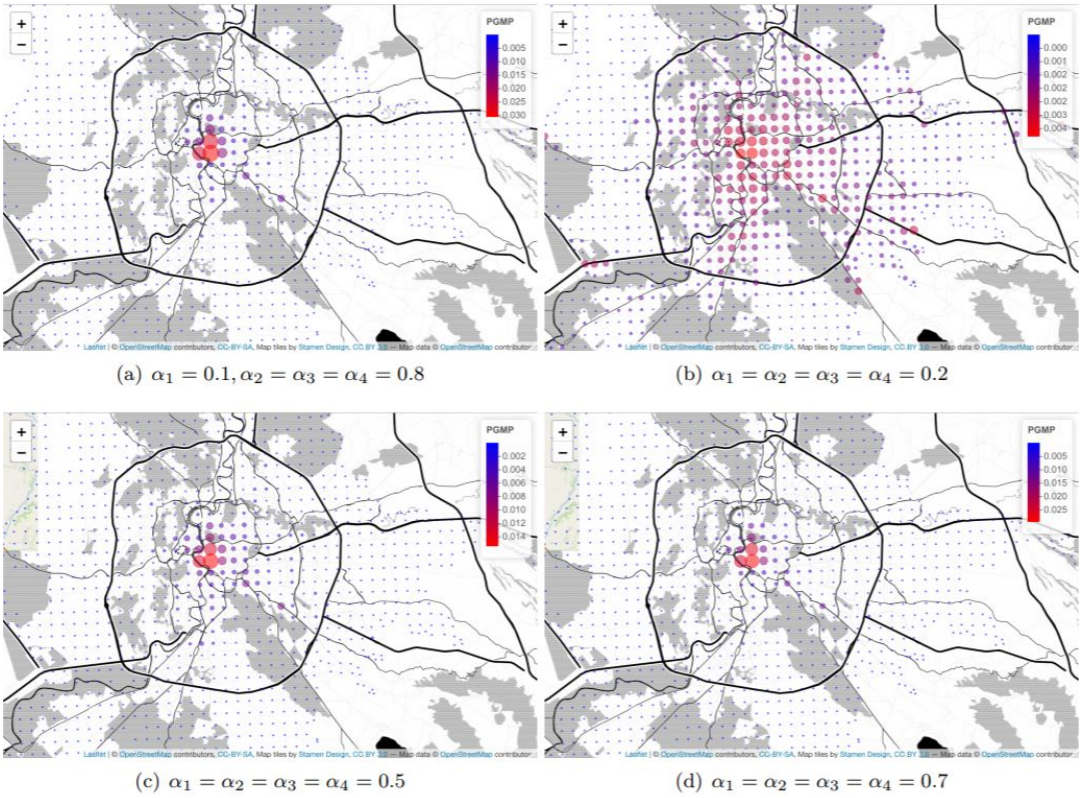
Case 3: $\alpha_1 = 0.5, \alpha_2 = 0.5, \alpha_3 = 0.5, \alpha_4 = 0.5$ In this case, as before, we choose the same value of the parameter for all layers, that is, $\alpha_i = 0.5$, for $i = 1, 2, 3, 4$. Note that now the situation is different because we are assigning the same value to each layer, but giving the same importance in the calculation to the data as to the network topology.

Case 4: $\alpha_1 = 0.7, \alpha_2 = 0.7, \alpha_3 = 0.7, \alpha_4 = 0.7$ In this last case, we assign much greater importance to the amount of the data than to the network. We are essentially interested in measuring centrality based on the data evaluated through the mode data in the different network layers.

The centrality algorithm has been executed for these four cases and the results are shown below.

In Figure 8-3 we show a map of Rome with the computed multiplex centrality of nodes for all the cases studied, for different values of the parameter α.

In the image representing case 1 (Figure 8-3(a)) it can be seen that there are a few nodes with very high centrality in the historical and tourist center of the city. This is a consequence of giving maximum importance to the data in layers 2, 3 and 4. These data summarize the commercial and recreational activity (bars, restaurants, shops, etc.) in layers 2 and 3 and touristic activity in layer 4. You can clearly identify the area where there is a greater probability of finding tourists in the city, for instance.



(a) $\alpha_1 = 0.1, \alpha_2 = \alpha_3 = \alpha_4 = 0.8$ (b) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.2$

(c) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.5$ (d) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.7$

Figure 7-4: Multiplex centrality (PGMP) for all the cases analyzed.

The difference in the image that we have represented of case 2 is worth noting. Now the importance in the nodes is not given to the data but to the network topology. The study is now based on the importance of the connections of each node with the rest. As the connectivity is determined by urban transport networks and vehicle mobility, the behaviour of the centrality distribution is now quite different. A much more uniform distribution of the centralities is observed, clearly showing the influence of the urban and metro transport networks.

In the lower figures (c) and (d), that corresponding to cases 3 and 4, we return to a pattern similar to that of case 1, characterized by the existence of a few nodes with very high centrality in the historic center of the city. The closest similarity in terms of results is found between cases 1 and 4, although in case 4 the differences in centralities are not as noticeable as in case 1 in which the *hubs* with high centrality are more discernible. In sum, the greatest differences are given when the importance is focused on the network or on the data (cases 1 and 2, respectively).

In Figure 8-4 we have graphically represented the centrality values for all the

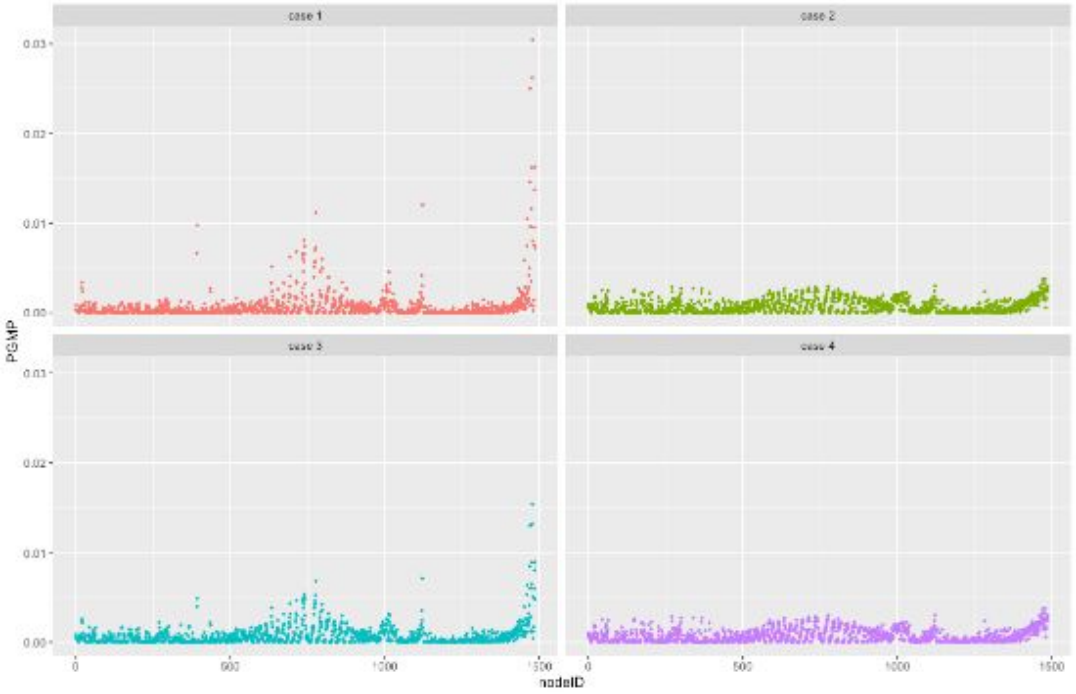
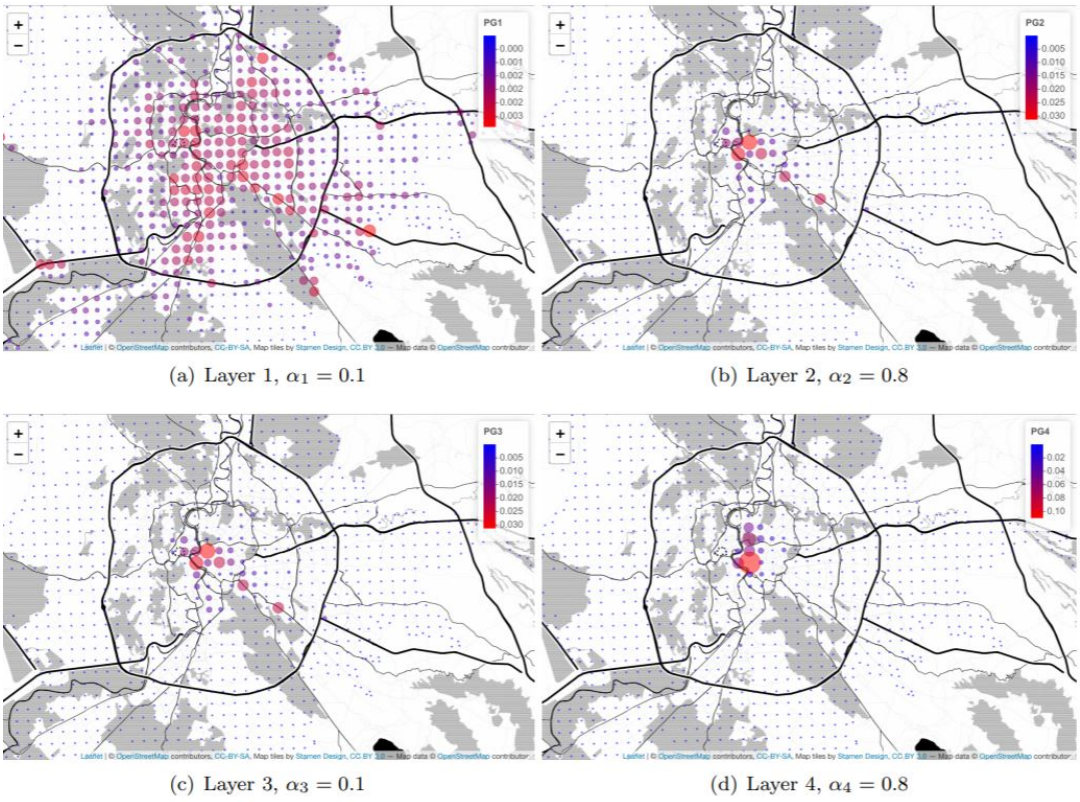


Figure 7-5: The multiplex centrality distribution for the cases studied.

nodes of the graph, differentiating the four cases that we have studied by varying the parameter α. It is important to highlight the small dispersion of the centrality values, which is especially clear in cases 2 and 4, where there are hardly any values beyond the third quartile. In cases 1 and 3, which display great similarity in terms of the distribution of degrees, one can observe a few nodes that trigger the value of centrality and become small hubs that absorb much of the importance within the network.

A multiplex network with four layers has been constructed where the relationships between the nodes have been established according to various types of public transport and mobility in vehicles through the urban traffic network. In addition, data on commercial and restaurant activities, as well as tourist interest, have been used to determine the importance of the different nodes. The study of centrality in a multiplex network with four layers is not the same as if the four layers are considered as four individual monoplex networks without any interaction between them. In this regard, we determine the centrality of the layers individually to compare them with the multiplex centrality. The individual centrality is calculated using the APA centrality algorithm for networks with node data (see [5]).



(a) Layer 1, $\alpha_1 = 0.1$ (b) Layer 2, $\alpha_2 = 0.8$

(c) Layer 3, $\alpha_3 = 0.1$ (d) Layer 4, $\alpha_4 = 0.8$

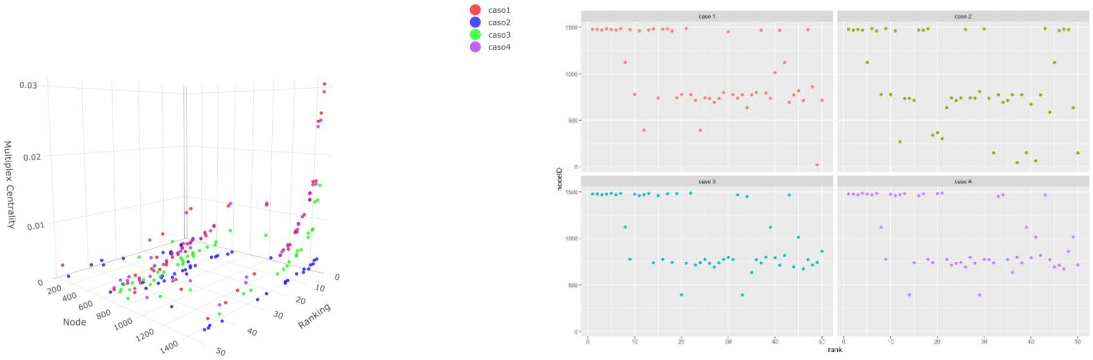
Figure 7-6: APA centrality for graphs in layers $1, 2, 3, 4$.

Let us take case 1, where $\alpha_{=}0.1, \alpha_2 = \alpha_3 = \alpha_4 = 0.8$.

Figure 8-8 shows four maps of Rome with the node centralities calculated individually in each layer, without taking into account the relationships between the layers. The centrality in layer 1 (a) presents the greatest spatial dispersion, as expected. This is the centrality of car displacements throughout the urban area of the city, with many displacements between the periphery and the rest of the areas, not only in the center. There exist no hubs of centrality in the historic and touristic center since within these central areas full of tourists and visitors people move massively in public urban transport, bicycle or simply walk.

The centralities of layers 2 and 3, PG1 and PG2, respectively in images (b) and (c) are very similar. This is because the data set used in both layers is identical: the commercial and restaurant activity associated with the city in the form of shops, bars, restaurants, and others. In addition, we are giving the utmost importance to data, so the role of the graph and its connectivity is not so relevant. The centrality observed in layer 4 (PG4) is based on the maximum importance of the data related

to the city's touristic attractions and museums. This causes the existence of a large central node located in the most touristic part of the city where all the main monuments are located next to the Vatican City. The only node that has a high centrality that is outside the historic city center is located in the Cinecitta World theme park.



(a) (b)

Figure 7-7: The 50 most important nodes of all the cases analyzed.

Since the distribution of centrality densities in all cases is quite monotonous, it may be interesting to work with the most important 50 nodes in each of the cases analyzed. For this task, a table has been constructed with this ranking in all the cases, taking as a reference the multiplex centrality (PGMP). The results can be seen in Figure 8-6. On the left side a 3D graph is shown where on the X axis we have the ranking from 1 to 50, on the Y axis we have the node and on the Z axis the corresponding centrality value is shown. The plot shows the centrality values for the four cases. In the right part (b) the centrality values are plotted against the ranking, making a graphic representation for each of the cases.

The 3D graph shows the great similarity between the data from case 1 and case 4, where the correlation of the most important nodes is very high. The case 2 is slightly different; if we analyze its 3D graph compared to other cases, the values of the centrality of all the nodes are very small and not even the most central nodes compare to that of the others. Homogeneity is the essential characteristic. We notice this in the plot (b) where the point cloud is a little different from the rest.

7.4 Conclusion

An algorithm for measuring the centrality in multiplex networks is used to determine the most important nodes in the metropolitan area of the city of Rome evaluating a data set obtained by aggregating origin-destination (OD) flows of private cars augmented with the bus and metro public transport system. A four-layer multiplex network is constructed using this mobility data set together with some attribute data associated to the nodes related to the commercial and tourist activity that takes place in the city.

The proposed algorithm constitutes a generalization of the APA algorithm for networks with data extended to multiplex networks, the main advantage of which is the possibility to choose for each layer the importance given to the network connectivity and to the node data present in the network. This task is done by the well-known parameter α in the PageRank model.

This work explores this possibility by analyzing the centrality of the network in four different cases, depending on the importance given to the data in the computation of centrality. When we give great importance to the commercial and tourist activity data, the areas of the historic city center are clearly marked, while when evaluating mobility in private vehicles, the centrality extends in a much more generalized way to a wider area, highlighting important nodes in some fundamental communication pathways.

The discussed example shows the possibilities offered by this study interrelating the nodes by layers with different data sets. The comparison carried out with respect to the centrality measurements calculated individually in each layer, without taking into account the relationships with other layers, show us the differences in centralities when calculated in multilayer networks and why the complex interdependence in urban networks is better represented with a multiplex model.

"Space is killed... and we are left
with time alone."

Heinrich Heine, 1843

Chapter 8

Spatio-temporal APA centrality

In the wake of the mobility challenges cities are facing worldwide, understanding the complex interactions between urban mobility patterns and the socio-economic activities in cities is of crucial importance for urban planning and policy making. Drawing on the recent advances in complex network theory, the mobility flow patterns, typically encoded as origin-destination (OD) matrices, can be represented as weighted directed graphs, with nodes denoting city locations and weighted edges the number of trips between them. Such a graph can further be augmented by node attributes denoting the various socio-economic characteristics at a particular location in the city, as described in Chapter 4. In this Chapter, we propose a generic workflow to study the spatio-temporal characteristics of "hotspots" of different types of socio-economic activities as characterised by the attribute-augmented network centrality measures introduced in Chapter 5. We apply these centrality measures to the urban OD networks in Rome and London to demonstrate the proposed workflow. Our results show structural similarities and distinctions between the spatial patterns of different types of human activity in the two cities. Our approach offers a workflow yielding simple indicators thus opening up opportunities for urban practitioners to develop tools for real-time monitoring and visualisation of interactions between mobility and economic activity in cities.

This Chapter is a modified version of our paper Gevorg Yeghikyan, Leandro Tortosa, Jose F Vicent, and Mirco Nanni. Ranking places in attributed temporal urban mobility networks. *PloS one*, 2020 (under review).

8.1 Introduction

The ever-growing availability of large scale data sources pertaining to human activities in contemporary cities and the fact that the socio-economic and technological systems lend themselves adequately to representation through discrete elements and interactions between them have led recent years to witness an unprecedented increase in modelling of such complex systems using network theory [25].

In urban science, there has been a significant research interest towards understanding urban systems particularly through modelling road structures, human mobility, traffic flow, and economic activity through a complex networks approach [26, 40, 55]. In such a setting, distinct elements in a city such as road junctions or neighbourhoods are typically represented as the network nodes, while the heterogeneous connections or interactions between them, such as road segments, passenger flows, activity correlations represent the edges in the network [218, 163]. Further, depending on the focus of the research, various statistical and graph-theoretical properties of the network can be studied to gain valuable insights about the urban spatial, temporal and socio-economic structures. Following this approach, several studies have analysed mobile phone usage, taxi or private car GPS trajectories, smart card, geo-located social media, and classical census data for inferring systemic patterns both at the individual and aggregate level [53, 227, 21, 45, 306].

An area of research of particular interest in complex network theory is the study of the importance of nodes or edges in a network through centrality measures. Such measures are typically based on local and global network connectivity structures and include a variety of types: degree [105], closeness [33], betweenness [104], eigenvector [189], PageRank [211], etc. However, these conventional centrality metrics measure the importance of nodes by considering only the network topology regardless of the intrinsic information on these nodes such as their behaviour, type or some other, domain-specific attribute. Since many kinds of real-world networks call for such node attributes, several centrality measures have recently been proposed extending the widely used centrality measures to accommodate node attributes [6, 10, 34]. This becomes especially relevant in urban modelling, as locations in a city possess

quantitative and qualitative characteristics irrespective of the connectivity structure of the network of interactions with other locations. Such characteristics may describe the availability and quantity of such urban features as parking lots, restaurants, real estate prices, population density, etc., qualitatively enhancing urban networks.

Another important line of research in complex networks is temporal network theory: the study of the evolution and behaviour of networks over time. Temporal networks integrate network science with time-series analysis and contribute greatly to the modelling of epidemic spreading, transportation optimization, biological systems, as well as social networks [127].

Although some recent work has focused on analysing the spatial patterns of different urban features [163, 221], studying urban networks with centrality measures [6, 217, 229], as well as modelling the evolution of urban interaction networks over time [293], we still have a poor understanding of the interplay between urban location characteristics and the networks of interactions between these locations. All the more so, the temporal evolution of this interplay remains an unexplored area of research.

Having this gap as motivation, the objectives of this Chapter are to analyse and study the spatial distribution of the central nodes by activity type over time in urban origin-destination (OD) networks. More specifically, we focus on the spatial arrangement of the most central nodes of the OD network as identified by the Adapted PageRank Algorithm (APA) [6] additionally considering activity related to food and retail services over time in Rome and London. We find that although the daily temporal patterns of the most central places in attributed OD flows in the two cities display structural similarity, the spatial distributions of food and retail related activity over time differ, indicating a more polycentric structure in London. The proposed pipeline from raw GPS and open source point-of-interest (PoI) data to the resulting data visualization offers a workflow with the potential for creating tools for monitoring the changes in mobility patterns and in their relations to various socio-economic activities over time. This would allow urban practitioners to monitor daily/weekly mobility patterns for analysing the effects of urban interventions or temporary events, but also to study long-term trends in these patterns for urban

policy making.

To achieve the objectives the structure of this Chapter is as follows: the theoretical tools employing graph theoretical methods for characterising centrality (Adapted PageRank Algorithm), a measure of statistical heterogeneity (the Gini coefficient) for describing the distribution of the obtained centrality values, a non-parametric technique for identifying "hotspots" of high centrality values, and a spreading index characterising the spatial spread of the "hotspots" in the two cities are presented in Section 2. Section 3 describes the dataset used for the proposed study and summarises the methodology underpinning the experiments. The proposed methodology is validated and the numerical results from studying real urban OD networks in London and Rome are discussed in Section 4. Finally, Section 5 concludes the Chapter.

8.1.1 Related work

The city is one of the most complex dynamic anthropogenic systems. To analyse this complexity, spatial networks have been widely used for modelling city objects and the interactions between them, and different approaches have been proposed with regards to the choice of objects and the various types of interactions between them to denote with nodes and edges, respectively [217, 216, 25]. In modelling cities with these simple mathematical objects called graphs, a variety of properties such as the relative importance of city locations through network centrality measures can further be studied.

Network centrality measures have been widely used in different problem settings across many research fields related to economic geography [51], road networks [217], and urban mobility [229]. [51], for instance, study the impact of social network structures exemplified by central nodes computed with the PageRank algorithm in the US startup mobility networks on the innovation performance of cities.

In studying street networks, [217, 217], for instance, analyse the distributions of various types of centrality measures computed on the street networks of different cities and find them to reveal the distinction between self-organized and planned cities. Another work [146] utilises betweenness centrality measures in street networks across cities worldwide to find universal bimodal betweenness regimes corresponding

to trees and loops explaining high and low centrality values, respectively. Similarly, conventional centrality measures have also been used in studying human mobility, particularly on inter- and intra-urban OD networks. In particular, [85] reveal node betweenness centrality in an inter-urban OD network displaying a positive correlation with population and wealth, while [229] study the statistical properties of betweenness centrality in intra-urban OD networks in different cities.

As mentioned in the previous section, conventional centrality measures suffer from the drawback of not taking account of exogenous information on the nodes. In this regard, there exist studies that have attempted to overcome this by extending centrality measures to include node attributes. [6, 10] propose an augmented centrality measure (APA) based on the PageRank centrality to study the key areas of city activity on the street network enriched with geo-referenced retail and services data on the nodes. [275] take a different approach, introducing distance decay and attractiveness modifications to the PageRank algorithm to incorporate the effects of distance and attractiveness in choosing a particular destination over another.

As we have seen, computing measures of network centrality gives us the relative importance of the nodes (locations) in an urban network. However, choosing the *most* important locations requires some discussion. In the field of spatial analysis, a "hotspot" usually refers to a location with an attribute value relatively higher than that of its neighbouring locations. The study of the spatial characteristics of "hotspots" has been the focus of research in such different fields as criminology [192], transportation [287], or epidemiology [256]. In the context of urban mobility, "hotspots" may be seen to reflect travel intensity between different areas [301, 299]. With the availability of large data streams of ever more granular location data, "hotspot" analysis is becoming a widely practiced tool in urban mobility research [12, 177].

Among many techniques for "hotspot" detection, there are two most commonly used techniques. The first is based on spatial statistical analysis, particularly on spatial autocorrelation indicators for detecting neighbouring areas with dissimilar value intensities [19]. The second "hotspot" detection method is based on kernel density estimation by using a spatial search method [128]. In [169], the authors have

applied this method to study the spatial distribution of popular locations.

So far, we have discussed static networks as the object of study with tools from network theory. However, since many real-world phenomena require modelling their behaviour over time, temporal network theory has become a valuable tool in many fields. This is the case with urban mobility which demonstrates important temporal patterns, the study of which could greatly inform urban planning, policy making, and management. A number of studies has attempted to analyze urban mobility from a temporal perspective. [272], for instance, use centrality measures for temporal prediction on OD networks built from cellular traffic data. [293], study temporal OD networks with change detection techniques for identifying "change points" in time, in which the entire structure of the graph changes.

There have also been recent applications of graph neural networks on temporal sequences of graphs, mostly in a prediction setting. For instance, based on the length of prediction windows, previous studies of traffic forecast can be divided in dynamical modelling [263] based on mathematical tools and physical knowledge, and data-driven methods [135, 69] based on classical statistical and machine learning.

8.2 Previous work

In this section, the centrality measure applied to rank the attributed nodes in the OD networks, statistical dispersion measures describing the centrality value distributions, as well as a measure of spatial spread are presented in detail.

8.2.1 The Adapted PageRank algorithm (APA)

The *PageRank* model [202] was proposed to compute a ranking for every Web page based on the graph of the Web. The objective of the model is the calculation of a vector, called *PageRank vector*, which establishes a ranking of all the pages analyzed according to their importance.

The PageRank vector is the dominant eigenvector of the matrix known as *Google matrix* G' (see [211] for an algebraic definition and characteristics of this matrix). Among its spectral features, G' is stochastic and positive, so it can be directly

applied the Perron-Frobenius theorem to assure the existence and uniqueness of the PageRank vector $\mathbf{x}$. To delve into the characteristics of the PageRank model, see [37, 43].

In 2012, [6] proposed an adaptation of the original PageRank model called Adapted PageRank Algorithm (APA) for spatial networks with data, although the original algorithm was initially thought for urban street networks. Afterwards, the APA model was modified introducing small variants [10]. The base of the APA model is, following the core of the original PageRank, the construction of an stochastic and positive matrix M_{APA} that keeps the excellent spectral properties of the Google matrix. From this new matrix, it is possible to compute a unique eigenvector that constitutes the classification of the nodes according to their importance in the network.

As the Google matrix had two terms, one related to the node's connections and the other related to the probability of surfing among the pages, the matrix M_{APA} has two terms, the first related to the connectivity and the second term related to the data associated to every node. So, a data matrix D of size $n \times k$ is constructed where the rows are the nodes and the columns are the attributes of the node's information object of the analysis.

Therefore, M_{APA} is constructed from the adjacency matrix A and the data matrix D as

$$M_{APA} = (1 - \alpha)P + \alpha V, \tag{8.1}$$

where P is the probability matrix computed from the adjacency matrix, and V is a matrix that collects the whole data associated to the nodes. Regarding the probability matrix P, it is constructed from the adjacency matrix A, as

$$p_{ij} = \begin{cases} \frac{1}{c_j} & \text{if } a_{ij} \neq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad 1 \leq i, j \leq n, \tag{8.2}$$

where c_j represents the sum of the j-th column of the adjacency matrix.

Remark that P has the following characteristics: it is nonnegative and stochastic by columns (see [6] to know more details about the spectral properties of P).

The APA algorithm proposed by the authors can be summarized as:

Input: Let $G = (V, E)$ be a primary graph representing a network with n nodes.

Output: $\mathbf{x}$ representing the network centrality

begin

Compute the matrix P from the graph G according to (8.2)

Construct the data matrix D

Construct the weighted vector $\mathbf{v}_0$

Compute $\mathbf{v}$ as $D\mathbf{v}_0 = \mathbf{v}$

Normalize $\mathbf{v}$, and denote it as $\mathbf{v}^*$

Construct V as $V = \mathbf{v}^*\mathbf{e}^T$

Construct the matrix M_{APA} following the expression (8.1)

Compute the eigenvector $\mathbf{x}$ of the matrix M_{APA} associated to eigenvalue $\lambda = 1$. The components of the resulting eigenvector $\mathbf{x}$ represent the ranking of the nodes in the graph G

end

Algorithm 8: APA algorithm for attributed networks.

Vector $\mathbf{x}$ constitutes the Adapted PageRank vector and provides a classification or ranking of the network nodes according to both the connectivity and the presence of data.

8.2.2 Gini coefficients

After computing the node rankings with the APA centrality for each activity type for each hour of the day, we need measures of heterogeneity to assess their distributions in time and space.

The first type of measure commonly used to assess how heterogeneous a variable is distributed, is the Gini coefficient, borrowed from economics. It is defined as

$$GI = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |\mathbf{x}_i - \mathbf{x}_j|}{2n^2 \bar{\mathbf{x}}}, \tag{8.3}$$

where $\mathbf{x}_i$ is the APA value at location $i = [1, 2, \ldots, n]$ and $\bar{\mathbf{x}} = (1/n) \sum_i \mathbf{x}_i$.

The Gini coefficient, originally used to measure wealth and income inequality, can be applied to quantify the heterogeneity of other variables as well. In the case of characterising heterogeneity of values at different locations in a city, the Gini coefficient will take on the value of zero if the variable of interest is distributed uniformly across city locations. Conversely, it takes on its maximum value when all of the variables of interest are concentrated in a single location, leading to a Gini coefficient of $GI = 1 - 1/n$, which is very close to 1 for large n.

However, being a measure of statistical dispersion, the Gini coefficient is agnostic to the spatial arrangement of the APA values in the city. As demonstrated in [223] and [264], a reshuffling of the spatial configuration can yield the exact same Gini coefficient.

In order to obtain a Gini coefficient that carries meaningful spatial information, we further use the Spatial Gini index proposed in [223]. In essence, it is a decomposition of the classical Gini with the aim of considering the joint effects of inequality and spatial autocorrelation. More specifically, it exploits the fact that the sum of all pairwise differences can be decomposed into sums of geographical neighbors and non-neighbours:

$$GI = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}^{A} \left|\mathbf{x}_i - \mathbf{x}_j\right|}{2n^2\bar{\mathbf{x}}} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \left(1 - w_{i,j}^{A}\right) \left|\mathbf{x}_i - \mathbf{x}_j\right|}{2n^2\bar{\mathbf{x}}}, \tag{8.4}$$

where $w_{i,j}^{A}$ is an element of the binary spatial adjacency matrix.

The Spatial Gini index can be interpreted as follows: as the positive spatial autocorrelation increases, the second term in equation 8.4 increases relative to the first, since geographically adjacent values will tend to take on similar values. On the contrary, negative spatial autocorrelation will cause an opposite decomposition, since the difference between non-neighbours will tend to be less than that between geographical neighbours. In either case, this offers the possibility to quantify the relative contributions of these two terms. The results obtained from this approach can further be tested for statistical significance by using random spatial permutations to obtain a sampling distribution under the null hypothesis that the APA variate is randomly distributed in space.

8.2.3 Spreading index

Despite their informative relevance, the Gini coefficient and its spatial variant exploit the mean $\bar{\mathbf{x}}$, which, under fat-tailed distributions, as many socio-economic variables tend to be, may be undefined. In such cases, as shown in [97], the Gini coefficient cannot be reliably estimated with non-parametric methods and will result in a downward bias emerging under fat tails.

Another downside of measuring heterogeneity of the obtained APA values with the Gini approach is that it does not offer the possibility to study the spatial arrangement of the "hotspots" - locations with very large APA values. The "hotspots" are defined as the grid cells with an APA value above a certain threshold $\mathbf{x}^*$ (see Figure 8-3). For choosing this threshold we resort to a non-parametric method introduced in [172]. Once we have identified the "hotspots" as cells with APA values larger than the chosen threshold $\mathbf{x}^*$, we can use the *spreading index* introduced in [264] for measuring the average distance between the "hotspots", normalized by the average city distance to enable cross-city comparisons:

$$\eta\left(\mathbf{x}^*\right) = \frac{\frac{1}{N(\mathbf{x}^*)} \sum_{i,j} d(i,j) \mathbf{1}_{(\mathbf{x}_i > \mathbf{x}^*)} \mathbf{1}_{(\mathbf{x}_j > \mathbf{x}^*)}}{\frac{1}{N} \sum_{i,j} d(i,j)}, \tag{8.5}$$

where $N(\mathbf{x}^*)$ is the number of pairwise distances of grid cells with an APA value greater than $\mathbf{x}^*$, N is the number of pairwise distances between all grid cells covering the city, $d(i,j)$ is the distance between cell i and cell j, and $\mathbf{1}_{(\mathbf{x}_i > \mathbf{x}^*)}$ is the indicator function for identifying the cells with APA values greater than $\mathbf{x}^*$ for computing the distances. The *spreading index* is essentially the average distance between cells with $\mathbf{x}_i > \mathbf{x}^*$, normalized by the average distance between all city cells. If the cells with large APA values are spread around across the city, this ratio will be large. Conversely, if the high value cells are concentrated close to each other, as in a monocentric city, this ratio will be small.

Now, we present a synoptic description of the overall workflow.

After the city territories have been tessellated into 1x1km grid cells, the raw GPS data has been processed, trip origins and destinations have been extracted and the OD networks have been built for each hour of the day both in London

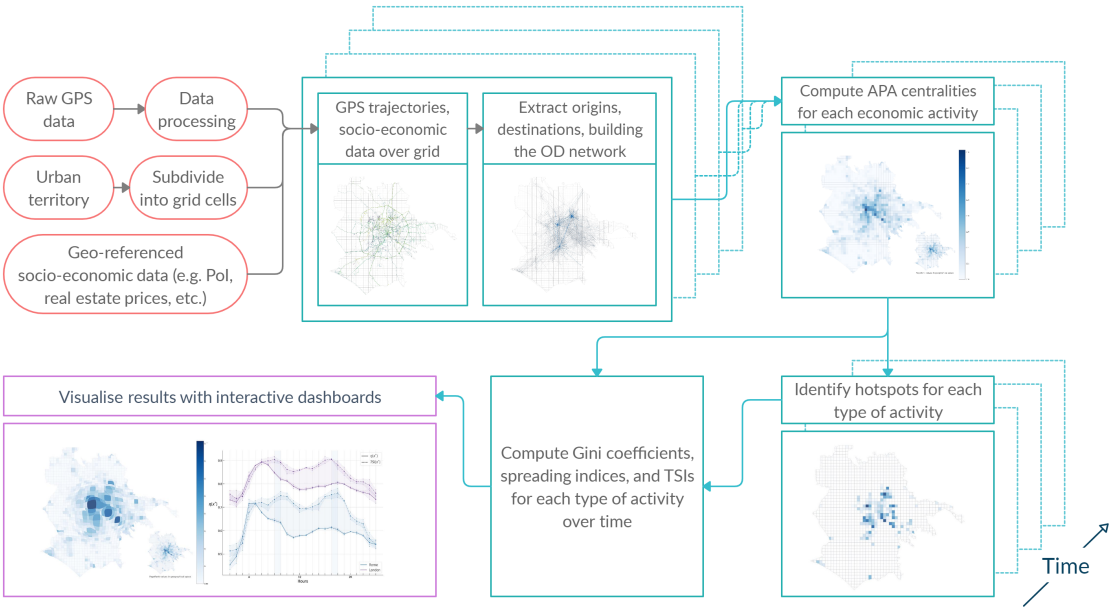


Figure 8-1: Workflow flowchart from raw data input to analysis and visualisation

and Rome, we proceed to computing the location centralities with the Adapted PageRank Algorithm. Then, we analyze the heterogeneity of the APA values in both cities during a typical day and during a typical week by using the Gini coefficient. Finally, in order to obtain a clearer picture of the spatial distribution of the APA values, we calculate the *spreading index* and its modification introduced in Section 8.3.5.

The methodology can be summarized in Figure 8-1.

8.3 Numerical results

In this section we conduct the numerical experiments for the study and outline the principal findings. We then undertake a detailed discussion of the results in the forthcoming section.

8.3.1 Computing the APA centrality

We proceed to compute the APA values using Algorithm 8 for the following three kinds of networks:

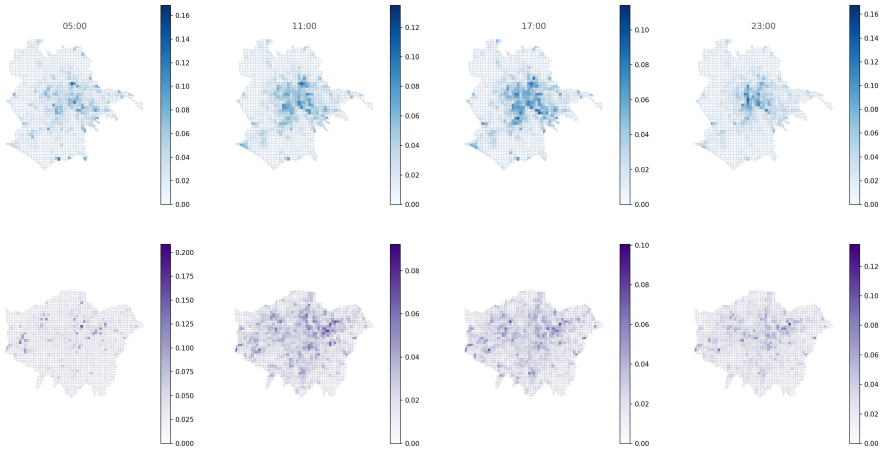


Figure 8-2: The APA values for the mobility flow network in Rome (up row) and London (down row) at different times of the day.

1. Mobility flow network only.

2. Flow network with nodes attributed with information related to retail (number of shops, shopping malls, retail stores).

3. Flow network with nodes attributed with information related to food services (number of bars, restaurants, cafes).

The APA values of the Rome and London grid cells at different times of the day can be seen in Figure 8-2. In this figure, the values of the APA centrality of each of the nodes with respect to the mobility flows have been calculated using Algorithm 2.1 and have been quantitatively represented. In the upper row the most central nodes in the city of Rome are clearly shown, at different times of the day; in the lower row the same calculations made in London are shown. Without delving into details, for now, a greater concentration of the most important nodes in the city of Rome is observed for all the chosen times, while in London the most central nodes are in much more dispersed locations. Precisely the study of this dispersion and the characteristics associated with the distribution of centrality values will be one of the axes of this work.

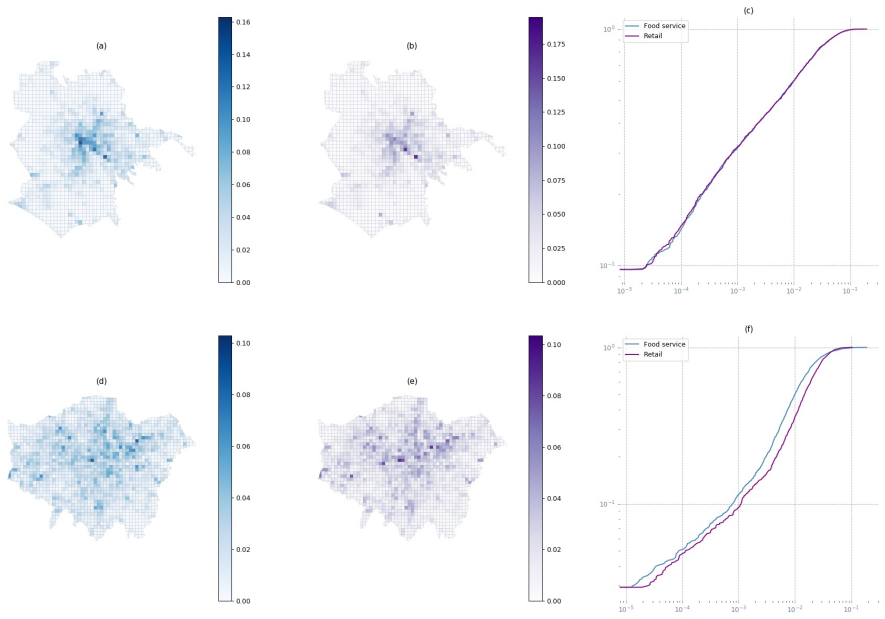


Figure 8-3: (a)-(b) Food service and retail activity APA distributions in Rome, (d)-(e) in London, (c)-(f) Log-log plots of empirical ECDFs in Rome and London at 12:00pm.

The spatial as well as empirical cumulative distributions (ECDF) of the computed APA values in Rome and London are presented in Figure 8-3. As can be seen from the ECDFs, the APA distributions in both cities are asymmetrically distributed: most of the grid cells have a very low centrality value, while only a handful of cells have a large centrality value. However, experiments aimed at identifying the analytical distributions yielded different results in the two cities. We conducted the fitting with the Python package "powerlaw" [16]. Parameters obtained via maximum likelihood estimation and the statistical goodness-of-fit measure quantified by the KS (Kolmogorov-Smirnov) test show differing results for the two cities: a truncated power law distribution for Rome ($p = 0.004$), and a log-normal-like distribution in the case of London ($p = 0.06$). Although the exact distribution is irrelevant here, this finding suggests that different data-generating mechanisms might be in place in the two cities.

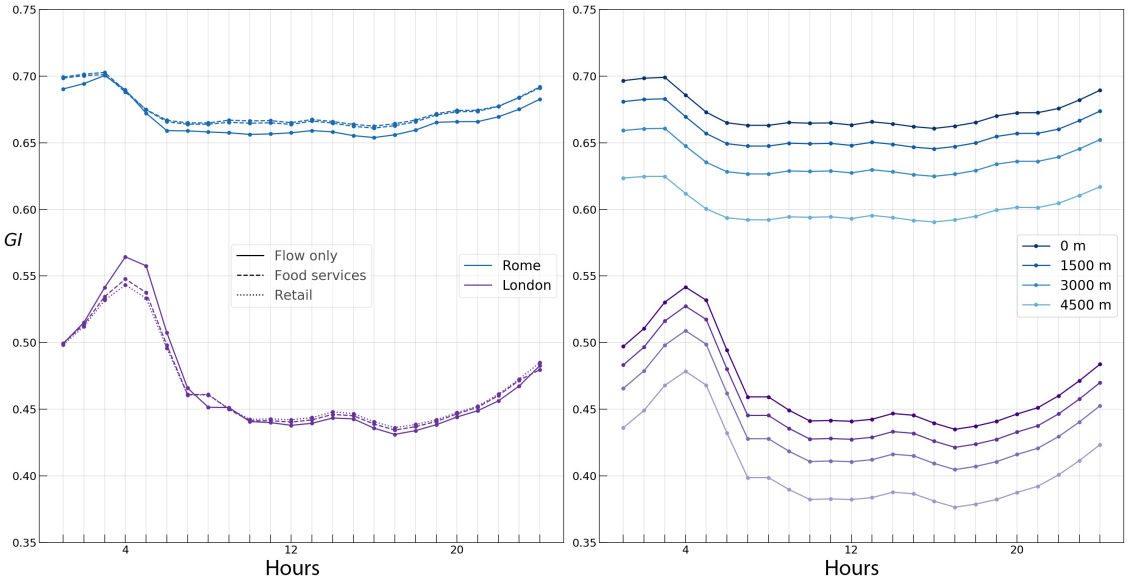


Figure 8-4: Gini (left) and Spatial Gini (right) coefficients during the day for flow only, food service, and retail activity in Rome and London.

8.3.2 Computing the Gini coefficients

We now proceed to analyzing the heterogeneity of the APA values in both cities, as described in section 8.2.2. In particular, as it is shown in Figure 8-4 (left), the daily average Gini coefficients in Rome and London take on values roughly 0.67 and 0.48, respectively. The temporal variation of the data is higher in London. In the same figure, we further observe a slightly higher Gini coefficient during the night hours in both cities, in accordance with the fact that most flows are associated with much fewer areas and thus yield a larger degree of concentration of activity during these hours.

With the aim of finding whether the Gini and Spatial Gini coefficients capture any difference between working days and weekends, both coefficients computed daily are represented in Figure 8-5. No significant change across the days of the week can be observed neither in Rome nor in London, while only a negligible rise of the coefficient on the weekend can be seen in London.

Despite the fact that some conclusions can be drawn from observing a relatively higher Gini coefficient during the night hours in both cities and on the weekends in London, the temporal evolution of the Gini coefficient, as can be seen in Figures 8-4

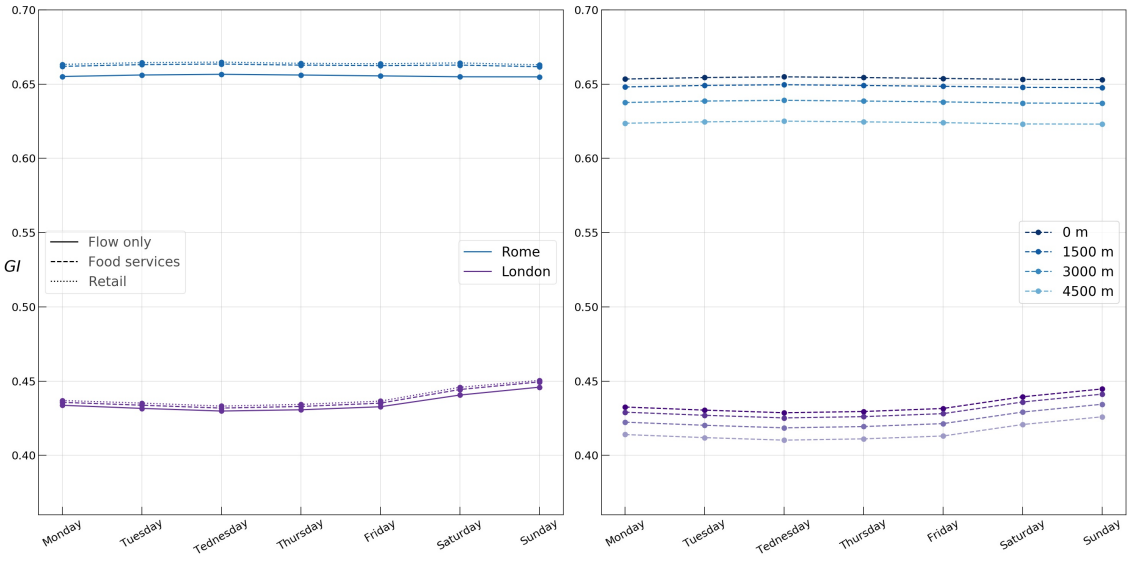


Figure 8-5: Gini (left) and Spatial Gini (right) coefficients during the week for flow only, food service, and retail activity in Rome and London.

and 8-5, conveys little significant information. Also, as mentioned in Section 8.2.2, it tells us nothing about the spatial distribution of the APA values.

In order to understand the temporal behaviour of the spatial component of the Gini coefficient, we resort to decomposing the Gini coefficient as described in Section 8.2.2. In essence, we are interested in finding how much of the Gini coefficient is due to non-neighbour heterogeneity. To achieve this, we follow the approach described in [223] and use the non-neighbour term in the Gini decomposition as a statistic to test for spatial autocorrelation:

$$GI_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(1 - w_{i,j}^{A}\right)|\mathbf{x}_i - \mathbf{x}_j|}{2n^2\bar{\mathbf{x}}}. \tag{8.6}$$

The expression (8.6) can be interpreted as the portion of overall heterogeneity associated with non-neighbour pair of grid cells. Inference on this statistic is carried out by computing a pseudo p-value by comparing the GI_2 obtained from the observed data to the distribution of GI_2 values obtained from random spatial permutations. It should be noted that this inference based on random spatial permutations is on the spatial decomposition of the Gini coefficient given by the expression (8.4), and not the value of the Gini coefficient itself.

Following the described approach, we proceed to the numerical experiments, varying the neighbourhood radius in the expression (8.6) from 1.5 to 6 kilometers. Both in Rome and London, the random spatial permutation approach yielded a statistically significant spatial decomposition for all hours of the day (p = 0.01). As demonstrated in Figures 8-4 and 8-5, the temporal profiles of the Spatial Gini coefficients closely follow the Gini profile. As the neighbourhood radius increases, the inequality due to non-neighbour APA values decreases, since the growing neighbourhood captures more and more of the inequality. We find a superlinear growth in the rate of decline of the Spatial Gini coefficient with increasing the neighbourhood radius, with a faster decline in Rome, suggesting a higher spatial concentration of urban flow in Rome.

8.3.3 Identifying urban hotspots

In order to obtain a clearer picture of the spatial structure of the "hotspot" cells with high APA values over time, we aim to compute the *spreading index* for flow, food services, and retail activity at different hours of the day in both Rome and London.

Figure 8-6 shows "hotspot" locations with APA values greater than the 50th, 75th, and 90th percentiles in Rome (a) and London (b). Remark the differences in "hotspot" locations in both cities for several percentiles. The "hotspot" concentration in Rome is significantly higher than in London, where we see spatial spread.

It is essential to perform a meaningful choice of the $\mathbf{x}^*$ for identifying the "hotspots" in equation (8.5). With the aim of choosing a threshold which will retain information without turning to noisy behaviour, we resort to a heuristic technique proposed in [172] based on the Lorenz curve from economics, see figure 8-7.

For a given distribution of data, the construction of the Lorenz curve proceeds as follows. For a set of values of cardinality n, the values are ordered in a non-decreasing sequence $\mathbf{x_i}$ with $i = 1...n$. The incomplete sums $L_i \equiv \left(\sum_{j=1}^{i} \mathbf{x_j}\right) / \left(\sum_{j=1}^{n} \mathbf{x_j}\right)$ are then plotted against $F_i \equiv i/n$. As described in [172], we note that the mean value $\bar{\mathbf{x}}$ corresponds to the projection point of the tangent of slope 1 on the x-axis and inverting $F(\overline{x}) = F_x$. The $\mathbf{x_{LB}}$ value is found from the intersection of the $x$axis with

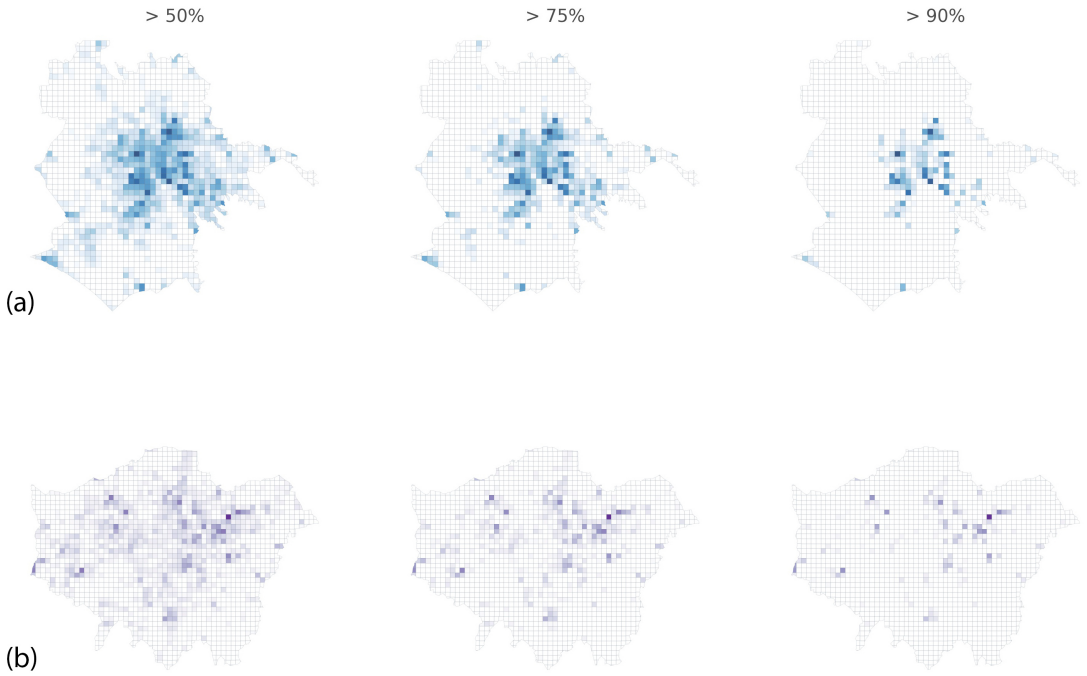


Figure 8-6: Hotspot locations with APA values greater than the 50th, 75th, and 90th percentiles in (a) Rome and (b) London.

the tangent of the Lorenz curve at $F_i = 1$ (red line). This method, called "LouBar", is inspired by the classical technique for determining the scale for an exponential decay. Indeed, if the decay from $F = 1$ were an exponential $\exp -(1 - F)/a$ where a is the scale to be determined, the described method would yield $1 - \mathbf{x_{LB}} = a$.

In Figure 8-8 we plot the spreading indices for different threshold values $\mathbf{x}^*$ over time in Rome and London. For low values of $\mathbf{x}^*$, the plots show relatively constant, low variance spreading indices over time, while for very large threshold values the spreading indices tend to become noisy.

In fact, the thresholds $\mathbf{x}^* = \bar{\mathbf{x}}$ and $\mathbf{x}^* = \mathbf{x_{LB}}$ form an interval $[\bar{\mathbf{x}}, \mathbf{x_{LB}}]$ containing all reasonable choices for determining the "hotspots". However, since the lower bound $\bar{\mathbf{x}}$ results in a curve with little variation during the day, and since values from the interval closer to the LouBar value give similar results to the Loubar value itself, we will proceed with this choice (see Figure 8-8).

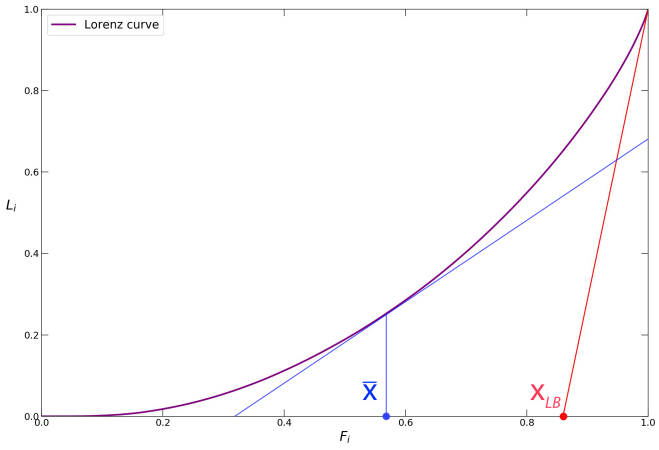


Figure 8-7: Lorenz curve for a data distribution.

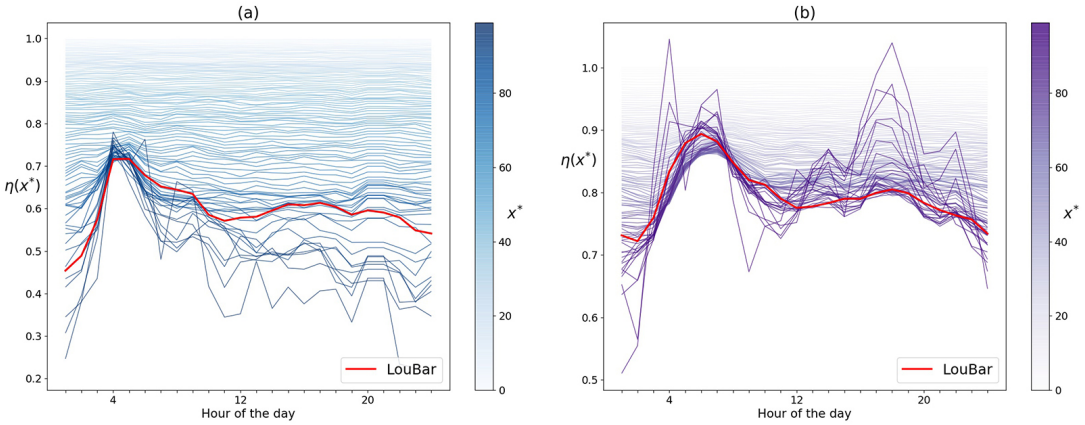


Figure 8-8: Spreading indices over time for various thresholds $\mathbf{x}^*$ in (a) Rome and (b) London.

8.3.4 Computing the spreading index

In this section, we present the results of studying the spreading index profiles on a typical day in Rome and London, and build hypotheses regarding their interpretations.

Having chosen the threshold value $\mathbf{x}^*$, we compute the hourly profiles of the spreading indices for flows only, food services, and retail activities in Rome and London. Since the data sets of raw GPS trajectories at our disposal span two years, we extract hourly OD networks across the working days and obtain sampling distributions and corresponding 95% confidence intervals of spreading indices at

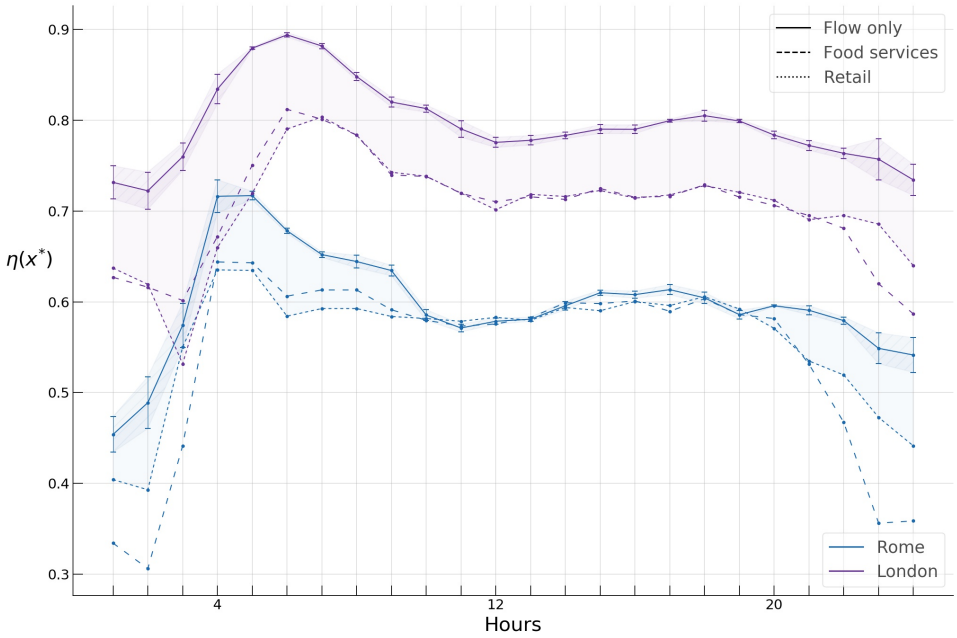


Figure 8-9: Spreading indices for flow only, food services, and retail activity in Rome and London during a typical day.

each hour with the aim of testing our results for robustness (Figure 8-9). The wider confidence intervals in the night hours are due to less available data for these hours.

First, we find a significant difference in the *spreading index* hourly profiles of Rome and London. During a typical day, the former varies from around 0.4 to 0.7, while the latter varies from around 0.65 to almost 0.9, suggesting a considerably higher concentration of "hotspots" during the day in Rome compared to London.

Next, we see structural similarities in the hourly patterns of the spreading indices in both cities. As shown in Figure 8-9, the spreading indices for all types of activities demonstrate a similar inverted U-like pattern, with the spreading index increasing considerably during the night hours, bulging during the morning and evening hours, and declining during the late evening hours. The rapid rise of the index during the night hours could possibly be attributed to the fact that most mobility during these hours is due to flows on highways located in the periphery of both cities, thus yielding a higher η, while the bulging of the index at morning and evening hours is

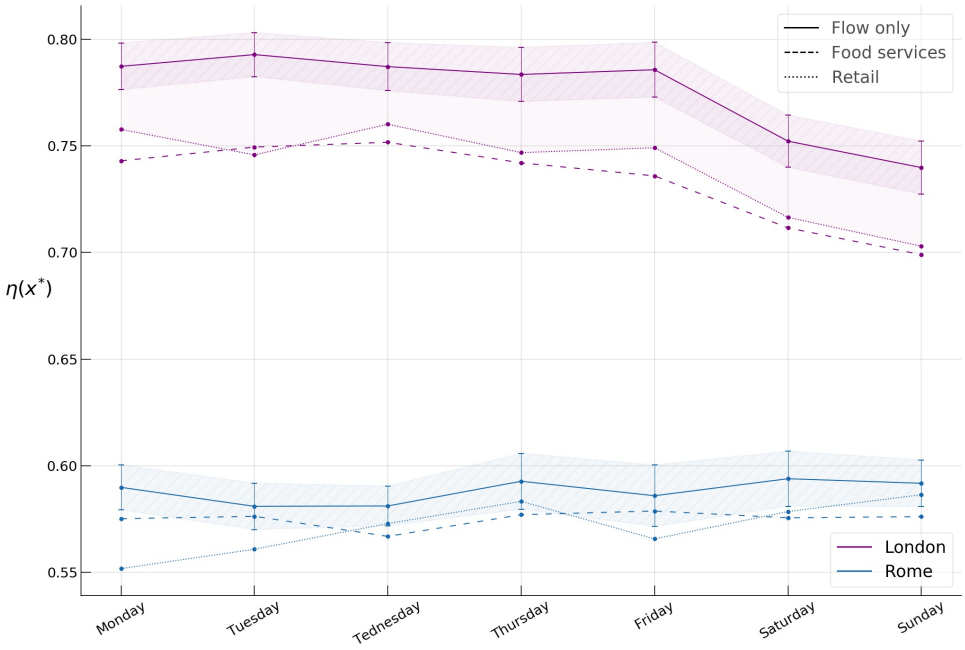


Figure 8-10: *Spreading indices* for flow only, food services, and retail activity in Rome and London during the week.

likely due to core-periphery commuting flows.

We further observe a large gap of around 0.1 between the flow only η profile and those of food services and retail in London, while similar, albeit smaller gaps in Rome can be observed only during the late evening and night hours (shaded areas in Figure 8-9), whereas the profiles for all types of activities collapse very close to each other during the working hours. This gap can likely be attributed to the "London congestion charge"[1], which has dramatically reduced private cars in central London since its introduction, while most of food services and retail stores and shops are located in the central part of London, bringing the spreading index down for these activities. In Rome, on the other hand, a similar gap exists only during the night and early morning hours, which one can intuitively expect since most of the food services and shops have a central location, decreasing η, while during these hours most of the flows are due to inter-peripheral highway flows which increase η.

[1]https://tfl.gov.uk/modes/driving/congestion-charge

In Figure 8-10, the spreading indices across the days of the week are shown. We use the 104 weeks of available data to build an empirical 95% confidence interval for the *spreading index*. We see the already familiar gap between the flow only and the other two types of activities in London. Further, we detect a statistically significant ($p < 1e^{-5}$) change in the index for London, while no significant change appears to be present in Rome.

8.3.5 The time-space spreading index (TSI)

We have previously computed and tracked the *spreading index* η over a typical day in Rome and London. The *spreading index*, being based on Euclidean distances between the cell centroids, represents geographic space, but fails to capture urban mobility. In particular, due to congestion in cities at peak hours, travel times can be said to distort the perception of space. If travel times are considered as a measure of distance, geographically very close locations in the city center might turn out to be further away than geographically further placed locations in the city periphery with low traffic. For this reason, we enable the spreading index to capture urban mobility by introducing the *time-space spreading index* (TSI), essentially replacing the distances in the calculation of the spreading index η by the pairwise average travel times:

$$TSI\left(\mathbf{x}^*\right) = \frac{\frac{1}{N(\mathbf{x}^*)} \sum_{i,j} t(i,j) \mathbf{1}_{(\mathbf{x}_i > \mathbf{x}^*)} \mathbf{1}_{(\mathbf{x}_j > \mathbf{x}^*)}}{\frac{1}{N} \sum_{i,j} t(i,j)}, \tag{8.7}$$

where $t(i,j)$ is the average travel time from cell i to cell j, and is obtained using the Google Distance Matrix API. This constitutes an important dimension for studying the spatio-temporal characteristics of the "hotspots" in the mobility networks.

Therefore, we then proceed to analyze the time-space spreading index TSI of the three activities during a typical day in Rome and London.

The spreading indices and TSIs for Rome and London are shown in Figure 8-11. While the two measures are very close to each other during the night hours, they start to deviate significantly during the rest of the day. At these hours, the TSI in both cities is considerably higher than the *spreading index*, hinting at the above-mentioned

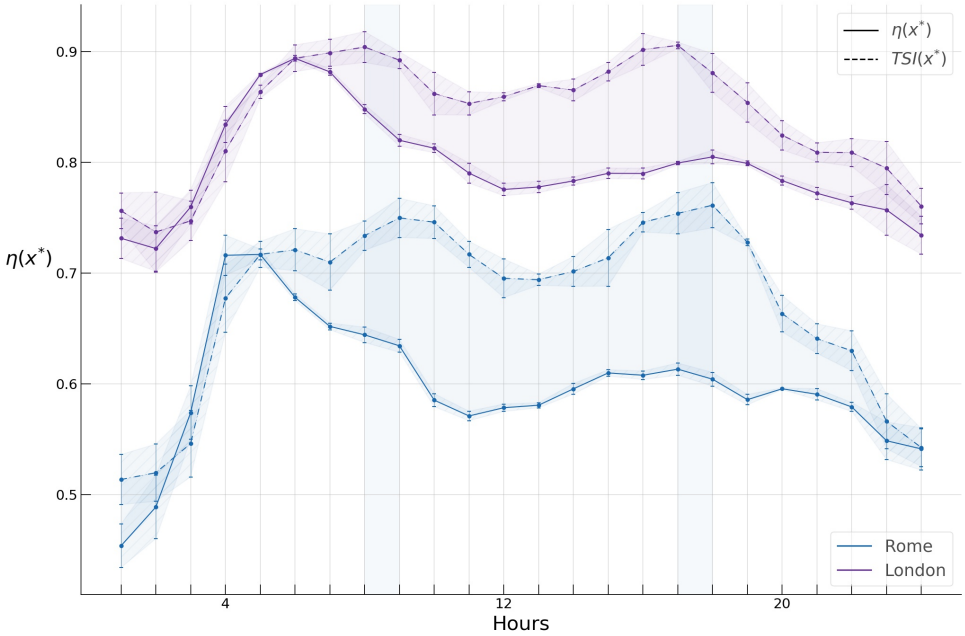


Figure 8-11: Spreading index and time-space spreading index (TSI) with corresponding 95% confidence intervals during a typical day in Rome and London.

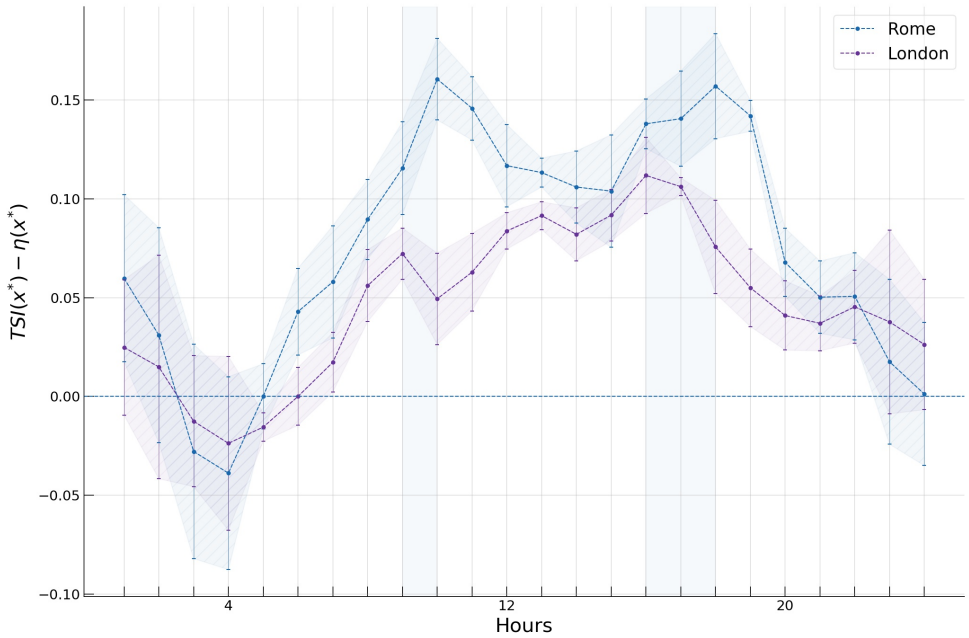


Figure 8-12: Tracking the difference $TSI(\mathbf{x}^*) - \eta(\mathbf{x}^*)$ in Rome and London during a typical day.

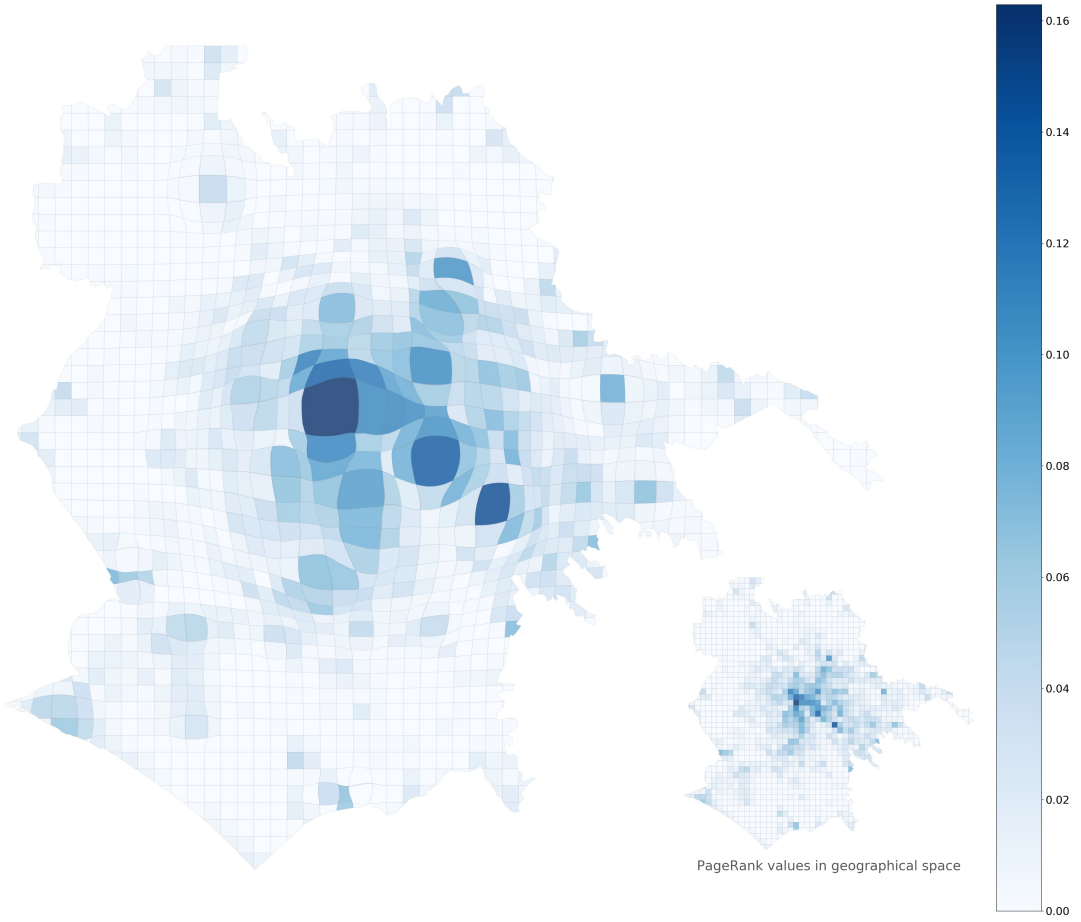


Figure 8-13: Retail APA values at 18:00 in Rome represented with pairwise time-weighted distances between grid cells using multidimensional scaling (MDS). The inset shows the same set of values in geographical space.

space-time distortion, in which geographically close central locations become further apart because of longer travel times due to traffic, effectively increasing the TSI compared to the *spreading index*. This effect is shown in Figure 8-13, where the time-weighted distances used in computing the TSI are visualised with multidimensional scaling (MDS) [13].

Note that the confidence intervals for the TSI values are wider than those of the *spreading indices* since additional uncertainty is introduced in the calculation of the TSI by including travel times contingent on volatile traffic conditions (Figure 8-11).

We also note the two peaks of higher TSI values during the morning and evening commuting hours forming a circadian rhythm in both cities. A peculiar observation is the mismatch of the peaks between the two cities. Rome seems to be "late" by roughly an hour (vertical shaded areas in Figures 8-11 and 8-12).

In Figure 8-12 we plot the differences $TSI(\mathbf{x}^*) - \eta(\mathbf{x}^*)$ during the day in Rome and London. We observe this difference during the day to be consistently greater in Rome, suggesting congestion to have a larger impact on the spatio-temporal characteristics of the "hotspots" in Rome.

The TSI for the hotspots of the three types of activities during a typical day in both cities are displayed in Figure 8-14. One can note a gap in London between the flow only temporal TSI profile, and the food services and retail TSI profiles, consistent with a similar gap in the case of the spreading index discussed in Section 8.3.4.

8.4 Conclusion

In this Chapter, we have proposed a generic end-to-end workflow for analyzing spatio-temporal characteristics of urban mobility induced "hotspots" for different types of activities in cities, and have demonstrated it in case studies in Rome and London. The proposed workflow comprised data mining of GPS data, the subdivision of the urban territory into regular grid cells, construction of temporal OD networks, addition of socio-economic activity attributes to the OD network nodes

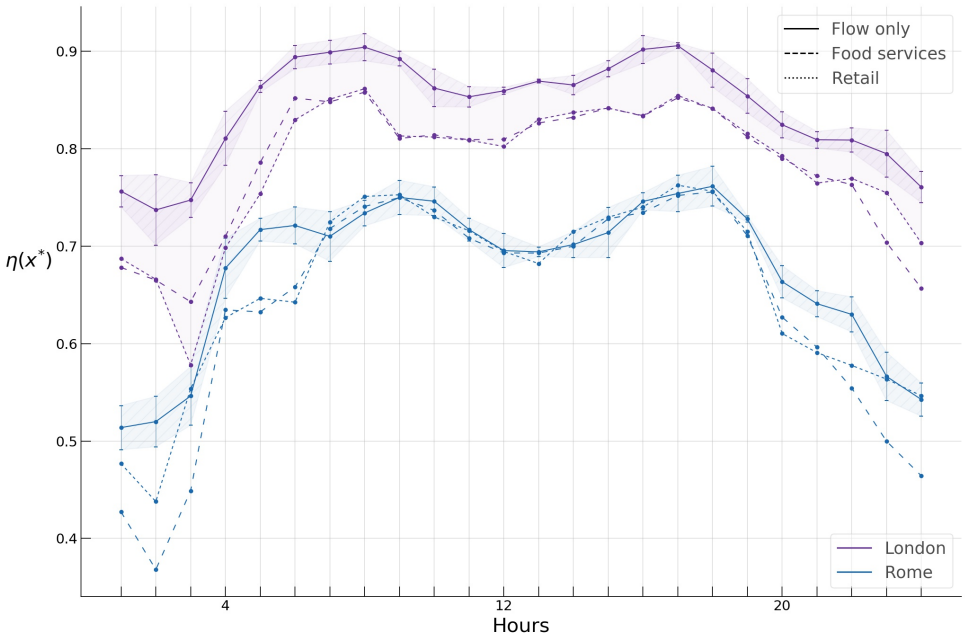


Figure 8-14: TSI for flow only, food services, and retail activity in Rome and London during a typical day.

from PoI data, computation of the attribute-enhanced APA centralities in the OD networks on an hourly or daily basis, identification of "hotspots", and visualisation and analysis of measures of their spatial heterogeneity. The obtained results led us to a series of hypotheses regarding their nature, the study of which will be the target of future work.

In particular, we observed an increase in both the *Gini coefficients* as well as the *spreading indices* during the night hours, suggesting higher inequality and spatial spread, respectively. However, a further decomposition of these measures would be required to determine what share of these inequality and spatial spread is due to core-periphery, inter-peripheral, or highway transit flows. Also, future work will be aimed at understanding whether there is a hierarchy of "hotspots" and how it evolves over time. Further, the hypothesis that the peculiar gap between the flow only and food services and retail *spreading index* profiles in London has to do with the congestion charge, and whether our approach can be adopted as a traffic management indicator, requires further study. Further yet, we note that a methodology needs to be developed and tested for using the measures proposed in this Chapter as monitoring tools in connection with specific urban planning policies in a particular city. For instance, deciding critical values of the proposed measures, beyond which action would be required on the part of the urban planners.

Notwithstanding the mentioned shortcomings, our approach has direct utility to urban planners and policy makers. It highlights the road map for creating analysis, visualisation, early warning, or trend detection tools with simple information-rich measures for monitoring city-wide spatial characteristics of mobility related to various socio-economic activities. The proposed workflow from raw data input to analysis and visualisation is generic enough to accommodate other types of spatial movement data (e.g., call detail records (CDR), smart card, etc.) as well as other socio-economic activities in cities over both short and long terms.

"The Network is the Message."

Manuel Castells, 2001

Chapter 9

Explaining mobility from urban attributes

9.1 Introduction

The conventional model-driven approaches to human mobility, such as (constrained) gravity, maximum entropy, intervening opportunities and radiation models ([93], [280]) have recently been challenged and augmented by machine learning, specifically deep learning techniques ([251], [94]). The latter focus on improving predictive power, but often fail to provide insights about how the observed mobility is related to urban form and socio-economic dynamics. Existing studies on these kind of relations, linking urban functions and mobility ([269], [298], [289]) are narrowly tailored to a specific question, e.g., quantifying flows between a selected number and type of Points-of-Interest. In this Chapter, we will attempt to *explain* urban mobility flow networks from urban socio-economic attributes by means of developing a network regression model respecting the network topology and offering a statistical framework for parameter estimation.

9.2 Background and related work

9.2.1 Goodness-of-fit measures

Coefficients of determination It has been suggested that a variety of goodness-of-fit measures be used for evaluating spatial interaction model performance (Knudsen, Fotheringham 1986), among which we will focus our attention to the pseudo R^2 statistic based on the likelihood function (McFadden 1974),

$$R^2_{pseudo} = 1 - \frac{\ln \hat{L}\left(M_{full}\right)}{\ln \hat{L}\left(M_{intercept}\right)}, \tag{9.1}$$

where $\hat{L}$ is the model likelihood, M_{full} is the model with all the covariates included, and $M_{intercept}$ is the model with only the intercept (i.e., no explanatory variables). To account for model complexity, we also use an adjusted version of this measure,

$$R^2_{\text{adj-pseudo}} = 1 - \frac{\ln \hat{L}\left(M_{full}\right) - K}{\ln \hat{L}\left(M_{Intercept}\right)} \tag{9.2}$$

where K is the number of covariates in the model. Both model fit measures achieve a maximum at a value of 1, with higher values meaning better model fit.

Akaike information criterion Another goodness-of-fit statistic utilised in the study is the Akaike information criterion (AIC),

$$AIC = -2 \ln \hat{L}\left(M_{full}\right) + 2K, \tag{9.3}$$

which also accounts for model complexity and in which lower values denote a better fit (Akaike 1974). This statistic has its roots in information theory, with the AIC being interpreted an asymptotic estimate of the information lost by using the full model to describe a given process.

Standardized root mean square error Since the R^2 and AIC goodness-of-fit measures have been devised for model selection, and hence should not be used to compare different models, we also use the standardized root mean square error (SRMSE),

$$\text{SRMSE} = \sqrt{\frac{\sum_i \sum_j \left(F_{ij} - \hat{F}_{ij}\right)^2}{nm}} \Big/ \frac{\sum_i \sum_j F_{ij}}{nm}, \tag{9.4}$$

where the numerator shows the root mean squared error between the observed flows, F_{ij}, and the predicted flows, $\hat{F}_{ij}$, while the denominator denotes the mean of the observed flows and is the quantity by which the root mean squared error is standardized by. nm is the number of origin-destination pairs (Knudsen, Fotheringham 1986).

Sorensen similarity index We also utilize the modified Sorensen similarity index (SSI), as it is being increasingly used in spatial interaction studies (Lenormand et al. 2012, Masucci et al. 2012, Yan et al. 2013). This statistic is defined as,

$$SSI = \frac{1}{(nm)} \sum_i \sum_j \frac{2 \min\left(F_{ij}, \hat{F}_{ij}\right)}{F_{ij} + \hat{F}_{ij}}, \tag{9.5}$$

and takes on values between 0 and 1, with values closer to 1 denoting a better fit.

9.2.2 Gravity, Poisson, and Negative Binomial models

	exponent	**Std.Err**	***p*-value**
intercept	-5.6118	0.0066	<1e-16 ***
outflow	0.6574	0.0006	<1e-16 ***
inflow	0.6547	0.0006	<1e-16 ***
distance	-0.2752	0.0001	<1e-16 ***

Figure 9-1: Estimated parameters of the gravity model for the mobility flows in London

Following the discussion of gravity models of human mobility in Chapter 2, we show in Table 9-1 the parameter estimates for the doubly constrained gravity model applied to the attributed urban mobility OD network in London. In the regression performance Table 9.1 presented in subsection 9.4.2, we see that the constrained gravity models perform progressively better on all goodness-of-fit metrics discussed in subsection 9.2.1.

Continuing the discussion of residual plots for the Poisson regression, we apply the model to the London data, and observe that the model yields a result for which

the majority of the Pearson residuals are concentrated close to zero with the greatest divergence from the model also occurring there (see Figure 9-2).

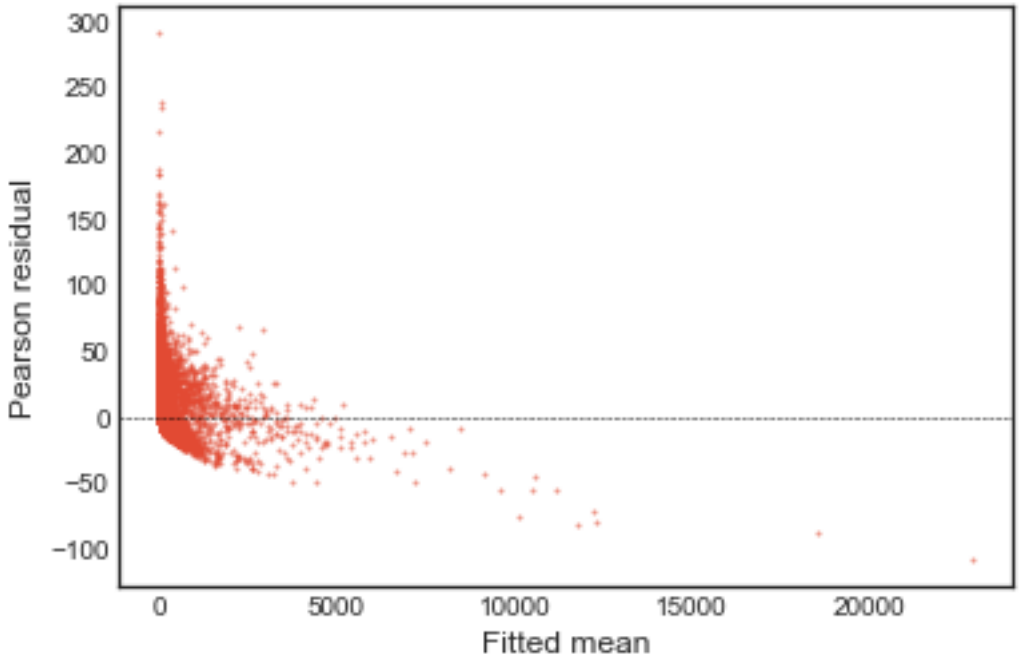


Figure 9-2: Pearson residuals plotted against fitted means. Rome and London data.

Since a Poisson random variable has a variance equal to its mean, so what, we would expect to observe this for our data if our assumption about the underlying distribution is correct.

As already discussed, the Pearson residual is scaled to compensate for non-uniform variance and hence calls for a more uniform vertical scattering than obtained. The Pearson statistic, the ratio between the sum of squares of the residuals and their degrees of freedom, captures this, being approximately 1 if the data is drawn from a Poisson distribution. A value greater than 1, as in our case ($>> 1$ for London, Rome, and Shenzhen), suggests the presence of overdispersion, which can arise when some unobserved variables are contributing to the mean but are not captured by the model.

The Negative Binomial regression model brings the Pearson statistic significantly down closer to $n - k$, but still keeps it at least an order of magnitude larger for both cities. However, as we see from the results (Table 9.1), it yields a better performance

on the adjusted R^2 and SSI measures, but performs worse on the SRMSE and AIC metrics. This suggests that loosening the mean-variance constraint does not contribute to obtaining an adequate model, and that neither the Poisson nor the Negative Binomial models succeed in explaining the variability in intra-urban OD flows.

The spatial autoregressive model (SAR), with its spatially adjusted Poisson and Negative Binomial variants described in Section 2.3.4, yields a further improvement of an order of magnitude for the Pearson statistic in both London and Rome, a negligible change in SRMSE, and a slight improvement for the SSI metric.

Thus, despite producing plausible coefficients, the models discussed so far fail to capture some key mechanism at play in generating OD flows, and which we conjecture to be an interplay between non-spatial and spatial network effects.

9.3 Multilayer Network Regression

9.3.1 Generalised Hypergeometric Ensembles (gHypE)

Thus far, we have extensively discussed regression models aiming to explain the dependent variable (the observed flows) as a function of independent ones (the dyadic relationships between city locations), accounting for random effects. However, mobility OD flows form a network and the dyadic relations in it are not independent. Because of this, ordinary least squares regression models are unsuitable for analysing network data ([151]). To overcome this, several different network regression models have been proposed ([237]). However, these models suffer from either having been developed for *unweighted* graphs or not taking into account *combinatorial* effects in the network of interactions. Combinatorial effects refer to the fact that elements interacting more in general are also more likely to interact among themselves. In network theory, this problem is known as *degree correction*. For example, two city locations might have a high mobility flow between them because one of them is located in a dense residential area and the other in the central business district, or because the other is an airport, or simply because one of them has large total out-flow and the other large total in-flow, and hence have a high chance to have

flow between them. Therefore, significant relations have to be disentangled from combinatorial effects when modelling the system.

To overcome the mentioned problems, we follow the approach for statistical regression on networks introduced by [63]. The proposed method builds upon generalised hypergeometric ensembles (gHypE), a recently developed class of statistical random graph ensembles ([64]). gHypEs provide an elegant formulation of the well-known configuration model in terms of an urn problem, incorporating arbitrary tendencies to form relations with combinatorial effects.

The urn problem to which the process of drawing edges in the proposed network model is mapped to is based on the following intuition. Suppose there is an urn filled with 100 balls of 3 different colors: 20 blue, 30 red, and 50 yellow. If 40 balls are chosen at random without replacement, what is the probability that 10 of them are red, 10 yellow, and 20 blue? The gHypE statistical random graph can then be thought in terms of the total number of edges corresponding to the total number of colored balls in the urn problem, each dyadic pair of nodes corresponding to a different color, and each multi-edge between the nodes of the dyadic pair corresponding to a single ball in the urn. This probability is described by the multivariate hypergeometric distribution. However, it assumes a uniform probability of drawing balls of each color. As we saw in Chapter 4, urban mobility flows, i.e. the edges (dyadic pairs) in the OD network have an extremely unequal distribution. This suggests that the original urn problem be transformed into balls of different colors having different propensities - probabilities of being sampled. These propensities are encoded by the various dyadic relationships between nodes and used as independent variables in the network regression setting, where the dependent variable - the OD flows - is modelled as a realisation of the gHypE statistical random graph ensemble. Furthermore, its compact analytical formulation allows for statistically testing the significance of the regression model against the observed interactions. A more detailed description of gHypE can be found in the Appendix B.1 and in the original paper.

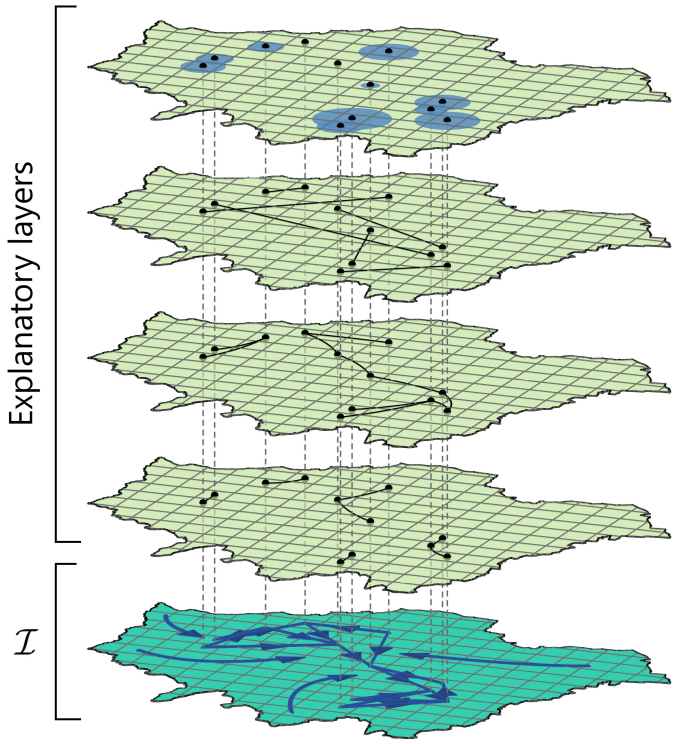


Figure 9-3: The multilayer network representation of the attributed OD network in London. The bottom layer (dark green) captures the observed flow counts between the cells. The top layers (light green) encode different types of relations, such as network distance, average of Airbnb prices, product of population densities, bus or subway network, etc. The gHypE network regression model allows us to *explain* the impact of these relational layers on the OD flows.

9.3.2 Multilayer Network Representation

The urban mobility OD flow network, together with the collection of p dyadic relations, can be represented as a multilayer network with the flow counts forming the interaction layer, and the dyadic relations between nodes - the covariates, forming the explanatory layers.

In the proposed framework, each unit of flow between locations in the interaction layer $\mathcal{I}$ is represented as a separate edge, thus turning the interaction layer $\mathcal{I}$ into a multi-edge graph. Further, each type of dyadic relation can be seen as a weighted graph in which the weight of each edge encodes the strength of that dyadic relation. These p graphs constitute the relational layers R_l with $l \in [1, p]$. The $n = |V|$ nodes and $p + 1$ layers then form the multilayer network $\mathcal{M}$. Figure 9-3 illustrates the

multilayer approach we take.

Given the multilayer network representation of the attributed urban OD flow network with its interaction layer a realisation of the generalised hypergeometric graph ensemble **gHypE**(n, m), we follow the framework proposed in [63] for statistical regressions with network layers.

We assume the multi-edged network layer $\mathcal{I}$ to be the dependent variable regressed on the remaining layers R_l which we consider to be the independent, explanatory variables. The resulting model then assumes the following form:

$$\mathcal{I} = f\left(\mathcal{R}_1, \ldots, \mathcal{R}_r; \beta_1, \ldots, \beta_r\right) \tag{9.6}$$

where the parameters $\beta_l, l \in [1, p]$ are the model parameters corresponding to each explanatory layer R_l.

9.3.3 Statistical model

Our aim is to model the interaction layer $\mathcal{I}$, represented as a multi-edged graph with fixed total number of edges m. As mentioned in the previous section, we model the interaction layer $\mathcal{I}$ as a realisation from a generalised hypergeometric ensemble, with n vertices and m edges. We denote with $\mathbf{A}$ the adjacency matrix of the interaction layer $\mathcal{I}$ and its elements with A_{ij}, $i, j \in V$. The relational layers are represented in a similar way: let $\mathbf{R}_l$ denote the adjacency matrix of the relational layer R_l and $\beta \in \mathbb{R}^p$ be the p-vector of coefficients in the regression. In this setting, $\mathcal{I}$ is distributed according to the Wallenius non-central hypergeometric distribution [64]:

$$\Pr(\mathcal{I}|\mathcal{R}) = \left[\prod_{i,j} \binom{\Xi_{ij}}{A_{ij}}\right] \int_0^1 \prod_{i,j} \left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{A_{ij}} dz \tag{9.7}$$

with $S_\Omega = \sum_{i,j} \Omega_{ij} \left(\Xi_{ij} - A_{ij}\right)$, where Ω encodes the propensity of pairs of nodes to interact, and Ξ the probability that pairs of nodes interact due to combinatorial effects, as described in [64].

The entries of the matrix of possible edges Ξ are assumed to be constructed according to the configuration model which randomly shuffles and rewires the topology

of a network while preserving the node degrees [71]. This is the most general way to model the combinatorial effect generated by the various activities of nodes represented by their degrees. This formalises the idea that more active nodes, i.e. those with a higher degree, are more likely to interact. Therefore, Ξ is completely defined by $\mathcal{I}$. The nodes' interaction propensity Ω, an important element in equation 9.7, depends on the explanatory layers $\{\mathcal{R}_l\}_{l\in[1,p]}$:

$$\Omega := \prod_{l=1}^{p} \mathbf{R}_l^{\beta_l}. \tag{9.8}$$

The statistical model in equation 9.6 can now be specified. f is considered to be the expected value of the hypergeometric graph ensemble that maximises the probability of observing $\mathcal{I}$, given the explanatory layers $\{\mathcal{R}_l\}_{l\in[1,p]}$:

$$\mathcal{I} = \mathbb{E}\left[\mathbf{gHypE}(n, m)|\mathcal{R}_1, \ldots, \mathcal{R}_p\right]. \tag{9.9}$$

Equation 9.9 is therefore equivalent to finding maximum likelihood estimators (MLE) for the parameter vector β in equation 9.7.

Following equation 9.7, given the observed interaction network $\mathcal{I}$, the likelihood of β is then defined as

$$L(\beta|\mathcal{I}) = \left[\prod_{i,j}\begin{pmatrix}\Xi_{ij}\\ A_{ij}\end{pmatrix}\right]\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\prod_{l=1}^{p} R_{l,ij}^{\beta_l}}{S_\beta}}\right)^{A_{ij}} dz, \tag{9.10}$$

with $S_\beta = \sum_{i,j}\prod_{l=1}^{r} R_{l,ij}^{\beta_l}\,(\Xi_{ij} - A_{ij})$.

As explained in [63], the numerical maximisation in equation 9.10 is not straightforward, and for $m \ll \sum_{ij}\Xi_{ij}$, which is the case in the urban OD network, the multivariate hypergeometric distribution can be approximated up to constants by the multinomial distribution:

$$L(\beta|\mathcal{I}) \sim \prod_{i,j\in V}\left(\frac{\Xi_{ij}\prod_{l=1}^{p} R_{l,ij}^{\beta_l}}{\sum_{i,j\in V}\Xi_{ij}\prod_{l=1}^{p} R_{l,j,j}^{\beta_l}}\right)^{I_{ij}}. \tag{9.11}$$

The MLE $\hat{\beta} = \text{argmax}_\beta(L(\beta|\mathcal{I}))$ is then obtained by finding numerical solutions to $\nabla L(\beta) = 0$. The components of the gradient of the log-likelihood $\nabla\log(L(\beta))$

are given as

$$\frac{\partial \log(L(\beta|\mathcal{I}))}{\partial \beta_l} = -m\frac{\sum_{ij} \log\left(R_{l,jj}\right) \Xi_{ij} \prod_{l=1}^{p} R_{l,ij}^{\beta_l}}{\sum_{ij} \Xi_{ij} \prod_{l=1}^{p} R_{l,ij}^{\beta_l}} + \sum_{ij} I_{ij} \log\left(R_{l,ij}\right). \tag{9.12}$$

After computing the MLEs $\left\{\hat{\beta}_l\right\}_{l\in[1,p]}$ for the p explanatory layers $\{\mathcal{R}_l\}_{l\in[1,p]}$ representing the *strength* of the effect each layer exerts on the interaction layer $\mathcal{I}$, we carry out statistical significance tests for the obtained parameters. This procedure is described in Appendix B.2.

9.4 Results

In this section, we present the principal results of the gHypE network regression on the urban mobility OD networks in Rome and London. First, the multilayer networks for both cities are constructed as described in Section 9.3.2. Then, the MLE estimates for the network regression parameters best explaining the observed OD network are obtained. Finally, the evolution of the obtained regression parameters across hours of a typical day are tracked and compared between the two cities.

9.4.1 Data

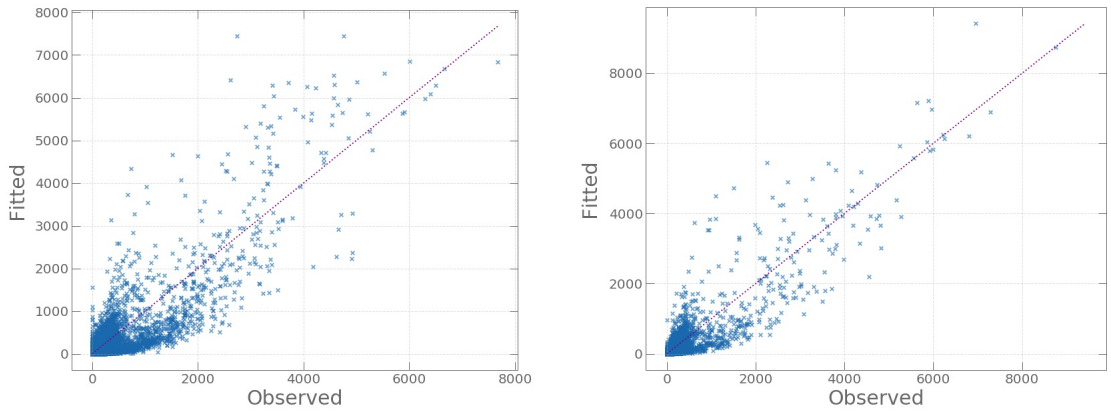
The multilayer network data used in this regression setting is essentially the same dataset as descibed in Chapter 4, except that all node attributes are transformed into dyadic relations through the simple operations of averages or products to capture the relation's strength. Since the OD flow matrices were obtained from individual crs GPS trajectories, which naturally extend beyond the city administrative boundaries, those Also, the APA centrality measures obtained in the previous chapters and transformed into dyadic relations through pairwise averaging are added as relational explanatory layers as well.

Model	Adjusted R^2	AIC	SRMSE	SSI
gravity	0.4698	1.2124e+07	16.4691	0.3789
origin-constrained gravity	0.4881	1.1704e+07	16.3256	0.3841
destination-constrained gravity	0.4879	1.1709e+07	16.3283	0.3884
doubly-constrained gravity	0.4924	1.1606e+07	15.8827	0.3937
Poisson log-linear regression	0.6291	8.2458e+05	13.2451	0.4615
Negative Binomial regression	0.5258	2.1854e+06	13.5257	0.5382
Spatially adjusted Poisson	0.6571	3.1108e+05	12.9981	0.5321
Spatially adjusted NB	0.5869	1.0201e+06	12.3859	0.5819
gHypE multilayer regression	**0.7228**	**8.2209e+03**	**7.4491**	**0.6194**

Table 9.1: Comparison of gHypE multilayer regression performance against baseline methods

9.4.2 Comparison with baseline models

Table 9.1 summarises the performance of our network regression model against the baseline models described in Section 2.3 as measured by goodness-of-fit metrics we introduced in subsection 9.2.1. It can be clearly seen that the gHypE network regression model outperforms the baseline models on all metrics.



(a) (b)

Figure 9-4: gHypE network regression fitted prediction values for (a) London (b) Rome

The fitted values of the gHypE network regression against the observed OD flows in London and Rome are shown in Figure 9-4. It should be noted, however, that the objective of this model is not to claim any predictive power but rather to focus on *explaining* the effects different relational layers have on the OD flows.

Coefficients	Estimate	Std. Err	p-value
betweenness	0.132	5.286e-02	1.995e-06
network distance	-1.256	1.149e-01	<1e-16
route factor	0.312	1.769e-01	2.124e-02
airbnb	-0.82	2.488e-02	4.321e-02
time	-0.631	2.717e-02	<1e-16
speed	-0.152	5.494e-02	1.112e-03
bus	0.209	3.361e-02	3.921e-02
subway	-0.121	3.481e-02	3.199e-02
densities	0.228	8.396e-03	4.289e-03
residential to rest	0.208	5.824e-02	1.119e-02
highway	0.062	1.098e-01	2.819e-03
correlation	-0.479	1.692e-02	<1e-16
APA flow only	0.670	3.611e-02	<1e-16
APA food services	0.607	2.001e-02	<1e-16
APA retail	0.634	1.452e-02	<1e-16

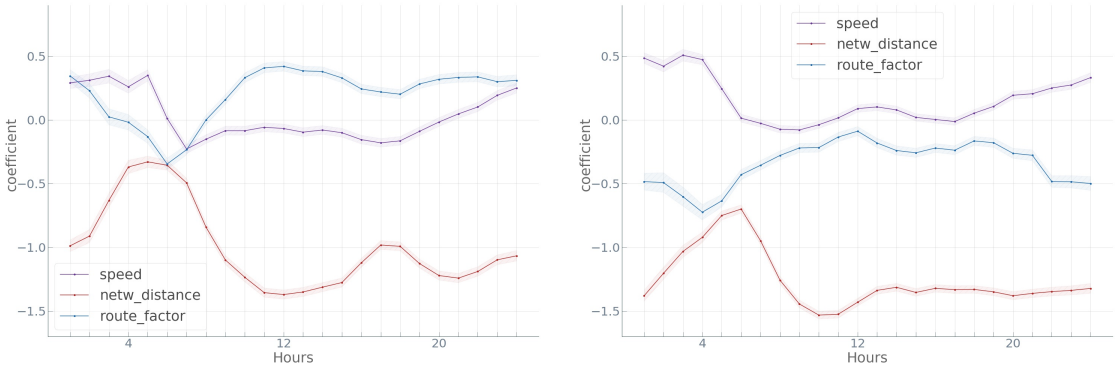
Table 9.2: MLE coefficients of the gHypE multilayer regression model in London

9.4.3 Estimated coefficients

Those explanatory layer coefficients obtained by MLE that are statistically significant are summarised in Table 9.2. Network distance and population densities - as expected, with negative and positive exponents, respectively - are treated as control variables. In order to account for the spatial boundary effects, the highway explanatory layer is also treated as a control variable - a small positive effect as can be seen in Table 9.2. We then focus our attention on the remaining relational explanatory layers of interest.

Figure 9-5 shows the temporal evolution of the estimated coefficients for speed, network distance, and the route factor over the hours on a typical weekday in London and Rome. We see structurally similar but locally differing results in both cities. The impact of network distance during the night hours is markedly weaker than during the day. This can be explained by the fact that the massive amount of short travels during the day do not take place in the night, and drivers cover longer distances during the day. To test this hypothesis, we conducted a two-sample t-test and compared the average distance travelled during the night hours (from 2am to

7am) to that of the day hours and found circa 17 vs. 7 km and 15 vs. 4 km in London and Rome, respectively. On the contrary, speed has a stronger positive effect during the night and late evening hours compared to working hours. This captures the intuition that average speed during the working hours is smaller due to traffic. Indeed, the difference between average driving speeds during night and day hours in London and Rome are 14 and 16 km/h respectively. The route factor, roughly representing the journey's divergence from a straight line, also displays an interesting behaviour with a significant dip during the night hours during which its negative effect is strengthened. Intuitively, this should be the case since during night hours drivers make less detours to reach a destination, both because of less traffic and because of longer journey distances.



(a) (b)

Figure 9-5: Speed, network distance, and route factor coefficients over time in (a) London (b) Rome

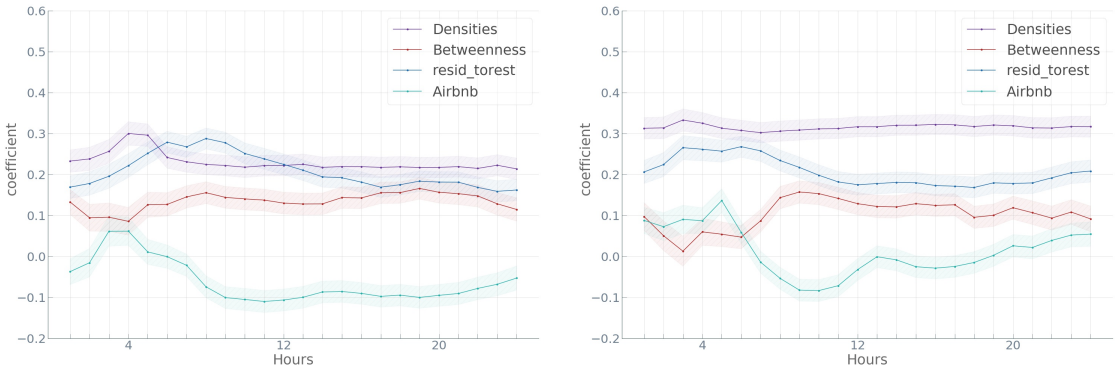
Next, we observe in Figure 9-6 that there is a more or less constant positive effect of population densities on the OD flows in both cities, with a small increase during the night hours. This can possibly be explained by the fact that since there are essentially no business trips to locations with a small population density during the night, its effect will be higher. We also observe a small difference in the effect between the cities, with the density effect being higher in Rome.

Average betweenness centrality also tends to have a positive constant effect in both cities, with a small decrease during the night hours. This can be attributed to the fact that most locations with a high street betweenness centrality value are relatively centrally located (Figure 4-4), while a significant amount of night trips

are peripheral to the city. Although this explanation requires further study and testing, we saw Chapter 8 that the spreading index related to urban mobility flows was indeed very high.

We also observe in the same figure that Airbnb prices have a negative, close to zero effect on OD flows, with the minimum value during the morning commuting time, and with a small positive effect during the night hours. This observation may be due to the fact that most flows at the beginning and end of the working day are core-periphery commuting trips, or, in other words, trips between locations with high and low Airbnb prices.

Finally, despite the fact that the "resid-to-X" type of explanatory layers were statistically insignificant by themselves, the aggregated "resid_torest" layer can be seen to have a significant positive effect on the flows with a slight increase during the night and early morning hours.



(a) (b)

Figure 9-6: Population densities, betweenness centrality, residential-to-other, and Airbnb coefficients over time in (a) London (b) Rome

Turning our attention to Figure 9-7, we see that the average travel time has a negative effect as a whole similar to network distance. However, unlike the network distance coefficient, we observe time to have a *stronger* negative effect during night hours. This is due to the fact that despite drivers travelling, on average, longer distances during the night, they do it in *less* time. We also see a similar temporal profile of the correlation coefficient. To recall, the correlation here measures the cross-correlation between the time series of car arrivals (destinations) in the two nodes forming the dyadic relationship in question. Similar to average travel time, it

enters overall with a negative exponent, with a markedly stronger effect during night hours. The overall negative effect is expected since most flows are core-periphery or residential-business area commuting flows where temporal patterns of car arrivals have either a weak or negative correlations. The effect is stronger during the night hours when most flow activity is spatially more spread out with arrival correlations being even weaker.

Next, we also see an overall positive effect of the existence of bus connections between nodes on the OD flows in both cities, with a slight drop in the effect during the night hours. Interestingly, the subway relational layer affects flows positively in Rome while its effect is negative in London. A reasonable hypothesis explaining this is the fact that London has banned the entry of cars in its center whereas Rome has not. Since subways typically connect the city center with the periphery, hence a positive coefficient in Rome and a negative one in London.

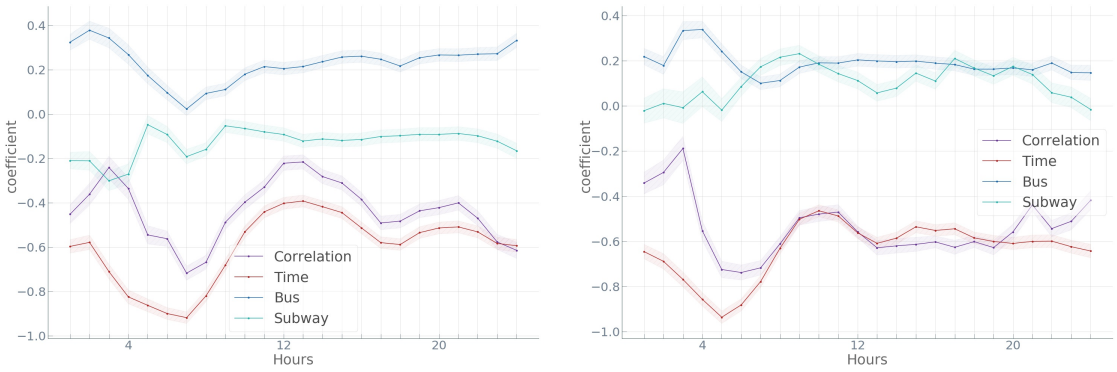


Figure 9-7: Time, correlation, subway, and bus coefficients over time in (a) London (b) Rome

We continue by discussing the effect of the APA centrality measures on the urban mobility flows in Figure 9-8. Our first observation is that the effect of all three measures is stronger in Rome compared to London. This raises the need for a deeper study of the relationship between network centrality measures and urban geography. Next, we note that despite this difference, the coefficients follow an overall similar temporal profile during the day. The influence of all measures in both cities display a sharp decrease during the night hours with two visible peaks during the morning and evening commute hours. We also note that food services

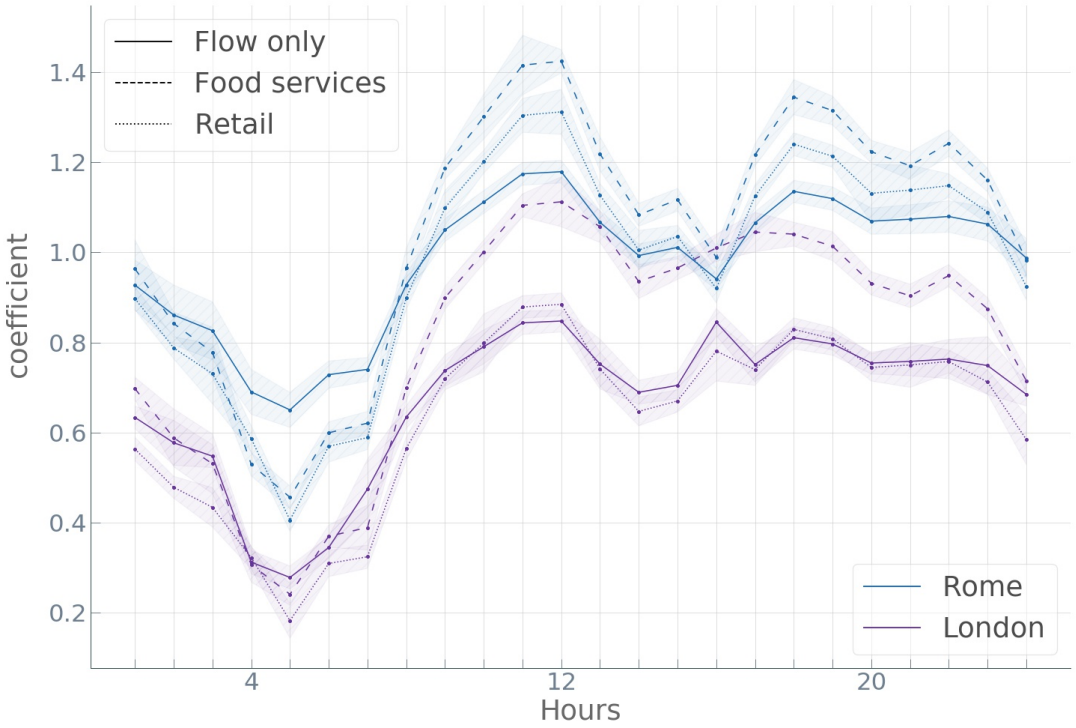


Figure 9-8: Flow only, food services, and retail APA centrality coefficients over time in (a) London (b) Rome

have the strongest effects in both cities, followed by retail activity.

Finally, we address the issue concerning the spatial resolution of the urban grid mentioned in Section 4.2. Since we do not have a theoretical apparatus to control for the effect of the grid resolution on the results of the regression, we took an empirical approach and tested our method for robustness in various grid resolution scenarios. All network layer parameters that were statistically significant under the default grid resolution of 500×500 meters turned out to be statistically significant under grid resolutions ranging from 250 to 2000 meters as well.

For example, Figure 9-9 shows the temporal behaviour of the parameter corresponding to the street network distance layer in London. As can be seen in the plot, the estimated parameter is remarkably robust to changes in scale ranging from 250 to 1500 meters, and visibly "breaks down" under grid sizes larger than that.

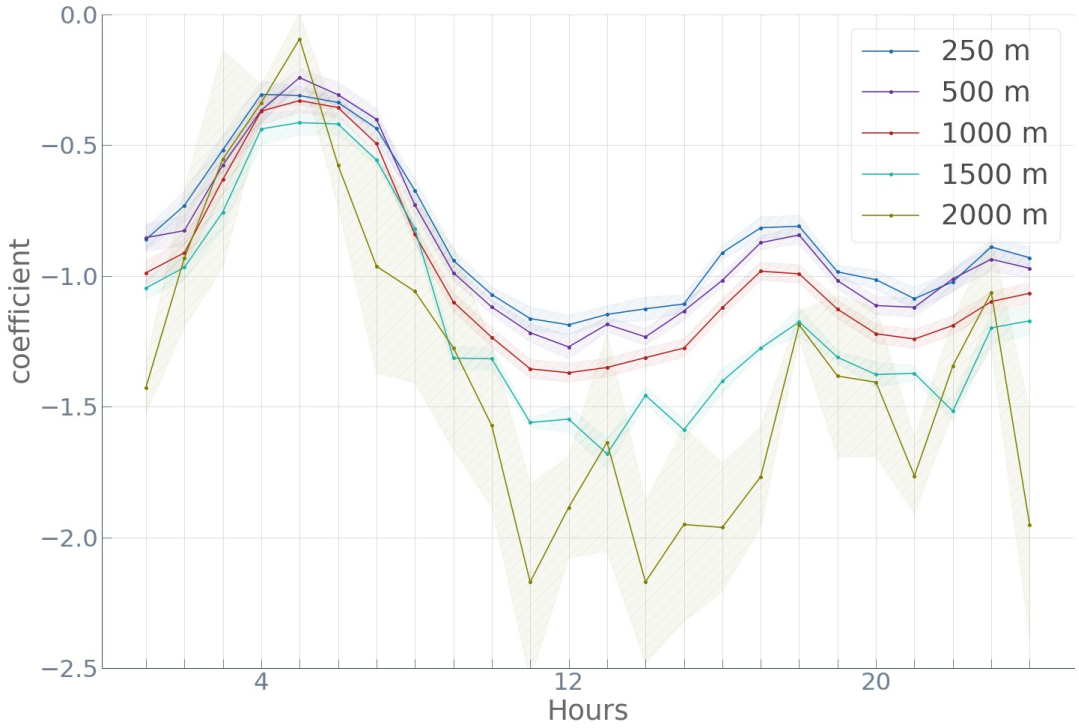


Figure 9-9: The network distance regression parameter over time under different spatial grid resolutions in London.

9.5 Discussion and Conclusions

In this Chapter, we built a framework for carrying out a regression analysis on multilayer networks, allowing to *explain* how each layer, built from the attributed urban network datasets (Chapter 4), contributes to explaining the observed urban mobility OD flows. The framework is based on the recently developed hypergeometric random graph ensemble (gHypE), has a closed form solution, and offers statistical model selection tools for testing the statistical significance of the parameters and for selecting the best possible model explaining the data.

We then conducted network regression for different hours of the day in both London and Rome, and obtained the temporal profiles of the estimated coefficients throughout the day.

We found interesting points of comparison between the two cities, offered possible explanations for the temporal behaviour of some coefficients, and noted the need for a deeper analysis of some of them. In particular, our findings regarding the drop in the effect of betweenness centrality during the night hours require further study as

its relationship to the spatial structure of urban mobility remains unclear. Another important direction of study for future work is the observation that the influence of APA centrality measures in Rome are significantly stronger compared to that of London.

A final important contribution to the proposed framework would be developing "local" spatial network regression models where specific origins or destinations are spatially selected to study how spatial interaction processes vary across geographic space ([99]). In other words, instead of the expected value obtained through MLE, a further enquiry into the spatial distribution of the estimated coefficients is required to better understand local relationships between urban socio-economic and spatial structure and mobility.

Having provided a general multilayer network regression framework for *explaining* urban mobility flows, we can now turn to the final Chapter in the present study, focusing on *prediction* of OD flows in an urban environment.

"We begin with the concept of Movement, which underlies all mechanisation..."

Sigfried Giedion, 1969

Chapter 10

Urban Graph Neural Networks

A fundamental problem of interest to policy makers, urban planners, and other stakeholders involved in urban development projects is assessing the impact of planning and construction activities on mobility flows. This is a challenging task due to the different spatial, temporal, social, and economic factors influencing urban mobility flows. In this Chapter, we address the problem of assessing origin-destination (OD) car flows between a location of interest and every other location in a city, given their attributes and the structural characteristics of the graph. We propose three neural network architectures, including graph neural networks (GNN), and conduct a systematic comparison between the proposed methods and state-of-the-art spatial interaction models, their modifications, and machine learning approaches. The objective of the Chapter is to address the practical problem of estimating potential flow between an urban development project location and other locations in the city, where the attributes of the project location are known in advance. We evaluate the performance of the models on a regression task using a custom data set of attributed car OD flows in London as described in Chapter 4.

This Chapter is a modified version of our paper Gevorg Yeghikyan, Felix L Opolka, Mirco Nanni, Bruno Lepri, and Pietro Lio. Learning mobility flows from urban features with spatial interaction models and neural networks. *arXiv preprint arXiv:2004.11924*, 2020 which is to appear in the Proceedings of IEEE International Conference on Smart Computing (SMARTCOMP 2020).

10.1 Introduction

So far, our urban mobility modelling attempts have been focused on essentially reconstructing OD flow matrices. However, modelling OD flow matrices in their entirety, the discussed approaches do not address the problem of assessing flows between a specific location and every other location in the city, given all other flows, other location characteristics, as well as information on the dyadic relations between those locations.

More specifically, the motivation for this task comes from a scenario in which it is necessary to assess the impact of an urban development project on the OD flows in and out of the project's location. Examples of these motivating scenarios include retail location choice and consumer spatial behaviour prediction, which have been approached with the Huff model and its modifications [129]. These models, however, suffer from a series of drawbacks related mostly to overly restrictive assumptions. We take a different approach and focus on the problem of evaluating OD flows in and out of a location of interest. By modelling urban flows as attributed graphs as described in Chapter 4, in which the nodes represent locations in a city (i.e. each node is described by a vector of features such as population density, Airbnb prices, available parking areas, etc.), and the edges represent the car flows between them (each one described by a vector of features such as road distance, average time required to travel, average speed, etc.), this Chapter aims to offer an instrument for assessing flows between a specific location and all other locations in the city. We have already discussed machine learning and neural network approaches to urban mobility flow prediction in Section 2.3.6 where we outlined existing literature and introduced graph neural networks, which the method proposed in this Chapter is based on.

Since a rigorous experimental setting would have required difficult-to-obtain longitudinal data of OD flows *before* and *after* the completion of an urban development project, we set up a *quasi-experimental* setting. We randomly select locations in a city and the flows associated with them as a test set, and attempt to find a function that takes the urban features describing city locations and the remaining flows as

input, and predicts the flows in the test set as output.

In this work, we define neural network models that make use of stationary node and edge features and compare different neural network architectures based on fully connected networks and graph neural networks.

In sum, our approach makes the following contributions:

- We propose three neural network architectures for predicting car flows between a location of interest and every other location in a city. Two of the models use graph convolutional layers that pool information from geographical or topological neighbourhoods around relevant nodes to incorporate more information (Section 10.4).
- We evaluate and compare our models on a custom dataset of aggregate OD car flows in London, containing node and edge features (Section 10.5).
- We show that the proposed neural network models outperform well-known spatial interaction and machine learning models. A comparison among neural network models reveals that graph convolutions do not substantially improve prediction performance on the formulated task (Sections 10.3, 10.5).
- We describe our custom dataset and make it publicly available along with the code for this study (Section 10.2).

10.2 Data description

To enhance our approach with results obtained in previous Chapters, we extend the urban mobility network data set described in Chapter 4 by including as node attributes the APA centrality values computed with the algorithms described in Chapter 8, and muliplex APA values from Chapter 7, respectively (Figure 10-1). However, in order to make sure the centrality values are not computed by including node information from our test set, we recompute both centrality measures by using only the partial network available in our training set.

To recall the general urban mobility network dataset structure, its main components are shown in Figure 10-2.

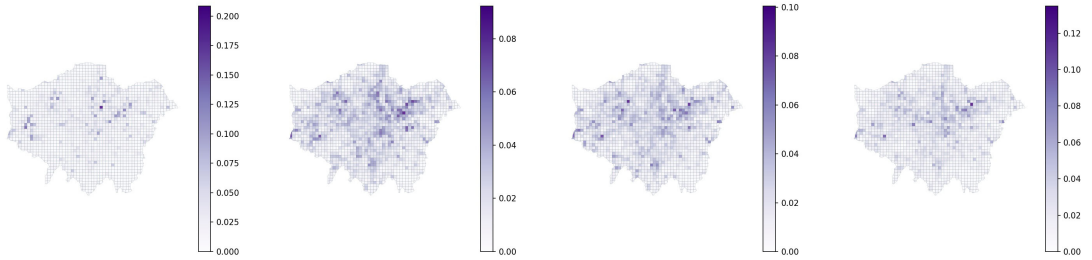
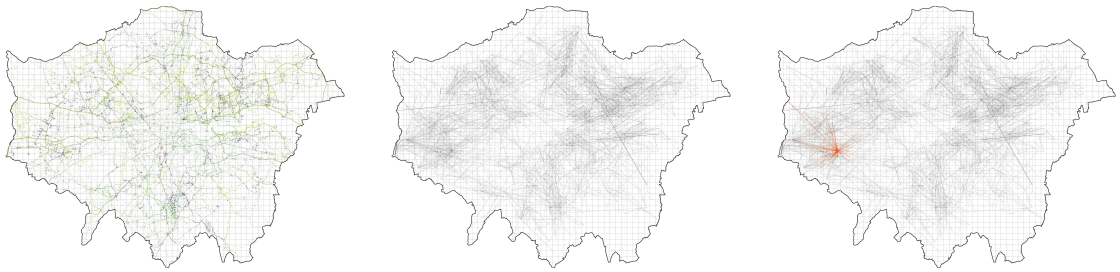


Figure 10-1: The APA centrality values for the mobility flow network in London at different times of the day.



(a) (b) (c)

Figure 10-2: (a) Car GPS trajectories over grid cells in London. (b) Origin-Destination (OD) flow network in London. (c) Target flows between a node of interest and every other node.

10.3 Problem statement

In this section, we describe the problem we are addressing and state definitions of important terms.

We define a **weighted attributed graph** $G = (\mathcal{V}, \mathcal{E}, \mathbf{W}, \mathbf{X}^v, \mathbf{X}^e)$ with feature information associated with both nodes and edges. More specifically, $\mathcal{V}$ is the set of n nodes, and $\mathcal{E} = \{e_{ij} = (i, j) : i, j \in \mathcal{V}\}$ represents the set of m edges in graph G. Furthermore, $\mathbf{W} \in \mathbb{R}^{n \times n}$ is the weighted adjacency matrix, essentially the OD matrix, with $\mathbf{W}_{ij} \geq 0 \ \forall i, j \in \mathcal{V}$ corresponding to the flow between cells i and j. Additionally, we denote the node feature matrix as $\mathbf{X}^v \in \mathbb{R}^{n \times p}$, where p is the number of node features. The edge feature matrix, on the other hand, is denoted as $\mathbf{X}^e \in \mathbb{R}^{m \times k}$, where k is the number of edge features.

The urban mobility flow network T is a weighted undirected attributed graph whose nodes are 500 $\times$ 500 m city grid cells, and the edges are the aggregate flows

between them. The nodes and edges are additionally augmented by feature vectors described in detail in Section 10.2. Furthermore, each edge e_{ij} in the urban mobility flow network T is associated with a target (or ground truth) flow w_{ij}, which is the corresponding entry in the weighted adjacency matrix $\boldsymbol{W}$ of T. It represents the aggregate mobility flow between cell (node) i and cell (node) j in the network.

In our prediction setting, we are given the urban mobility flow network $T = (\mathcal{V}, \mathcal{E}, \mathbf{W}, \mathbf{X}^v, \mathbf{X}^e)$ and a node of interest i for which the target flows $W_{i1}, \ldots, W_{in}$ are unknown. Hence, we aim to learn a mapping $f : \{\mathcal{V}, \mathcal{E}, \mathbf{W}, \mathbf{X}^v, \mathbf{X}^e\} \rightarrow \mathbb{R}^n$ from the urban mobility flow network to the missing flows, i.e. $[W_{i1}, \ldots, W_{in}] = f(i, \mathbf{W}, \mathbf{X}^v, \mathbf{X}^e)\ \forall i \in \mathcal{V}$. In other words, the aim is to predict the missing target flows (Figure 10-2c), given the features of node i and the rest of the graph.

10.4 Methodology

In the following, we describe three neural network models[1] that are trained to predict the unknown flows in the urban mobility flow network T. When a model makes a prediction for the flow associated with an edge going from a node of interest to another node in the graph, it can use all node and edge features in the graph, as these features are available even for nodes of interest, i.e. sites of prospective urban development projects. Furthermore, it may use the ground truth flows for edges that are not connected to a node of interest. In a practical situation, this corresponds to the flows between *existing* locations in the city for which flow information is therefore available.

The first neural network architecture is a fully connected neural network operating on the features of the target edge and the features of its two incident nodes. More specifically, when predicting the flow for target edge e_{ij}, we concatenate the node features $\boldsymbol{x}_i^v$ and $\boldsymbol{x}_j^v$ for incident node features, as well as the corresponding edge features $\boldsymbol{x}_{ij}^e$. The concatenated vector

$$\bar{\boldsymbol{x}} = [\boldsymbol{x}_i^v, \boldsymbol{x}_{ij}^e, \boldsymbol{x}_j^v] \tag{10.1}$$

[1] Code available at github.com/FelixOpolka/Mobility-Flows-Neural-Networks.

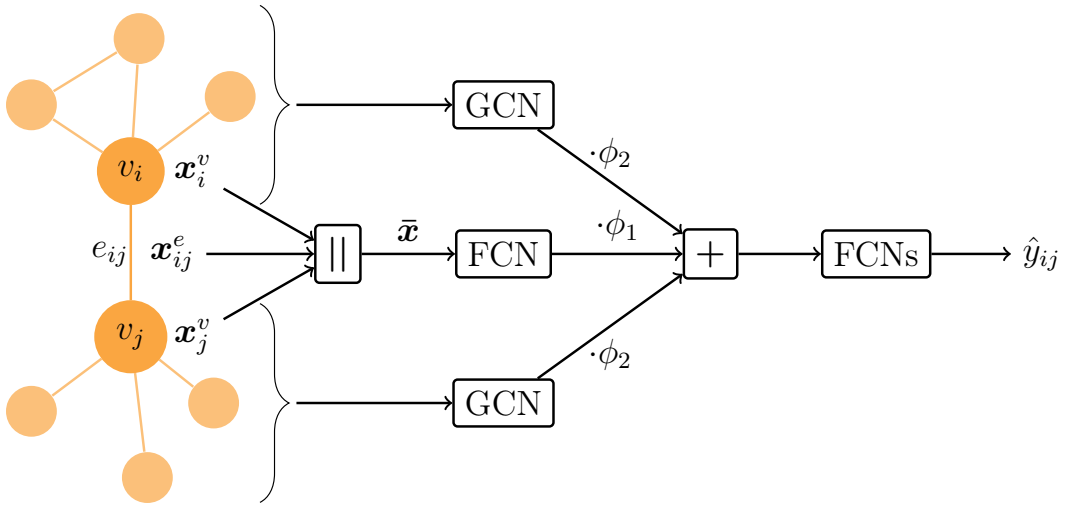


Figure 10-3: Overview of the neural network model architectures. When predicting the flow for edge e_{ij}, all three models concatenate the corresponding edge features $\boldsymbol{x}_{ij}^e$, and the node features $\boldsymbol{x}_i^v, \boldsymbol{x}_j^v$ of the incident nodes. The resulting vector is fed into a single fully connected layer. In case of the GNN-based models *GNN-geo* and *GNN-flow*, the network also perform graph convolutions on the neighbourhoods of v_i and v_j and computes a weighted sum of both neighbourhood embeddings and the edge embedding. A further set of fully connected layers maps the sum to the predicted flow $\hat{y}_{ij}$. The *FCNN* model skips the addition step and does not perform graph convolutions.

is passed into a fully connected neural network with ReLU-non-linearities, defined as $\text{ReLU}(z_j) = \max(0, z_j)$, where z_j is the j^{th} output of the linear transformation. Each fully connected layer is followed by batch normalisation [131] and dropout [245] to counter overfitting. We refer to this model as *FCNN*.

The second model builds upon the *FCNN* model through the additional use of graph convolutions to generate embeddings of node neighbourhoods. We use a graph convolutional neural network (GCN) [145] to generate node embeddings $\boldsymbol{h}_i, \boldsymbol{h}_j$ for the two nodes incident to the target edge e_{ij}. GCN layers extend fully-connected layers with an additional neighbourhood aggregation step before the non-linearity. The layer applies a linear transformation to all node features $\boldsymbol{h}_i^{(l-1)}$ in the graph and then, for each node, computes a weighted average of the resulting representations at the central node and in the 1-hop neighbourhood of the central node:

$$\boldsymbol{z}_i^{(l)} = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{\sqrt{(d_i + 1)(d_j + 1)}} \boldsymbol{h}_j^{(l-1)} \boldsymbol{\Theta}, \tag{10.2}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{D^{(l-1)} \times D^{(l)}}$ is a learned weight matrix, $\mathcal{N}(i)$ refers to the 1-hop neighbourhood of node i, and d_i denotes the degree of node i. This aggregation scheme is followed by a non-linearity and can be written more compactly using matrix multiplication as

$$\boldsymbol{H}^{(l)} = \mathrm{ReLU}(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l-1)}\boldsymbol{\Theta}). \tag{10.3}$$

where $\tilde{\boldsymbol{W}} = \boldsymbol{W} + \boldsymbol{I}$ and $\tilde{\boldsymbol{D}}$ is the degree matrix of $\tilde{\boldsymbol{W}}$. Equation 10.3 defines a graph convolutional layer and multiple such layers can be stacked to form a multi-layer graph neural network. A GNN with k layers allows us to compute embeddings encoding node feature information from within a k-hop neighbourhood.

For the second model, we apply multiple graph convolutions as defined above on the flow-weighted geographical adjacency matrix $\boldsymbol{W}^{\text{geo}}$ where W_{ij}^{geo} is non-zero if and only if node i is in the geographical neighbourhood of node j and $W_{ij}^{\text{geo}} = W_{ij}$, i.e. the flow between i and j. The resulting node embeddings $\boldsymbol{h}_i, \boldsymbol{h}_j \in \mathbb{R}^D$ for the two nodes incident to edge e_{ij} are added to the representation of $\bar{\boldsymbol{x}}$ (see Equation 10.1 after the first fully connected layer:

$$\boldsymbol{h}_{ij}^{(1)} = \phi_1 \mathrm{FCN}(\bar{\boldsymbol{x}}) + \phi_2 \left[\mathrm{GNN}(\boldsymbol{x}_i) + \mathrm{GNN}(\boldsymbol{x}_j)\right], \tag{10.4}$$

where ϕ_1, ϕ_2 are learned weighting coefficients. We note that both mentions of $\mathrm{GNN}(\cdot)$ refer to the same sequence of graph convolutional layers. We then feed $\boldsymbol{h}_{ij}^{(1)}$ into a number of fully connected layers, again with dropout and batch normalisation, such that the resulting model contains the same number of fully connected layers as the *FCNN* model. We call the resulting model *GNN-geo*.

Finally, we evaluate a third model, denoted by *GNN-flow*, which is equivalent to *GNN-geo* except graph convolutions are performed using the flow-based adjacency matrix $\boldsymbol{W}^{\text{flow}} = \boldsymbol{W}$, where W_{ij}^{flow} is the flow between i and j. Hence, the adjacency matrix used by *GNN-flow* will contain additional edges to those used by *GNN-geo*. A visualisation of the model architectures is given in Figure 10-3.

The graph based models *GNN-geo* and *GNN-flow* require flow information for the adjacency matrices. While this is readily available for edges between two regular nodes, we have to approximate flow between a regular node i and a node of interest

j. This is done by taking the average of the flows from node i to each node in the neighbourhood of j, i.e.

$$\tilde{W}_{ij} = \frac{1}{|\mathcal{N}(j)|} \sum_{k \in \mathcal{N}(j)} W_{ik}. \tag{10.5}$$

We note that even though the *FCNN* does not use graph convolutions and hence does not qualify as a common graph neural network, it does use graph structure information by concatenating specifically the features $\boldsymbol{x}_i^v, \boldsymbol{x}_j^v$ of the nodes incident to the target edge e_{ij}.

All models output the flow corresponding to the target edge e_{ij} and are trained to minimise the mean squared error between the predicted and the actual flow. More details on the experimental setup are provided in Section 10.5.3.

10.5 Experiments

We evaluate the described model on the London dataset described in Section 10.2. In the following, we describe the goodness-of-fit metrics we use to measure model performance, the baseline methods we compare our models to, and the experimental setup.

10.5.1 Goodness-of-fit measures

Mean absolute error (MAE). Let $\hat{y}_{ij}$ be the predicted flow between i and j, y_{ij} be the ground truth flow, then

$$\text{MAE} = \frac{1}{|\mathcal{E}|} \sum_i \sum_j |y_{ij} - \hat{y}_{ij}| \,. \tag{10.6}$$

Binned MAE. Due to the highly skewed distribution of the flow data, the vast majority of flows have a small flow count, with only a handful of flows with a very large flow value (see Figure 4-7b). Because of this, the total MAE will be biased downwards. To account for this, we additionally measure the MAE of all models within 4 bins with the following boundaries: $0 \leq 10.0 \leq 100.0 \leq 1000.0 \leq 10000.0$, corresponding to MAE_0, MAE_1, MAE_2, MAE_3, respectively. Finally, we define the

MAE bin mean as

$$\text{Bin mean MAE} = \frac{\text{MAE}_0 + \text{MAE}_1 + \text{MAE}_2 + \text{MAE}_3}{4}, \tag{10.7}$$

where MAE_i refers to MAE of the i^{th} bin.

Mean absolute percentage error (MAPE). To display the model accuracy with respect to the ground-truth flow values, we further use the mean absolute percentage error, defined as

$$\text{MAPE} = 100 \times \frac{1}{|\mathcal{E}|} \sum_i \sum_j \left| \frac{y_{ij} - \hat{y}_{ij}}{y_{ij}} \right|, \tag{10.8}$$

Sorensen similarity index. We use a modified version of the Sorensen similarity index (SSI), which has been extensively used in spatial interaction modelling [288, 162], and is defined as

$$SSI = \frac{1}{|\mathcal{E}|} \sum_i \sum_j \frac{2 \min(y_{ij}, \hat{y}_{ij})}{y_{ij} + \hat{y}_{ij}}, \tag{10.9}$$

and takes on values between 0 and 1, with values closer to 1 denoting a better fit.

Common part of commuters. Further, we use a similar metric, the common part of commuters, used specifically for mobility OD flow networks [162]:

$$CPC = \frac{2 \sum_{i,j=1}^{n} \min(y_{ij}, \hat{y}_{ij})}{\sum_{i,j=1}^{n} y_{ij} + \sum_{i,j=1}^{n} \hat{y}_{ij}}. \tag{10.10}$$

This measure takes on the value 0, when the flows in the two networks completely differ, and 1, when they are in perfect agreement.

Common part of links. Finally, to measure the degree to which the topological structure of the original network has been reconstructed, we use the common part of links (CPL) [161] defined as

$$CPL = \frac{2 \sum_{i,j=1}^{n} \mathbb{1}_{y_{ij}>0} \cdot \mathbb{1}_{\hat{y}_{ij}>0}}{\sum_{i,j=1}^{n} \mathbb{1}_{y_{ij}>0} + \sum_{i,j=1}^{n} \mathbb{1}_{\hat{y}_{ij}>0}}, \tag{10.11}$$

where $\mathbb{1}_A$ is the indicator function of condition A. The common part of links shows

the proportion of links between the observed and predicted networks such that $y_{ij} > 0$ and $\hat{y}_{ij} > 0$. It takes on the value zero if the two networks have no common links and one if the networks are topologically equivalent.

10.5.2 Baseline models

In this study, we compare the proposed model to the following baselines, using the same experimental setup for all models:

- **Doubly constrained gravity model (DC-GM)**: The classical gravity model with a power law decay has several formulations with respect to preserving the total in- nor out-flows during model calibration: unconstrained, origin-constrained, destination-constrained, and doubly constrained. Here we take the latter.
- **Huff model**: A probabilistic formulation of the gravity model described in Section 2.3.5.
- **Poisson regression**: An instance of the Generalized Linear Modelling framework, in which the dependent variable, being count data, is assumed to be drawn from a Poisson distribution.
- **Negative Binomial regression (NB)**: A generalization of the Poisson regression in which the restrictive assumption that the mean and the variance of the dependent variable are equal is loosened.
- **Spatial Autoregressive Model (SAM)**: An extension to the Generalized Linear Modelling framework by accounting for spatial dependence among the flows by using spatial lags represented by spatial weight matrices built from observed data [166].
- **Generalised hypergeometric ensemble multilayer network regression (gHypE)**: This recent random graph approach [63] provides a statistical ensemble of all possible flow networks under the constraints of preserving in- and out-flows from each node, as well as respecting pairwise flow propensities of

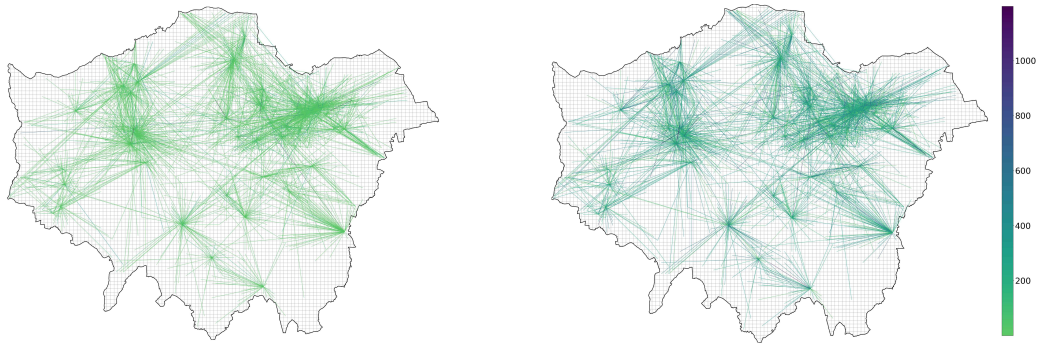
nodes. The multilayer network regression considers these propensities as latent variables, inferred from the edge features describing the dyadic relations between city locations. As opposed to conventional regression methods, this method intrinsically respects the network constraints.

- **Random Forest regression (RF)**: We follow the approach proposed in [244] aimed at predicting inter-city mobility flows with a set of attributes describing each city. We adapt the same approach to our problem of intra-city flow prediction. Following the described method, we use a Random Forest approach with eXtreme Gradient Boosting (*XGBoost*) [70] through 5-fold cross-validation, model and feature selection, and hyperparameter tuning.

10.5.3 Experimental setup

For training and evaluating the three proposed models, we divide the dataset into a training, validation, and test set of edges. The subsets contain 70%, 10%, and 20% of the edges respectively. To construct the test set, we randomly select nodes in the graph and add their incident edges to the test set. We ensure that an equal number of edges fall in each of the four bins split by flow magnitude. Hence, once a bin is full, no more edges are added to the test set that would fall into this bin. We use the same procedure to construct the validation set. Nodes in the validation and test set are considered nodes of interest, while nodes in the training set are considered regular nodes.

We train all models on the same training set. To address the imbalance between flows of different magnitude, we resample the data such that each bin contains the same number of samples. We perform hyperparameter search to determine the optimal dimension of the intermediate representations, i.e. the outputs of the GCN and fully connected layers, the dropout rate, and the number of fully connected and GCN layers. We select models based on the bin mean MAE (see Equation 10.7) achieved on the validation set. The selected models have a total of four fully connected layers. The GNN-based models use a single GCN layer. We use a dimensionality of 32 for intermediate representations and the dropout rate is set to 0.5.



(a) (b)

Figure 10-4: MAE residuals of flows associated with test nodes (a) *GNN-geo*. (b) *XGBoost*.

MAE	Total	$[0; 10)$	$[10; 10^2)$	$[10^2; 10^3)$	$[10^3; 10^4)$	bin mean
DC-GM	167.58	64.88	170.45	881.98	2176.35	823.42
Huff	122.89	48.21	99.86	511.41	1476.72	534.05
Poisson	106.74	40.69	88.56	475.23	1261.41	466.47
NB	92.62	33.02	76.96	431.44	1087.12	407.14
SAM	75.09	19.31	61.53	395.01	989.30	366.29
gHypE	58.11	9.02	53.10	346.96	832.26	310.34
XGBoost	31.59 ± 5.88	2.61 ± 0.89	45.12 ± 11.06	228.96 ± 39.96	549.83 ± 84.79	206.63 ± 34.18
FCNN	12.55 ± 0.91	0.33 ± 0.08	28.97 ± 4.93	161.12 ± 22.36	408.88 ± 36.59	149.82 ± 13.65
GNN-geo	13.34 ± 2.51	0.52 ± 0.40	31.63 ± 9.68	161.32 ± 9.09	422.04 ± 25.70	153.88 ± 9.74
GNN-flow	15.35 ± 4.23	0.63 ± 0.62	38.66 ± 16.65	170.06 ± 17.41	458.05 ± 64.56	166.85 ± 16.39
GNN-APA	9.51 ± 0.43	0.24 ± 0.05	20.26 ± 3.75	152.09 ± 7.14	399.90 ± 19.63	143.12 ± 6.88
GNN-APA-mpx	10.05 ± 0.51	0.28 ± 0.06	23.31 ± 4.30	155.89 ± 8.04	406.01 ± 21.12	146.37 ± 8.04

Table 10.1: Comparison of model performance in terms of mean absolute error grouped by flow magnitude.

We train for a total of 110 epochs using the Adam optimiser [144] with a batch size of 256 and a learning rate of 0.01. We reduce the learning rate by a factor of ten after 50 epochs and every 15 epochs after that. We stop training early once the performance of the model does no longer improve in terms of bin mean MAE on the validation set.

We have also experimented with using different types of graph neural network layers including GAT layers [262], GIN layers [285], and Jumping Knowledge layers [286]. We did not find these layers to improve performance on the validation data set and hence preferred the conceptually simpler GCN layers.

	SSI	MAPE $[10^3; 10^4)$	CPL	CPC
DC-GM	0.39	162.59	0.38	0.49
Huff	0.48	106.91	0.56	0.54
Poisson	0.46	102.10	0.57	0.54
NB	0.54	91.03	0.62	0.56
SAM	0.59	66.65	0.68	0.58
gHypE	0.62	52.99	0.79	0.60
XGBoost	0.67 ± 0.02	40.90 ± 5.85	0.86 ± 0.02	0.61 ± 0.01
FCNN	0.71 ± 0.00	27.16 ± 2.23	1.0 ± 0.00	0.69 ± 0.01
GNN-geo	0.70 ± 0.01	27.06 ± 1.65	1.0 ± 0.00	0.68 ± 0.04
GNN-flow	0.71 ± 0.02	30.67 ± 4.18	1.0 ± 0.01	0.65 ± 0.05

Table 10.2: Comparison of model performance in terms of MAPE, SSI, CPL, and CPC.

10.6 Results

We compare our models to the baseline ones in terms of MAE in Table 10.1. We find that all three neural network models outperform all the spatial interaction models (*DC-GM*, *Huff*, *Poisson*, *NB*, *SAM*) as well as *gHypE* and *XGBoost* in terms of total MAE by a large margin. Crucially, the MAEs per bin reveal that the neural network models achieve high accuracy across bins relative to the magnitude of flows, hence the neural network does not only perform well on small flows, which are highly overrepresented in the dataset.

We extend our analysis by creating two additional models corresponding to the node attributes augmented with the APA and multiplex APA centrality values, respectively, as described in Section 10.2. We see from Table 10.1 that including the APA centrality values improves overall model performance on the MAE metric across all bins.

We also observe that there is no clear difference in the performance between the three neural network based models. Surprisingly, the graph neural networks (*GNN-geo*, *GNN-flow*) do not outperform the fully connected neural network *FCNN*. This indicates that node neighbourhood information does not result in stronger predictive performance for this dataset and prediction task. We stress, however, that while *FCNN* is not a graph neural network, it does use graph structural information by

concatenating edge features with features of incident nodes. Furthermore, previous work on mobility flow prediction has omitted an explicit comparison of GNNs to fully connected neural networks, hence it remains unclear whether GNNs offer a predictive advantage in the urban mobility setting.

Finally, we compare the neural network models to the baselines in terms of SSI, MAPE of the largest bin, CPL, and CPC. These results also confirm that the neural network models find a better fit to the data compared to the state-of-the-art.

To further illustrate the effectiveness of the GNN models, we represent the MAE residuals on the London diagrammatic maps in Figure 10-4. These representations show the difference between predicted and ground-truth flows between the locations in the test set. We compare the state-of-the-art *XGBoost* model with our *GNN-flow* model and observe that the latter results in spatially smoother residuals.

10.7 Conclusion

In this Chapter, we formulated and addressed the problem of learning urban mobility flows between a location of interest and every other location in the city, given the array of socio-economic and structural features describing each location and the pairwise dyadic relations between them. We proposed three novel neural network architectures, using fully connected and graph convolutional layers, and compared them to a set of strong baseline models. We find that the neural network models achieve state-of-the-art performance and outperform the baselines by a large margin.

In fulfilment of the stated objective, our work has direct utility to urban planners and policy makers in offering a technique for assessing mobility flows between an urban development project location and other locations in the city.

"If you optimize everything, you will always be unhappy."

Donald Knuth

Chapter 11

Conclusion

11.1 Summary and conclusions

In this final Chapter, we summarise the results, discuss the limitations of our work, and sketch an outlook on future work. We began this thesis by introducing in Chapters 1, 2, and 3 the overarching theme and motivation for our research: that of finding valuable knowledge about the complex relationships between urban ***socio-economic structure*** and ***mobility***. To frame this task in a meaningful system for jointly representing urban structure and mobility, we described the city as a network in which the nodes represented urban locations (Cartesian grid cells of different spatial resolution) and the edges represented the OD mobility flows between them.

We then augmented the network with an array of socio-economic attributes describing the network nodes and edges. We described this process of building the attributed urban mobility network dataset in Chapter 4.

In Chapter 5, we introduced new network centrality measures based on Google's PageRank - the Adapted PageRank Algorithm (APA) - which allowed the incorporation of node attribute data in the computation of the centrality values. The proposed centrality measures also provided the possibility to assign the relative importance of attribute data in relation to the network topology when computing node rankings. This possibility, controlled via a parameter, offers the urban researcher with ample flexibility in studying urban locations through network centrality, given specific socio-economic attributes of interest to the researcher. Despite the fruit-

ful application of the developed centrality measures to urban mobility networks in Rome and London in subsequent chapters, a major theoretical shortcoming is the fact that we have not explored the behaviour of the vector of centrality values as a function of the mentioned parameter in order to gain a deeper understanding of the interplay between the network topology and the node attribute data. We intend to formulate this interesting question as future work.

In Chapters 6 and 7, we extended the APA centrality measures to Biplex and Multiplex networks to accommodate more types of dyadic relations between city locations treating them as additional layers in the complex urban networks. Being based on the previously introduced APA, these measures inherited its properties, notably the possibility to assign the relative importance of attribute data in each network layer. We applied the multiplex network centrality measures on the Rome urban mobility network and discussed the effects that different parametrisations of giving different importance to data in each network layer has on the spatial distribution of important nodes in the city. An interesting direction for future work in this regard would be a deeper study of how different layers affect the APA centrality in such a multilayer network setting.

Chapter 8 focused on unraveling the spatio-temporal behaviour of the most central nodes corresponding to different types of socio-economic activity in the city. We did this by computing the APA centrality values for flow only, food, and retail activities across the hours of the day, and across the days of the week in Rome and London. We identified the "hotspots" with high centrality values via a robust empirical method, and proposed simple metrics capturing the spatio-temporal structure of the studied socio-economic activities in both cities. We compared the results and findings in Rome and London, and formulated hypotheses offering possible explanations, which, however, will have to be studied further to get a more holistic picture of the spatio-temporal patterns of APA centrality in urban networks. In particular, the proposed Gini and TSI metrics fail to explain what share the spatial inequality and spread are due to core-periphery and inter-peripherial urban flows. As future work, we also intend to study the possible existence of a hierarchy of "hotspots" and its evolution over time. We see this application as a critical tool for monitoring

urban mobility and informing urban planning decisions. To that end, we see the necessity to develop and test a methodology for using the proposed measures as monitoring and policy informing tools. In particular, defining low, normal, high, or critical values of the proposed measures as simple indicators for urban planners to take action, should be formulated as part of methodological research in urban planning.

In Chapter 9, we developed a multilayer network regression model in which the OD flow layer is regressed on the other layers representing socio-economic relations between city locations. Within this framework, the observed OD flow network is modelled as a realisation from a recently developed family of statistical random graphs. This approach allowed to respect the network structure of the mobility flows - a major shortcoming of existing models - and to carry out statistical maximum likelihood estimation of parameters *explaining* the effect of each network layer on the observed OD flows. We saw how the dyadic relational layers built from APA centrality values obtained from our work in Chapter 5 contributed to this task. We conducted the network regression over the hours of a typical day, observing interesting similarities and differences in the temporal profiles of the obtained parameters between Rome and London. We see an important direction for future work in developing "local" network regression models to more accurately capture the variation of the regression parameters in geographical space. This is a crucial aspect since urban systems have high spatial and directional variation on different spatial scales, replacing and aggregating which with an average parameter valid for the city as a whole fails to capture the intricate local relationships so important in cities.

Finally, the contribution of Chapter 10 is twofold. On the one hand, it proposes several neural network architectures, including Graph Neural Networks (GNN), for *predicting* urban mobility flows by learning a mapping from the given OD flow network and its node and edge attributes to the missing flows in the OD network. On the other hand, it offers a solution to a practical problem arising in urban planning: that of assessing (missing) mobility flows to or from an urban development project in a particular city location, given its socio-economic attributes and the rest of the mobility flow network. The proposed neural network architectures preformed at

least an order of magnitude better compared to classical human mobility models and significantly better than more recent machine learning models. An interesting direction to develop this result in the future is enquiring into the reasons for why graph convolutions did not seem to offer a statistically significant improvement over simpler graph neural networks, despite their recent success in many other fields. In particular, we see a starting point in generating synthetic urban OD flow networks with any of the discussed mobility models and, being in control of the data generating mechanism, running experiments with various network sizes, parameters, and architectures.

11.2 Further research questions

In the present work, we touched upon but only barely scratched the surface of the relationship of urban spatial structure, socio-economic characteristics, and mobility. Naturally, enormous space for further research has been opened up, to which we deem important to devote a short discussion.

A general direction for future work concerning all proposed and discussed methods and techniques has to do with the urban grid resolution at the core of building the urban OD flow networks. Notwithstanding the few empirical robustness tests we carried out, the effect of the grid resolution on the OD network properties, APA centrality, multilayer network regression, and Graph Neural Networks remains to be explored and studied from a rigorous theoretical standpoint. For instance, the dependence of such an important indicator of urban mobility as the probability distribution characteristics on the spatial grid resolution is not at all trivial. Given the incredible richness and variability of spatial information in our cities, it is imperative to be careful and considerate with the spatial aggregation units we choose to study the urban environment. Finding the spatial resolution most suitable to the particular study at hand can be thought of as a problem of bias vs. variance tradeoff. Pertaining to the field of spatial statistics and to the Modifiable Aerial Unit Problem in particular, this question has important implications for studying all kinds of spatial phenomena dealing with spatial aggregations. However, it is

particularly important for urban data science, which, as mentioned in Section 1, requires rigorous theoretical and methodological foundations and dialogues with the research fields it borrows from and, by implication, heavily relies on.

Another interesting research question I intend to address in future work is the integration of qualitatively different data into the methodology proposed in this thesis. In particular, semantic information from geo-located social networks such as Twitter or Foursquare could prove very useful in enriching our network model of urban mobility and capture a qualitatively new dimension of social activity and its relationship to urban mobility in cities. First important steps in modelling social phenomena in a spatial context in cities have been undertaken [150, 14]. However, the link between the semantic dimension of social activity and urban mobility is still terra incognita and I see the network-theoretical modelling framework proposed in this thesis as a viable possibility to explore and incorporate this new dimension in urban data science.

In this thesis, urban mobility was primarily represented by *collective* mobility patterns aggregated in space and time. An important direction of research worth examining is the role and impact of *individual* mobility patterns [226, 52] on urban spatial structure and dynamics. Although our approach allows to integrate some indices like the radius of gyration capturing individual mobility into the network-based urban mobility framework presented in this thesis, further studies are necessary to disentangle the impact individual mobility choices and preferences have on urban land use, real estate prices, accessibility, and other urban socio-economic characteristics. This becomes especially important in the wake of the mass introduction of electric and autonomous vehicles in cities, since the individual and collective urban mobility patterns and their impact on urban economies, as well as the challenges posed to urban planning require urgent and thorough analysis [109].

A research question particularly interesting for urban management and policy making is the identification of urban functional zones as determined by the *actual* activity clusters in cities. Identifying such functional zones as shopping, working, entertainment, etc., can greatly inform and enhance administrative divisions in cities, improve the organization of various city services such as waste collection, repair

works, as well as determine catchment areas for retailers or government services. Existing work in this direction mostly utilises techniques from spatial statistics and conventional hotspot analysis [296, 294]. However, I see urban mobility networks as described and formulated in this thesis and network community detection techniques from network science as tools with great great potential to tap into this research question. A first step in this direction has been made in [107] in which the authors use network community detection algorithms for identifying densely connected pockets in inter-city mobility networks. Developing and applying network community detection techniques to intra-urban mobility flow networks evolving over time can offer great insights into the dynamics of urban functional zones. In fact, preliminary experiments on both London and Rome mobility network datasets built and described in Chapter 4 have shown promising results.

These are but some of the research questions this PhD thesis has opened up and provided the methodology and techniques to explore. Offering a method for building very specific and detailed urban models, the attributed urban mobility network approach presented in this thesis provides a fairly generic approach for addressing a wide range of other questions in urban data science, urban planning and design, as well as urban management and policy making. I conclude the present PhD thesis with the intention to delve deeper into the mentioned directions, discover new avenues of research, and contribute to urban data science.

Glossary

CNR National Research Council. 14

IMT IMT School for Advanced Studies Lucca. 14

OD Origin-Destination flow. 23

SNS Scuola Normale Superiore of Pisa. 14

SSSA Sant'Anna School of Advanced Studies. 14

UniPi University of Pisa. 14

Bibliography

[1] Taras Agryzcov, Leandro Tortosa, José F Vicent, and Richard Wilson. A centrality measure for urban networks based on the eigenvector centrality concept. *Environment and Planning B: Urban Analytics and City Science*, page 2399808317724444, 2017.

[2] T. Agryzkov, J.L. Oliver, L. Tortosa, and J.F. Vicent. An algorithm for ranking the nodes of an urban network based on the concept of pagerank vector. *Applied Mathematics and Computation*, 219:2186–2193, 2012.

[3] T. Agryzkov, F. Pedroche, L. Tortosa, and J.F. Vicent. Combining the two-layers pagerank approach with the apa centrality in networks with data. *International Journal of Geo-Information*, 7(480), 2018.

[4] T. Agryzkov, L. Tortosa, and J.F. Vicent. New highlights and a new centrality measure based on the adapted pagerank algorithm for urban networks. *Applied Mathematics and Computation*, 291(C):14–29, 2016.

[5] Taras Agryzkov, Manuel Curado, Leandro Tortosa, and Jose F Vicent. Extending the adapted pagerank algorithm centrality to multiplex networks with data using the pagerank two-layer approach. *Symmetry*, 11(2):284, 2019.

[6] Taras Agryzkov, Jose L Oliver, Leandro Tortosa, and Jose F Vicent. An algorithm for ranking the nodes of an urban network based on the concept of pagerank vector. *Applied Mathematics and Computation*, 219(4):2186–2193, 2012.

[7] Taras Agryzkov, José L Oliver, Leandro Tortosa, and José F Vicent. An algorithm for ranking the nodes of an urban network based on the concept of pagerank vector. *Applied Mathematics and Computation*, 219(4):2186–2193, 2012.

[8] Taras Agryzkov, José L Oliver, Leandro Tortosa, and José F Vicent. A new betweenness centrality measure based on an algorithm for ranking the nodes of a network. *Applied Mathematics and Computation*, 244:467–478, 2014.

[9] Taras Agryzkov, Leandro Tortosa, and José F Vicent. New highlights and a new centrality measure based on the adapted pagerank algorithm for urban networks. *Applied Mathematics and Computation*, 291:14–29, 2016.

[10] Taras Agryzkov, Leandro Tortosa, and Jose F Vicent. New highlights and a new centrality measure based on the adapted pagerank algorithm for urban networks. *Applied Mathematics and Computation*, 291:14–29, 2016.

[11] Taras Agryzkov, Leandro Tortosa, José F Vicent, and Richard Wilson. A centrality measure for urban networks based on the eigenvector centrality concept. *Environment and Planning B: Urban Analytics and City Science*, pages 239–, 2017.

[12] Rein Ahas, Anto Aasa, Siiri Silm, and Margus Tiru. Daily rhythms of suburban commuters' movements in the tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies*, 18(1):45–54, 2010.

[13] Nobbir Ahmed and Harvey J Miller. Time–space transformations of geographic space for exploring, analyzing and visualizing transportation systems. *Journal of Transport Geography*, 15(1):2–17, 2007.

[14] Fritz Akhmad Nuzir, Bart Julien Dewancker, et al. Dynamic land-use map based on twitter data. *Sustainability*, 9(12):2158, 2017.

[15] Réka Albert. R. albert and a.-l. barabási, rev. mod. phys. 74, 47 (2002). *Rev. Mod. Phys.*, 74:47, 2002.

[16] Jeff Alstott and Dietmar Plenz Bullmore. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1), 2014.

[17] Yaniv Altshuler, Rami Puzis, Yuval Elovici, Shlomo Bekhor, and AS Pentland. Augmented betweenness centrality for mobility prediction in transportation networks. In *International Workshop on Finding Patterns of Human Behaviors in NEtworks and MObility Data, NEMO11*, 2011.

[18] Alex Anas, Richard Arnott, and Kenneth A Small. Urban spatial structure. *Journal of economic literature*, 36(3):1426–1464, 1998.

[19] Luc Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115, 1995.

[20] Konstantin Avrachenkov, Nelly Litvak, Vasily Medyanikov, and Marina Sokol. Alpha current flow betweenness centrality. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 106–117. Springer, 2013.

[21] James P Bagrow and Yu-Ru Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5), 2012.

[22] Albert L. Barabasi and Jennifer Frangos. *Linked: the new science of networks,*. Basic Books,, 2014.

[23] Albert-László Barabási. A.-l. barabási and r. albert, science 286, 509 (1999). *Science*, 286:509, 1999.

[24] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[25] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.

[26] Marc Barthelemy. *The structure and dynamics of cities*. Cambridge University Press, 2016.

[27] Marc Barthelemy. *Morphogenesis of Spatial Networks*. Springer, 2018.

[28] Marc Barthelemy, Patricia Bordin, Henri Berestycki, and Maurizio Gribaudi. Self-organization versus top-down planning in the evolution of a city. *Scientific reports*, 3:2153, 2013.

[29] Michael Batty. Building a science of cities. *Cities*, 29:S9–S16, 2012.

[30] Michael Batty. *The new science of cities*. MIT press, 2013.

[31] Michael Batty. Urban analytics defined, 2019.

[32] Alex Bavelas. A mathematical model for group structures. *Applied anthropology*, 7(3):16–30, 1948.

[33] Alex Bavelas. Communication patterns in task-oriented groups. *The journal of the acoustical society of America*, 22(6):725–730, 1950.

[34] Oualid Benyahia and Christine Largeron. Centrality for graphs with numerical attributes. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1348–1353, 2015.

[35] M. Benzi and C. Klymko. On the limiting behavior of parameter-dependent network centrality measures. *SIAM Journal on Matrix Analysis and Applications*, 36(2):686–706, 2015.

[36] Pavel Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2(1):73–120, 2005.

[37] Pavel Berkhin. A survey on pagerank computing. *Internet mathematics*, 2(1):73–120, 2005.

[38] S Berroir, H Mathian, T Saint-Julien, and L Sanders. The role of mobility in the building of metropolitan polycentrism. *Modelling urban dynamics [Desrosiers, F. & Thériault, M.(eds)][1–25](ISTE-Wiley)*, 2011.

[39] Alain Bertaud and Stephen Malpezzi. The spatial distribution of population in 48 world cities: Implications for economies in transition. *Center for urban land economics research, University of Wisconsin*, 32(1):54–55, 2003.

[40] Luis Bettencourt and Geoffrey West. A unified theory of urban living. *Nature*, 467(7318):912–913, 2010.

[41] Luís MA Bettencourt. The origins of scaling in cities. *science*, 340(6139):1438–1441, 2013.

[42] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

[43] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

[44] G. Bianconi. *Multilayer networks. Structure and Functions*. Oxford University press, Oxford, UK, 2018.

[45] John Bingham-Hall and Stephen Law. Connected or informed?: Local twitter networking in a london neighbourhood. *Big Data & Society*, 2(2):2053951715597457, 2015.

[46] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[47] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin. Multilayer networks. *J. Complex Networks*, 2(3):203–271, 2014.

[48] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[49] P. Bonacich. Power and centrality: a family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1897.

[50] Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

[51] Moreno Bonaventura, Luca Maria Aiello, Daniele Quercia, and Vito Latora. Predicting urban innovation from the workforce mobility network in us. *arXiv preprint arXiv:1911.00436*, 2019.

[52] Agnese Bonavita, Riccardo Guidotti, and Mirco Nanni. Self-adapting trajectory segmentation.

[53] Federico Botta and Charo I del Genio. Analysis of the communities of an urban mobile phone network. *PloS one*, 12(3):e0174198, 2017.

[54] Enrico Bozzo and Massimo Franceschet. Resistance distance, closeness, and betweenness. *Social Networks*, 35(3):460–469, 2013.

[55] Anne Bretagnolle, Eric Daudé, and Denise Pumain. From theory to modelling: urban systems as complex systems. *Cybergeo: European Journal of Geography*, 2006.

[56] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[57] Giuseppe Bruno and Gennaro Improta. Using gravity models for the evaluation of new university site locations: A case study. *Computers & Operations Research*, 35(2):436–444, 2008.

[58] F. Caccioli, M. Shrestha, C. Moore, and J.D. Farmer. Stability analysis of financial contagion due to overlapping portfolios. *Journal of Banking & Finance*, 46:233–245, 2014.

[59] F. Calabrese, C. Ratti, M. Colonna, P. Lovisolo, and D. Parata. Real-time urban monitoring using cell phones: a case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 25(2):141–151, 2011.

[60] A Colin Cameron and Pravin K Trivedi. Essentials of count data regression. *A companion to theoretical econometrics*, 331, 2001.

[61] Alessio Cardillo, Salvatore Scellato, Vito Latora, and Sergio Porta. Structural properties of planar graphs of urban street patterns. *Physical Review E*, 73(6):066107, 2006.

[62] Alessio Cardillo, Massimiliano Zanin, Jesús Gómez-Gardenes, Miguel Romance, Alejandro J García del Amo, and Stefano Boccaletti. Modeling the multi-layer nature of the european air transport network: Resilience and passengers re-scheduling under random failures. *The European Physical Journal Special Topics*, 215(1):23–33, 2013.

[63] Giona Casiraghi. Multiplex network regression: How do relations drive interactions? *arXiv preprint arXiv:1702.02048*, 2017.

[64] Giona Casiraghi and Vahan Nanumyan. Generalised hypergeometric ensembles of random graphs: the configuration model as an urn problem. *arXiv preprint arXiv:1810.06495*, 2018.

[65] Oded Cats. Topological evolution of a metropolitan rail transport network: The case of stockholm. *Journal of Transport Geography*, 62:172–183, 2017.

[66] Di Chai, Leye Wang, and Qiang Yang. Bike flow prediction with multi-graph convolutional networks. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 397–400. ACM, 2018.

[67] Sonic HY Chan, Reik V Donner, and Stefan Lämmer. Urban road networks—spatial networks with universal geometric features? *The European Physical Journal B*, 84(4):563–577, 2011.

[68] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8 – 15, 2007.

[69] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. Learning deep representation from big and heterogeneous data for traffic accident inference. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[70] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[71] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145, 2002.

[72] L da F Costa, BAN Travencolo, MP Viana, and E Strano. On the efficiency of transportation systems in large cities. *EPL (Europhysics Letters)*, 91(1):18003, 2010.

[73] Thomas Courtat, Catherine Gloaguen, and Stephane Douady. Mathematics and morphogenesis of cities: A geometrical approach. *Physical Review E*, 83(3):036106, 2011.

[74] P. Crucitti, V. Latora, and S. Porta. The network analysis of urban streets: a dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2):853–866, 2006.

[75] P. Crucitti, V. Latora, and S. Porta. The network analysis of urban streets: a primal approach. *Planning and design*, 33(5):705–725, 2006.

[76] Paolo Crucitti, Vito Latora, and Sergio Porta. Centrality in networks of urban streets. *Chaos: an interdisciplinary journal of nonlinear science*, 16(1):015113, 2006.

[77] Paolo Crucitti, Vito Latora, and Sergio Porta. Centrality measures in spatial networks of urban streets. *Physical Review E*, 73(3):036125, 2006.

[78] Manuel Curado, Leandro Tortosa, Jose F Vicent, and Gevorg Yeghikyan. Analysis and comparison of centrality measures applied to urban networks with data. *Journal of Computational Science*, page 101127, 2020.

[79] Alan Ricardo da Silva and Thais Carvalho Valadares Rodrigues. Geographically weighted negative binomial regression—incorporating overdispersion. *Statistics and Computing*, 24(5):769–783, 2014.

[80] Peter Davis. Spatial competition in retail markets: movie theaters. *The RAND Journal of Economics*, 37(4):964–982, 2006.

[81] G.F. De Arruda, A.L. Barbieri, P.M. Rodríguez, F.A. Rodrigues, Y. Moreno, and L. da Fontoura Costa. Role of centrality for the identification of influential spreaders in complex networks. *Physical Review E*, 90(3):032812, 2014.

[82] Matthias De Beule, Dirk Van den Poel, and Nico Van de Weghe. An extended huff-model for robustly benchmarking and predicting retail network performance. *Applied Geography*, 46:80–89, 2014.

[83] Manlio De Domenico, Albert Solé, Sergio Gómez, and Alex Arenas. Random walks on multiplex networks. *arXiv preprint arXiv:1306.0519*, 2013.

[84] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A Porter, Sergio Gómez, and Alex Arenas. Mathematical formulation of multilayer networks. *Physical Review X*, 3(4):041022, 2013.

[85] Andrea De Montis, Marc Barthélemy, Alessandro Chessa, and Alessandro Vespignani. The structure of interurban traffic: a weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924, 2007.

[86] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc.

[87] Rui Ding. The complex network theory-based urban land-use and transport interaction studies. *Complexity*, 2019, 2019.

[88] Rui Ding, Norsidah Ujang, Hussain bin Hamid, Mohd Shahrudin Abd Manan, Yuou He, Rong Li, and Jianjun Wu. Detecting the urban traffic network structure dynamics through the growth and analysis of multi-layer networks. *Physica A: Statistical Mechanics and its Applications*, 503:800–817, 2018.

[89] Rui Ding, Norsidah Ujang, Hussain Bin Hamid, and Jianjun Wu. Complex network theory applied to the growth of kuala lumpur's public urban rail transit network. *PloS one*, 10(10):e0139961, 2015.

[90] César Ducruet. Multilayer dynamics of complex spatial networks: The case of global maritime flows (1977–2008). *Journal of Transport Geography*, 60:47–58, 2017.

[91] Nadav Eiron, Kevin S McCurley, and John A Tomlin. Ranking the web frontier. In *Proceedings of the 13th international conference on World Wide Web*, pages 309–318. ACM, 2004.

[92] Emu Analytics. DATAPACK MARKETPLACE retrieved from http://www.emu-analytics.com/products/datapacks.php . http://www.emu-analytics.com/products/datapacks.php, 2020.

[93] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[94] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 world wide web conference*, pages 1459–1468, 2018.

[95] Guilherme Ferraz de Arruda, André Luiz Barbieri, Pablo Martín Rodriguez, Yamir Moreno, Luciano da Fontoura Costa, and Francisco Aparecido Rodrigues. The role of centrality for the identification of influential spreaders in complex networks. *arXiv preprint arXiv:1404.4528*, 2014.

[96] Donald L Foley. The daily movement of population into central business districts. *American sociological review*, 17(5):538–543, 1952.

[97] Andrea Fontanari, Nassim Nicholas Taleb, and Pasquale Cirillo. Gini estimation under infinite variance. *Physica A: Statistical Mechanics and its Applications*, 502:256–269, 2018.

[98] A Stewart Fotheringham. A new set of spatial-interaction models: the theory of competing destinations. *Environment and Planning A: Economy and Space*, 15(1):15–36, 1983.

[99] A Stewart Fotheringham and Chris Brunsdon. Local forms of spatial analysis. *Geographical analysis*, 31(4):340–358, 1999.

[100] A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

[101] A Stewart Fotheringham and Morton E O'Kelly. *Spatial interaction models: formulations and applications*, volume 1. Kluwer Academic Publishers Dordrecht, 1989.

[102] L. Freeman. Centrality in social networks conceptual clarifications. *Social networks*, 1(3):215–239, 1972.

[103] L.C. Freeman. Centrality in social networks' conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[104] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[105] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[106] Linton C Freeman, Stephen P Borgatti, and Douglas R White. Centrality in valued graphs: A measure of betweenness based on network flow. *Social networks*, 13(2):141–154, 1991.

[107] Lorenzo Gabrielli, Daniele Fadda, Giulio Rossetti, Mirco Nanni, Leonardo Piccinini, Dino Pedreschi, Fosca Giannotti, and Patrizia Lattarulo. Discovering mobility functional areas: A mobility data analysis approach. In *International Workshop on Complex Networks*, pages 311–322. Springer, 2018.

[108] R. Gallotti and M. Barthelemy. Anatomy and efficiency of urban multimodal mobility. *Scientific reports*, 4:6911, 2014.

[109] Nikolaos Gavanas. Autonomous road vehicles: Challenges for urban planning in european cities. *Urban Science*, 3(2):61, 2019.

[110] Corrado Gini. Measurement of inequality of incomes. *The Economic Journal*, 31(121):124–126, 1921.

[111] Corrado Gini. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208:73–79, 1936.

[112] Segun Goh, MY Choi, Keumsook Lee, and Kyung-min Kim. How complexity emerges in urban systems: Theory of urban morphology. *Physical Review E*, 93(5):052309, 2016.

[113] Gomez-Gardenes J Perez-Vicente CJ Moreno Y Arenas A Gomez S, Diaz-Guilera A. Diffusion dynamics on multiplex networks. *Physical review letters*, 110(2):028701, 2013.

[114] Michael F Goodchild and Donald G Janelle. The city around the clock: Space—time patterns of urban ecological structure. *Environment and Planning A*, 16(6):807–820, 1984.

[115] The World Bank Group. Urban population (% of total), 2018.

[116] Marianne Guérois and Denise Pumain. Built-up encroachment and the urban field: a comparison of forty european cities. *Environment and Planning A*, 40(9):2186–2203, 2008.

[117] Jingyi Guo, Xianghua Li, Zili Zhang, and Junwei Zhang. Traffic flow fluctuation analysis based on beijing taxi gps data. In *International Conference on Knowledge Science, Engineering and Management*, pages 452–464. Springer, 2018.

[118] A. Halu, R.J. Mondragón, P. Panzarasa, and G. Bianconi. Multiplex pagerank. *PLoS ONE*, 8(e78293), 2013.

[119] Arda Halu, Raúl J Mondragón, Pietro Panzarasa, and Ginestra Bianconi. Multiplex pagerank. *PloS one*, 8(10), 2013.

[120] Walter G Hansen. How accessibility shapes land use. *Journal of the American Institute of planners*, 25(2):73–76, 1959.

[121] Taher Haveliwala, Sepandar Kamvar, Dan Klein, Chris Manning, and Gene Golub. Computing pagerank using power extrapolation. Technical report, Stanford, 2003.

[122] Taher H Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796, 2003.

[123] Jingrui He, Yan Liu, and Richard Lawrence. Graph-based transfer learning. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 937–946. ACM, 2009.

[124] Bill Hillier. A configurational theory of architecture, 1996.

[125] Bill Hillier and Julienne Hanson. *The social logic of space*. Cambridge university press, 1989.

[126] Petter Holme and Jari Saramaki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[127] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[128] Yujie Hu, Harvey J Miller, and Xiang Li. Detecting and analyzing mobility hotspots using surface networks. *Transactions in GIS*, 18(6):911–935, 2014.

[129] David L Huff. A probabilistic analysis of shopping center trade areas. *Land economics*, 39(1):81–90, 1963.

[130] J. Iacovacci and G. Bianconi. Extracting information from multiplex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 335(6071):065306, 2016.

[131] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

[132] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *Proceedings of 2013 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–274, 08 2013.

[133] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016.

[134] Junteng Jia and Austin R Benson. Detecting core-periphery structure in spatial networks. *arXiv preprint arXiv:1808.06544*, 2018.

[135] Yuhan Jia, Jianping Wu, and Yiman Du. Traffic speed prediction using deep learning method. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1217–1222. IEEE, 2016.

[136] Bin Jiang and Christophe Claramunt. Topological analysis of urban street networks. *Environment and Planning B: Planning and design*, 31(1):151–162, 2004.

[137] Andreas Justen, Francisco J Martínez, and Cristián E Cortés. The use of space–time constraints for the selection of discretionary activity locations. *Journal of Transport Geography*, 33:146–152, 2013.

[138] Vamsi Kalapala, Vishal Sanwalani, Aaron Clauset, and Cristopher Moore. Scale invariance in road networks. *Physical Review E*, 73(2):026130, 2006.

[139] Sepandar Kamvar, Taher Haveliwala, and Gene Golub. Adaptive methods for the computation of pagerank. *Linear Algebra and its Applications*, 386:51–65, 2004.

[140] Chaogui Kang, Xiujun Ma, Daoqin Tong, and Yu Liu. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, 391(4):1702–1717, 2012.

[141] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[142] H.A. Khanday, H. Ganaiand, and R. Hashmy. A comparative analysis of identifying influential users in online social networks. In *Proceedings of 2018 International Conference on Soft-computing and Network Security (ICSNS)*, pages 1–6, 02 2018.

[143] Jiwon Kim, Kai Zheng, Sanghyung Ahn, Marty Papamanolis, and Pingfu Chao. Graph-based analysis of city-wide traffic dynamics using time-evolving graphs of trajectory data. In *Australasian Transport Research Forum (ATRF), 38th*, 2016.

[144] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[145] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017.

[146] Alec Kirkley, Hugo Barbosa, Marc Barthelemy, and Gourab Ghoshal. From the betweenness centrality in street networks to structural invariants in random planar graphs. *Nature communications*, 9(1):1–12, 2018.

[147] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010.

[148] M. Kivela, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, and M.A. Porter. Multilayer networks. *J. Complex Networks*, 2(3):203–271, 2014.

[149] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[150] Andreas Koch and Peter Mandl. *Modeling social phenomena in spatial context*, volume 2. LIT Verlag Münster, 2013.

[151] David Krackhardt. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social networks*, 10(4):359–381, 1988.

[152] Stefan Lämmer, Björn Gehlsen, and Dirk Helbing. Scaling laws in the spatial structure of urban road networks. *Physica A: Statistical Mechanics and its Applications*, 363(1):89–95, 2006.

[153] Amy N Langville and Carl D Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[154] Amy N Langville and Carl D Meyer. A survey of eigenvector methods for web information retrieval. *SIAM review*, 47(1):135–161, 2005.

[155] Amy N Langville and Carl D Meyer. A survey of eigenvector methods for web information retrieval. *SIAM review*, 47(1):135–161, 2005.

[156] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701, 2001.

[157] Vito Latora and Massimo Marchiori. Vulnerability and protection of infrastructure networks. *Physical Review E*, 71(1):015103, 2005.

[158] Vito Latora and Massimo Marchiori. A measure of centrality based on network efficiency. *New Journal of Physics*, 9(6):188, 2007.

[159] Florent Le Néchet. Urban spatial structure, daily mobility and energy consumption: a study of 34 european cities. *Cybergeo: European Journal of Geography*, 2012.

[160] Jaekoo Lee, Hyunjae Kim, Jongsun Lee, and Sungroh Yoon. Transfer learning for deep learning on graph-structured data. In *AAAI*, pages 2154–2160, 2017.

[161] Maxime Lenormand, Aleix Bassolas, and José J Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–169, 2016.

[162] Maxime Lenormand, Sylvie Huet, Floriana Gargiulo, and Guillaume Deffuant. A universal model of commuting networks. *PloS one*, 7(10), 2012.

[163] Maxime Lenormand, Miguel Picornell, Oliva G Cantú-Ros, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, Maxi San Miguel, and José J Ramasco. Comparing and modelling land use organization in cities. *Royal Society open science*, 2(12):150449, 2015.

[164] Maxime Lenormand, Miguel Picornell, Oliva G Cantú-Ros, Antonia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frias-Martinez, and José J Ramasco. Cross-checking different sources of mobility information. *PLoS One*, 9(8):e105184, 2014.

[165] Picornell M. Cantú-Ros-O.G. Louail T.-Herranz R. Barthelemy M. Frías-Martínez E. San Miguel M. Lenormand, M. and J.J. Ramasco. Comparing and modelling land use organization in cities. *Royal Society open science*, 2(12):150449, 2015.

[166] James P LeSage and R Kelley Pace. Spatial econometric modeling of origin-destination flows. *Journal of Regional Science*, 48(5):941–967, 2008.

[167] Yingru Li and Lin Liu. Assessing the impact of retail location on store performance: A comparison of wal-mart and kmart stores in cincinnati. *Applied Geography*, 32(2):591–600, 2012.

[168] Xiao Liang, Jichang Zhao, Li Dong, and Ke Xu. Unraveling the origin of exponential law in intra-urban human mobility. *Scientific reports*, 3:2983, 2013.

[169] Moshe Lichman and Padhraic Smyth. Modeling human location data with mixtures of kernel densities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 35–44, 2014.

[170] Q. Liu, Z. Ma, X. Yangand, and H. Zhu. Identifying influential actors in social network platforms. In *Proceedings of 2015 IEEE 11th International Conference on Computational Intelligence and Security (CIS)*, pages 223–226, 02 2015.

[171] Mussone Lorenzo and Matteucci Matteo. Od matrices network estimation from link counts by neural networks. *Journal of Transportation Systems Engineering and Information Technology*, 13(4):84–92, 2013.

[172] Thomas Louail, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4:5276, 2014.

[173] Rémi Louf and Marc Barthelemy. A typology of street patterns. *Journal of The Royal Society Interface*, 11(101):20140924, 2014.

[174] B. Ma, T. Xu, and J. Zhou. Imagerank: A novel sorting algorithm with relevance feedback in application of national costume image retrieval. In *Proceedings of 2nd International Conference on Signal and Image Processing (ICSIP)*, pages 166–171, 08 2017.

[175] Baiyou Ma, Tianwei Xu, and Juxiang Zhou. Imagerank: A novel sorting algorithm with relevance feedback in application of national costume image retrieval. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, pages 166–171. IEEE, 2017.

[176] Matteo Magnani and Luca Rossi. The ml-model for multi-layer social networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 5–12. IEEE, 2011.

[177] Feng Mao, Minhe Ji, and Ting Liu. Mining spatiotemporal patterns of urban dwellers from taxi trajectory data. *Frontiers of Earth Science*, 10(2):205–221, 2016.

[178] David Martin, Christopher Gale, Samantha Cockings, and Andrew Harfoot. Origin-destination geodemographics for analysis of travel to work flows. *Computers, Environment and Urban Systems*, 67:68–79, 2018.

[179] A Paolo Masucci, Joan Serras, Anders Johansson, and Michael Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2):022812, 2013.

[180] H. Migallón, V. Migallón, J.A. Palomino, and J. Penadés. A heuristic relaxed extrapolated algorithm for accelerating pagerank. *Advances in Engineering Software*, 120:88–95, 2018.

[181] Yuriy Mileyko, Herbert Edelsbrunner, Charles A Price, and Joshua S Weitz. Hierarchical ordering of reticular networks. *PLoS One*, 7(6):e36715, 2012.

[182] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[183] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010.

[184] Lewis Mumford. *The city in history: Its origins, its transformations, and its prospects*, volume 67. Houghton Mifflin Harcourt, 1961.

[185] Murray Cox. Inside Airbnb retrieved from http://insideairbnb.com/get-the-data.html . http://insideairbnb.com/get-the-data.html, 2019.

[186] United Nations. World urbanization prospects: The 2005 revision, 2005.

[187] Mark Newman. *Networks*. Oxford university press, 2018.

[188] Mark EJ Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

[189] Peter Newman. *The new Palgrave dictionary of economics and the law*. Springer, 1998.

[190] T.T. Nguyen, H.L. Nguyen, D. Hwang, and J.J. Jung. Pagerank-based approach on ranking social events: A case study with flickr. In *Proceedings of 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, 10 2015.

[191] Tuong Tri Nguyen, Hoang Long Nguyen, Dosam Hwang, and Jason J Jung. Pagerank-based approach on ranking social events: a case study with flickr. In *2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, pages 147–152. IEEE, 2015.

[192] Ke Nie, Zhensheng Wang, Qingyun Du, Fu Ren, and Qin Tian. A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: A case study from wuhan, china. *Sustainability*, 7(3):2662–2677, 2015.

[193] Juhani Nieminen. On the centrality in a graph. *Scandinavian journal of psychology*, 15(1):332–336, 1974.

[194] M. Nykl, K. Jezek, D. Fiala, and M. Dostal. Pagerank variants in the evaluation of citation networks. *Journal of Informetrics*, 8(3):683–692, 2014.

[195] Michal Nykl, Karel Ježek, Dalibor Fiala, and Martin Dostal. Pagerank variants in the evaluation of citation networks. *Journal of Informetrics*, 8(3):683–692, 2014.

[196] Société EIRL Oalley. Oalley how far can i go ?, 2018.

[197] Stan Openshaw. The modifiable areal unit problem (vol. 38). *Norwich: Geo Books*, 1983.

[198] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

[199] Organizers, Wei Kang, Taylor Oshan, Levi J Wolf, Discussants, Geoff Boeing, Vanessa Frias-Martinez, Song Gao, Ate Poorthuis, and Wenfei Xu. A roundtable discussion: Defining urban data science. *Environment and Planning B: Urban Analytics and City Science*, 46(9):1756–1768, 2019.

[200] Arthur O'sullivan. *Urban economics*. McGraw-Hill/Irwin Boston, MA, 2007.

[201] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Technical report 1999-66, Stanford InfoLab*, 66, 1999.

[202] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[203] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[204] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[205] Robert Paluch, Xiaoyan Lu, Krzysztof Suchecki, Bolesław K Szymański, and Janusz A Hołyst. Fast and accurate detection of spread source in large complex networks. *Scientific reports*, 8(1):1–10, 2018.

[206] Wei Pan, Gourab Ghoshal, Coco Krumme, Manuel Cebrian, and Alex Pentland. Urban characteristics attributable to density-driven tie formation. *Nature communications*, 4:1961, 2013.

[207] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6:8166, 2015.

[208] R. Pastor-Satorras, C. Castellano, P.t Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87:925–979, 2015.

[209] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2007.

[210] F Pedroche. Métodos de cálculo del vector pagerank. *Bol. Soc. Esp. Mat. Apl*, 39:7–30, 2007.

[211] F Pedroche. Métodos de cálculo del vector pagerank. *Bol. Soc. Esp. Mat. Apl*, 39:7–30, 2007.

[212] F. Pedroche, M. Romance, and R. Criado. A biplex approach to pagerank centrality: From classic to multiplex networks. *Chaos*, 26(065301), 2016.

[213] Rafael Henrique Moraes Pereira, Vanessa Nadalin, Leonardo Monasterio, and Pedro HM Albuquerque. Urban centrality: a simple index. *Geographical analysis*, 45(1):77–89, 2013.

[214] L.H.M. Piraveenan and M. Prokoprnko. Percolation centrality: Quan- tifying graph-theoretic impact of nodes during percolation in networks. *PLoS ONE*, 8(1):1789–1802, 2013.

[215] M.J. Pocock, D.M. Evans, and J. Memmott. The robustness and restoration of a network of ecological networks. *Science*, 335(6071):973–977, 2014.

[216] Sergio Porta, Paolo Crucitti, and Vito Latora. The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2):853–866, 2006.

[217] Sergio Porta, Paolo Crucitti, and Vito Latora. The network analysis of urban streets: a primal approach. *Environment and Planning B: planning and design*, 33(5):705–725, 2006.

[218] Sergio Porta, Emanuele Strano, Valentino Iacoviello, Roberto Messora, Vito Latora, Alessio Cardillo, Fahui Wang, and Salvatore Scellato. Street centrality and densities of retail and services in bologna, italy. *Environment and Planning B: Planning and design*, 36(3):450–465, 2009.

[219] Denise Pumain, Elfie Swerts, Clémentine Cottineau, Céline Vacchiani-Marcuzzo, Cosmo Antonio Ignazzi, Anne Bretagnolle, François Delisle, Robin Cura, Liliane Lizzi, and Solène Baffi. Multilevel comparison of large urban systems. *Cybergeo: European Journal of Geography*, 2015.

[220] Aithal B.H. Ramachandra, T.V. and D.D. Sanna. Insights to urban dynamics through landscape spatial pattern analysis. *Royal Society open science*, 18:329–343, 2012.

[221] TV Ramachandra, Bharath H Aithal, and Durgappa D Sanna. Insights to urban dynamics through landscape spatial pattern analysis. *International Journal of Applied Earth Observation and Geoinformation*, 18:329–343, 2012.

[222] G. Ranjan and Z.L. Zhang. Geometry of complex networks and topological centrality. *Physica A: Statistical Mechanics and its Applications*, 392(17):3833–3845, 2013.

[223] Sergio J Rey and Richard J Smith. A spatial decomposition of the gini coefficient. *Letters in Spatial and Resource Sciences*, 6(2):55–70, 2013.

[224] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of 2002 eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70, 07 2002.

[225] Matthew Richardson and Pedro Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *Advances in neural information processing systems*, pages 1441–1448, 2002.

[226] Salvatore Rinzivillo, Lorenzo Gabrielli, Mirco Nanni, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. The purpose of motion: Learning activities from individual mobility networks. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pages 312–318. IEEE, 2014.

[227] Salvatore Rinzivillo, Simone Mainardi, Fabio Pezzoni, Michele Coscia, Dino Pedreschi, and Fosca Giannotti. Discovering the geographical borders of human mobility. *KI-Künstliche Intelligenz*, 26(3):253–260, 2012.

[228] Martin Rosvall, Ala Trusina, Petter Minnhagen, and Kim Sneppen. Networks and cities: An information perspective. *Physical Review Letters*, 94(2):028701, 2005.

[229] Meead Saberi, Hani S Mahmassani, Dirk Brockmann, and Amir Hosseini. A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. *Transportation*, 44(6):1383–1402, 2017.

[230] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[231] Mohamed Salheen and Leslie Forsyth. Addressing distance in the space syntax syntactical model. *Urban Design International*, 6(2):93–110, 2001.

[232] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *Trans. Neur. Netw.*, 20(1):61–80, 2009.

[233] Nina Schwarz. Urban form revisited—selecting indicators for characterising european cities. *Landscape and urban planning*, 96(1):29–47, 2010.

[234] C Silverstein. Mathematical analysis of hyperlinks in the world wide web. In *SIAM 50th anniversary and 2002 annual meeting. July, 8*, volume 12, 2002.

[235] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96, 2012.

[236] Alex D Singleton, Seth Spielman, and David Folch. *Urban analytics*. Sage, 2017.

[237] Tom AB Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37, 2011.

[238] Luis Solá, Miguel Romance, Regino Criado, Julio Flores, Alejandro García del Amo, and Stefano Boccaletti. Eigenvector centrality of nodes in multiplex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(3):033131, 2013.

[239] A. Solé-Ribalta, M. De Domenico, S. Gómez, and A. Arenas. Centrality rankings in multiplex networks. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*, pages 149–155, 06 2014.

[240] Albert Solé-Ribalta, Manlio De Domenico, Sergio Gómez, and Alex Arenas. Centrality rankings in multiplex networks. In *Proceedings of the 2014 ACM conference on Web science*, pages 149–155, 2014.

[241] Albert Solé-Ribalta, Manlio De Domenico, Sergio Gómez, and Alex Arenas. Random walk centrality in interconnected multilayer networks. *Physica D: Nonlinear Phenomena*, 323:73–79, 2016.

[242] Albert Solé-Ribalta, Sergio Gómez, and Alex Arenas. Congestion induced by the structure of multiplex networks. *Physical review letters*, 116(10):108701, 2016.

[243] Michael Southworth and Eran Ben-Joseph. *Streets and the Shaping of Towns and Cities*. Island Press, 2013.

[244] Gabriel Spadon, Andre CPLF de Carvalho, Jose F Rodrigues-Jr, and Luiz GA Alves. Reconstructing commuters network using machine learning and urban indicators. *Scientific reports*, 9(1):1–13, 2019.

[245] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[246] G. Stergiopoulos, P. Kotzanikolaou, M. Theocharidou, and D. Gritzalis. Risk mitigation strategies for critical infrastructures based on graph centrality analysis. *International Journal of Critical Infrastructure Protection*, 10:34–44, 2015.

[247] Emanuele Strano, Vincenzo Nicosia, Vito Latora, Sergio Porta, and Marc Barthélemy. Elementary processes governing the evolution of road networks. *Scientific reports*, 2:296, 2012.

[248] Emanuele Strano, Matheus Viana, Luciano da Fontoura Costa, Alessio Cardillo, Sergio Porta, and Vito Latora. Urban street networks, a comparative analysis of ten european cities. *Environment and Planning B: Planning and Design*, 40(6):1071–1086, 2013.

[249] Steven H Strogatz. Exploring complex networks. *nature*, 410(6825):268, 2001.

[250] Li Sun, Ximan Ling, Kun He, and Qian Tan. Community structure in traffic zones based on travel demand. *Physica A: Statistical Mechanics and its Applications*, 457:356–363, 2016.

[251] Lijun Sun and Kay W Axhausen. Understanding urban mobility patterns with a probabilistic tensor factorization framework. *Transportation Research Part B: Methodological*, 91:511–524, 2016.

[252] M. Sydow. Random surfer with back step. In *WWW Alt. '04 Proceedings of the 13th international World Wide Web conference*, pages 352–253, 2004.

[253] Wu T and Chen X. Pagerank-based analysis and visualization of ethnic entrepreneurship and innovation. In *Proceedings of 2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, pages 724–728, 07 2017.

[254] Marius Thériault and François Des Rosiers. *Modeling Urban Dynamics*. Wiley Online Library, 2013.

[255] Isabelle Thomas, Pierre Frankhauser, and Christophe Biernacki. The morphology of built-up landscapes in wallonia (belgium): A classification using fractal indices. *Landscape and urban planning*, 84(2):99–115, 2008.

[256] Neeraj Tiwari, CMS Adhikari, Ajoy Tewari, and Vineeta Kandpal. Investigation of geo-spatial hotspots for the occurrence of tuberculosis in almora district, india, using gis and spatial scan statistic. *International journal of health geographics*, 5(1):33, 2006.

[257] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[258] Florian Toqué, Etienne Côme, Mohamed Khalil El Mahrsi, and Latifa Oukhellou. Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1071–1076. IEEE, 2016.

[259] I. Trajkovski. Pagerank-like algorithm for ranking news stories and news portals. In *Proceedings of Innovations conference on Macedonian Society in Information and Communication Technologies*, pages 87–96, 09 2014.

[260] Transport for London. Open data retrieved from https://tfl.gov.uk/info-for/open-data-users/ . https://tfl.gov.uk/info-for/open-data-users/, 2019.

[261] Yu-Hsin Tsai. Quantifying urban form: compactness versus' sprawl'. *Urban studies*, 42(1):141–161, 2005.

[262] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[263] Eleni I Vlahogianni. Computational intelligence and optimization for transportation big data: challenges and opportunities. In *Engineering and Applied Sciences Optimization*, pages 107–128. Springer, 2015.

[264] Valerio Volpati and Marc Barthelemy. The spatial organization of the population density in cities. *arXiv preprint arXiv:1804.00855*, 2018.

[265] I Vragović, E Louis, and Albert Díaz-Guilera. Efficiency of informational transfer in regular and complex networks. *Physical Review E*, 71(3):036122, 2005.

[266] Neng Wan, Bin Zou, and Troy Sternberg. A three-step floating catchment area method for analyzing spatial access to health services. *International Journal of Geographical Information Science*, 26(6):1073–1089, 2012.

[267] Bao Wang, Xiyang Luo, Fangbo Zhang, Baichuan Yuan, Andrea L Bertozzi, and P Jeffrey Brantingham. Graph-based deep modeling and real time forecasting of sparse spatio-temporal data. *arXiv preprint arXiv:1804.00684*, 2018.

[268] Fahui Wang, Anzhelika Antipova, and Sergio Porta. Street centrality and land use intensity in baton rouge, louisiana. *Journal of Transport Geography*, 19(2):285–293, 2011.

[269] Pengfei Wang, Yanjie Fu, Guannan Liu, Wenqing Hu, and Charu Aggarwal. Human mobility synchronization and trip purpose detection with mixture of hawkes processes. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 495–503, 2017.

[270] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2:1001, 2012.

[271] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, and C. Peng. Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Transactions on Mobile Computing*, 18(9):2190–2202, 2019.

[272] Xu Wang, Zimu Zhou, Fu Xiao, Kai Xing, Zheng Yang, Yunhao Liu, and Chunyi Peng. Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Transactions on Mobile Computing*, 18(9):2190–2202, 2018.

[273] Y. Wang, Y. Tong, and M. Zeng. Ranking scientific articles by exploiting citations, authors, journals, and time information. In *Proceedings of 27th AAAI Conference on Artificial Intelligence*, pages 933–939, 07 2013.

[274] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[275] Tzai-Hung Wen et al. Geographically modified pagerank algorithms: Identifying the spatial concentration of human movement in a geospatial network. *PloS one*, 10(10):e0139509, 2015.

[276] K. Wenfeng, T. Guangming, S. Yifeng, and W. Shuo. Identifying influential nodes in complex network based on weighted semi-local centrality. In *Proceedings of 2016 2nd International Conference on Computer and Communications (ICCC)*, pages 2467–2471, 10 2016.

[277] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

[278] AG Wilson. Complex spatial systems: The modeling foundations of urban and regional systems. *Harlow: Pearson Education*, 2000.

[279] Alan Geoffrey Wilson. A family of spatial interaction models, and associated developments. *Environment and Planning A*, 3(1):1–32, 1971.

[280] Alan Geoffrey Wilson. Urban and regional models in geography and planning. 1974.

[281] Jianjun Wu, Rong Li, Rui Ding, Tongfei Li, and Huijun Sun. City expansion model based on population diffusion and road growth. *Applied Mathematical Modelling*, 43:1–14, 2017.

[282] Tao Wu and Xi Chen. Pagerank-based analysis and visualization of ethnic entrepreneurship and innovation. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, pages 724–728. IEEE, 2017.

[283] Feng Xie and David Levinson. *Evolving transportation networks*. Springer Science & Business Media, 2011.

[284] Peng Xie, Tianrui Li, Jia Liu, Du Shengdong, Yang Xin, and Junbo Zhang. Urban flows prediction from spatial-temporal data using machine learning: A survey, 2019. arXiv preprint arXiv: 1908.10218.

[285] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[286] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5453–5462, 2018.

[287] Ikuho Yamada and Jean-Claude Thill. Comparison of planar and network k-functions in traffic accident analysis. *Journal of Transport Geography*, 12(2):149–158, 2004.

[288] Xiao-Yong Yan, Chen Zhao, Ying Fan, Zengru Di, and Wen-Xu Wang. Universal predictability of mobility patterns in cities. *Journal of The Royal Society Interface*, 11(100):20140834, 2014.

[289] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. Representing urban functions through zone embedding with human mobility patterns. In *IJCAI*, pages 3919–3925, 2018.

[290] Gevorg Yeghikyan, Felix L Opolka, Mirco Nanni, Bruno Lepri, and Pietro Lio. Learning mobility flows from urban features with spatial interaction models and neural networks. *arXiv preprint arXiv:2004.11924*, 2020.

[291] M. Yildirimoglu and J. Kim. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. *Transportation Research Part C: Emerging Technologies*, 89:254–267, 2018.

[292] Mehmet Yildirimoglu and Jiwon Kim. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. *Transportation Research Procedia*, 27:1034–1041, 2017.

[293] Mehmet Yildirimoglu and Jiwon Kim. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. *Transportation Research Part C: Emerging Technologies*, 89:254–267, 2018.

[294] Beibei Yu, Zhonghui Wang, Haowei Mu, Li Sun, and Fengning Hu. Identification of urban functional regions based on floating car track data and poi data. *Sustainability*, 11(23):6541, 2019.

[295] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting.

[296] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):712–725, 2014.

[297] Lianfa Zhang, Jianquan Cheng, and Cheng Jin. Spatial interaction modeling of od flow data: Comparing geographically weighted negative binomial regression (gwnbr) and ols (gwolsr). *ISPRS International Journal of Geo-Information*, 8(5):220, 2019.

[298] Wangsheng Zhang, Shijian Li, and Gang Pan. Mining the semantics of origin-destination flows using taxi traces. In *UbiComp*, pages 943–949, 2012.

[299] Pengxiang Zhao, Kun Qin, Xinyue Ye, Yulong Wang, and Yixiang Chen. A trajectory clustering approach based on decision graph and data field for detecting hotspots. *International Journal of Geographical Information Science*, 31(6):1101–1127, 2017.

[300] X. Zhao, H Zhang, M. Zhang, Cheng L., , and S. Ma. Identifying influential nodes in large-scale software networks. In *Proceedings of 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)*, pages 764–767, 10 2017.

[301] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.

[302] Chen Zhong, Stefan Müller Arisona, Xianfeng Huang, Michael Batty, and Gerhard Schmitt. Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28(11):2178–2199, 2014.

[303] J. Zhou, A. Zeng, Y. Fan, and Z. Di. Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, 106(2):805–816, 2016.

[304] Jianlin Zhou, An Zeng, Ying Fan, and Zengru Di. Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, 106(2):805–816, 2016.

[305] Y.B. Zhou, L. Lu, and M. Li. Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. *New Journal of Physics*, 14(3):033033, 2012.

[306] Yuren Zhou, Billy Pik Lik Lau, Chau Yuen, Bige Tunçer, and Erik Wilhelm. Understanding urban human mobility through crowdsensed data. *IEEE Communications Magazine*, 56(11):52–59, 2018.

[307] Di Zhu and Yu Liu. Modelling spatial patterns using graph convolutional networks. In *10th International Conference on Geographic Information Science (GIScience 2018)*, 2018.

[308] Guohun Zhu, Jonathan Corcoran, Paul Shyy, Salvatore Flavio Pileggi, and Jane Hunter. Analysing journey-to-work data using complex networks. *Journal of Transport Geography*, 66:65–79, 2018.

Appendix A

Data

Table A.1 contains an overall summary of the OD network structures in both London and Rome.

City	Nodes	Node attributes	Edges	Edge attributes
London	6791	36	23062231	13
Rome	5432	35	15012127	13

Table A.1: Summary of the OD network datasets

Table A.2 contains a detailed description of the edge attributes of the OD flow networks built from private car GPS trajectory data.

Table A.3 contains a detailed description of the node attributes of the OD flow networks built from private car GPS trajectory data, augmented with.

Table A.3: Node attributes of the500×500m OD network

Attribute	Description
NodeID	IDs of the London grid cells
airbnb_price	average airbnb price in a cell
universities	total area of universities in a cell
tourism	number of touristic attractions

Continued on next page

Table A.3 – continued from previous page

Attribute	Description
theatres	number of theatres
shops	number of shops
shopping_malls	number of shopping malls
restaurants	number of restaurants
residential	total residential area
pubs_cafes	number of pubs and cafes
post	number of postal offices
parking	total parking area
offices	number of office buildings
museums	number of museums
medical	total area of medical facilities
schools	number of secondary schools
industrial	total industrial area
government	number of governmental institutions
fuels	number of gas stations
fast_foods	number of fast food restaurants
commercial	number of commercial firms
cinemas	number of cinemas
bars_cafes	number of bars
banks	number of banks
atms	number of ATM machines
arts	number of arts venues
airport	binary dummy: 1 if an airport falls into the cell, 0 otherwise
in_total	total inflow to cell
out_total	total outflow from cell
street_density	number of street junctions per cell
gyration_radius	average radius of gyration of cars "housed" in a cell

Continued on next page

Table A.3 – continued from previous page

Attribute	Description
gyration_radius _spatial_lag	average radius of gyration of cars "housed" in the neighbours of the cell
highways	binary dummy: 1 if the cell is located on a highway or street on the edge of a city
metro_flow	number of passengers that have entered subway in the cell
avg_betw_ centrality	average betweenness centrality of the street junctions in the cell
in_total_ spatial_lag	total inflow to the geographic neighbours of a cell
out_total_ spatial_lag	total outflow from the geographic neighbours of a cell
APA_flow_only	Computed APA centrality for flow only
APA_food	Computed APA centrality for food
APA_retail	Computed APA centrality for retail

Attribute	Description
location_1, location_2	IDs of the London grid cells
flows	mobility flow counts between cells location1 and location2
netw_distance	the physical road distance between the centroids of location_1 and location_2 extracted from OpenStreetMap
total_loc_flow	the total in/out-flow associated to location1
route_factor	the ratio between netw_distance and the Euclidean distance between the centroids of location_1 and location_2. It is greater or equal to 1
subway	the number of subway lines between location_1 and location_2
bus	the number of bus lines between location_1 and location_2
airbnb	average of the Airbnb prices of location_1 and location_2
speed	average travel speed between location_1 and location_2
time	average travel time between location_1 and location_2
corr_at_ destinations	correlation between the time series of car arrivals at location_1 and location_2
corr_incidence	correlation between the time series of car incidences at location_1 and location_2
location1_to_ neighbourhood	the aggregate car flow count from location_1 to the immediate geographic neighbours of location_2
neighbourhood_to _location2	the aggregate car flow count from the neighbours of location_1 to location_2

Table A.2: Edge attributes of the 500 × 500 m OD network

Appendix B

gHypE statistical random graphs

B.1 Illustrating gHypE intuition

The intuition behind the gHypE statistical random graphs described in Section 9.3.1 is illustrated in Figures B-1 and B-2.

B.2 gHypE regression model selection

The gHypE network regression parameters, estimated via MLE as described in Section 9.3.3 need to be tested for statistical significance. In particular, we want to obtain the statistical significance of the regression model with all parameters $\left\{\hat{\beta}_l\right\}_{l\in[1,p]}$, and test it against simpler variants of the model with a smaller number of parameters. This is known in statistics as *model selection*, which allows to select the regression model with highest statistical significance, and to drop the layers corresponding to non-significant parameters in the model.

Our aim is to compare two structures of statistical models defined by subsets of layers $\{\mathcal{R}_l\}_{l\in[1,q]}$ and $\{\mathcal{R}_l\}_{l\in[1,q+s]}$ as described in equation 9.6, with q and $q+s$ parameters, respectively, and to see which of the two better describes the observed interaction layer $\mathcal{I}$, i.e. the OD flow network.

We note that one model is a special case of the other, and recall that they are described in equation 9.10 which we present here again:

Figure B-1: The configuration model illustrated (left) as a typical edge rewiring exercise and (right) as analogous to the urn problem. In the first case, in order to obtain a new multi-edge, first an out-stub $(A, \cdot)$ is sampled for rewiring, then an in-stub is sampled uniformly at random from those available. If each possible combination of out- and in-stubs is represented as a ball, we get the urn problem without replacement. In this setting, the probability of observing a multi-edge (A, B) is three times as high as that of observing a multi-edge between (A, D) and 1.5 times as high as that of observing a multi-edge between (A, C) in both the edge rewiring and the urn schemes.

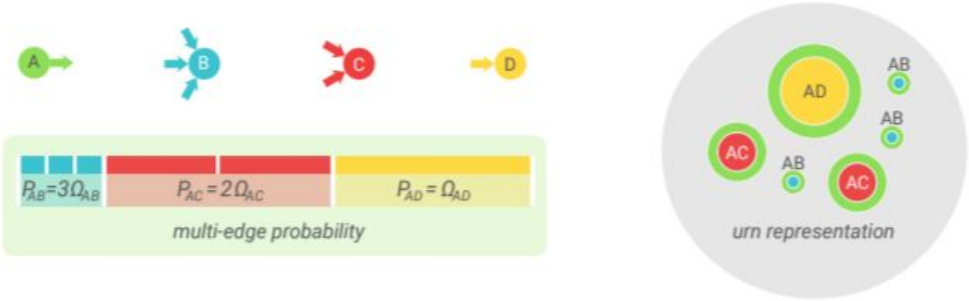


Figure B-2: Edge propensities driving the selection process in the configuration model. As opposed to the conventional configuration model, in this case the stubs are not sampled uniformly at random as in Figure B-1. Once an out-stub has been sampled, each in-stub is then described by a propensity Ω_{ij} of being sampled to form the new multi-edge. This results in the odds of wiring the out-stub $(A, \cdot)$ to the node D being higher than that of B because of a very large edge propensity Ω_{AD},despite node B having three times more in-stubs than node D.

$$L(\beta|\mathcal{I}) = \left[\prod_{i,j}\begin{pmatrix}\Xi_{ij}\\ A_{ij}\end{pmatrix}\right]\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\prod_{l=1}^{p} R_{l,ij}^{\beta_l}}{S_\beta}}\right)^{A_{ij}} dz, \tag{B.1}$$

with $S_\beta = \sum_{i,j}\prod_{l=1}^{r} R_{l,ij}^{\beta_l}\,(\Xi_{ij} - A_{ij})$.

We proceed by conducting model selection via the *likelihood ratio test*, which, as stated by the Neyman-Pearson lemma, is the most powerful statistical test at significance level α. We do this by defining the null hypothesis H_0 by the first model $\{\mathcal{R}_l\}_{l\in[1,q]}$ with q parameters, and the alternative hypothesis H_1 by the second model $\{\mathcal{R}_l\}_{l\in[1,q+s]}$, with s more parameters. We can then use the likelihood ratio test statistic to identify whether the more complex model with $q+s$ parameters has enough explanatory power to justify the addition of complexity by s more parameters.

The likelihood statistic $\Lambda(\mathcal{I})$ is defined as

$$\Lambda(\mathcal{I}) = \frac{L\left(\beta_0|\mathcal{I}\right)}{L\left(\beta_1|\mathcal{I}\right)} = \frac{L\left(\beta_q|\mathcal{I}\right)}{L\left(\beta_{q+s}|\mathcal{I}\right)}. \tag{B.2}$$

Recalling the likelihoods from equation 9.10, the likelihood ratio $\Lambda(\mathcal{I})$ is then given by

$$\Lambda(\mathcal{I}) = \frac{\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\prod_{l=1}^{q} R_{l,ij}^{\beta_l}}{S_{\beta_0}}}\right)^{A_{ij}} dz}{\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\prod_{l=1}^{q+s} R_{l,ij}^{\beta_l}}{S_{\beta_1}}}\right)^{A_{ij}} dz}. \tag{B.3}$$

We can then calculate the p-value corrsponding to $\Lambda(\mathcal{I})$ and select the more complex model only if the null hypothesis can be rejected at a chosen significance level α.

According to Wilks' theorem ([277]), as the number of samples N and the number of dyadic relations $A_{ij} \to \infty$, the likelihood ratio distribution converges to a χ^2 distribution with d degrees of freedom. Moreover, as shown by [63], if the number of non-zero Ξ_{ij} is large enough, Wilks' theorem holds even for a single observed sample ($N = 1$). The number of degrees of freedom in the χ^2 distribution will be $d = (q+s) - q = s$ plus the number of degrees of freedom of the additional

network layers $\{\mathcal{R}_l\}_{l\in[q,q+s]}$. That said, we can do a stepwise selection and find the best model with only the statistically significant network layers included. In every selection step, the results of the likelihood ratio test tell us whether to add or remove a layer from the model. Following the procedure described above, the statistically significant layers were identified and are presented in Table 9.2 in Section 9.4.