Non-binary maximum entropy network ensembles and their application to the study of urban mobility

Oleguer Sagarra Pascual

NON-BINARY MAXIMUM ENTROPY NETWORK ENSEMBLES AND THEIR APPLICATIONS TO THE STUDY OF URBAN MOBILITY

OLEGUER SAGARRA PASCUAL

Supervisor & Tutor: Prof. Dr. Albert Díaz-Guilera
Programa de doctorat de Física
Departament de Física Fonamental
Facultat de Física
Universitat de Barcelona

Abril 2016 – versió 1.0

ο αματης επιεαινι, ο σοπης διστζι ϰαι σιλλοψζεται

— Αριστοτλις

The ignorant person affirms,

the wise man [or woman] hesitates and reflects. — Aristotle

Dedicat a la Berta per ser el meu present i futur, i a la meva família
per ser el meu origen i un constant suport.

FOREWORD

Stepping into the scientific world is a marvellous adventure. Such a world which should not be disconnected from the society we live in, its interests and challenges. Yet, it is a world that demands intellectual rigour, personal honesty (with others and most specially with oneself) and huge personal motivation and effort.

Achieving this balance is a difficult endeavour, and I will not be the one to pretend that science is objective: It shapes our inner self as we do it, and we improve it by mixing its method with our personal choices and experiences.

This thesis is the testimony of the beginning of my journey in this scientific world. A journey full of uncertainties, difficulties, happy moments, constant surprises and opportunities to meet all sorts of interesting people.

It has been long, it has been hard and it has pushed me to overcome my inner fears of failure. All in all, a great and rewarding experience.

Dear reader, I hope you enjoy this text while reading it as much as I have enjoyed while developing what is in it. But beware, be critical as you read, as (positive) critical spirit is what pushes science forward.

ACKNOWLEDGEMENTS

Escriure uns agraïments sempre és una feina difícil. Arribar a la fi de la redacció d'una tesi és un procès llarg, on forçosament hi intervenen moltes persones. El que dec a totes aquestes persones no és senzill de plasmar en un petit text, i sóc conscient que de manera injusta però involuntària, ometré molta gent que mereix ser mencionada aquí. Per si de cas, m'agradaria expressar un reconeixement a tots els i les que d'una manera o d'una altra, heu fet possible aquest document. Moltes gràcies per dedicar-me el vostre temps i esforç.

En primer lloc vull agrair la meva família. Al meu pare Ferran un reconeixement per llegir-se pacientment la tesi, a la meva mare Margot i els meus dos germans Boi i Enric un gran gràcies per ser un suport constant. I òbviament a la Berta, per ser la millor companya en aquest viatge, compartint tants i tants moments de qualitat i evitant que la feina m'empresonés.

Tot seguit, un gran reconeixement a l'Albert, per apostar des del primer dia per mi, per la seva experiència, els seus consells i la seva proximitat i sobretot per un tracte professional, però sobretot personal, extremadament generós. No em voldria deixar el Conrad en aquest apartat, per la seva precisió i finesa en les seves aportacions, ni els col·laboradors amb els qui he treballat, en Mario per tantes hores compartides dins i fora del despatx, així com en Josep i la Isabelle, i la Luce, la meva "mare" científica.

Als companys de despatx i de grup, menció molt especial al Pol per ser el millor company de fatigues que un pot desitjar en una aventura com aquesta, per tants congressos, birres, hores i diversions passades. Però també vull mencionar l'Oriol per l'inestimable i sempre disponible ajuda (especialment en temes informàtics), en Kolja, la Roberta, en Guille i l'Ignacio, així com els companys post-docs, en Nikos, l'Antoine, en Jan, el Ruben, en Tiago, l'Oriol i en Francesco i la resta de PI's del grup, els Marians.

Fer recerca en el nostre àmbit pot arribar a ser molt divertit, i dóna per conèixer molta gent: L'Emanuelle (¡gracias por la hospitalidad!) i la gent de Saragossa, tota la gent de Tarragona i de complexitat.cat (l'Àlex i el Jesús en especial), així com els de les escoles d'estiu i congressos del FISES (menció apart els amics de Benasque-people, en especial al tarat d'en Dani i l'Édgar), els co-organitzadors del Warmup (thanks Gio, Miche) i tota la gent amb qui he coincidit als diversos congressos (especialment als ECCS). En aquest sentit, no podré oblidar mai les estones passades amb tota la gent que ha format part de les JIPI, tants els organitzadors (tot l'equip en general però en especial l'Anna, en Joan, i en Rebled) com els assistents a cadascuna de les quatre jornades en que he participat. Tancant aquest apartat, una gran menció a la gent de D-Recerca, Doctorands Diagonal i Precarios-FJI per tot l'esforç en dignificar la feina dels joves investigadors, i per tot el que he après formant-ne part.

From my stay in Boston, I would like to thank Carlo for giving me the opportunity to be a part of such a big thing as the MIT community and all the people (specially Pierrick, Tony, Luis, Clara and the rest of the gang) at the SENSEable city lab for their warm welcome and nice time spent together. Most special thanks to Michael and Roberta, without whom the stay would have not been possible in the first place. I would not like to forget about all the people at Spain@MIT for the crazy times which we enjoyed together, as well as all the folks at the Minutemen hockey club. Last but not least, a big thanks to the one and only "Papito" Rémi: Merci pour être là, toujours et a n'importe quelle heure!

No sols de recerca es viu, i no voldria oblidar mai les magnífiques estones passades a la Ponderosa, entre festes, sopars o simplement compartint hores amb la Maria, l'Arnau, en Martí, en Peter i la Gisela, així com l'Ausiàs i la Vero. Tampoc puc oblidar la veïna de la cantonada i el tercer component del Filipino's-Team, l'Alba i en Fran, per les llargues estones orbitant al voltant del barri. La vida (ni la meva recerca) no seria igual sense la desconnexió que suposa el meu equip i tot l'univers de la gent de la secció de hockey herba del FCB (i el món de l'stick en general). Tampoc ho serien les estones d'oci sense la gent del lunchtime (Núria, Enric, Uri, Dani) ni el petats (Xavi, Eloi, Padu) ni tota la gent de físics pel món, ni molt menys la penya del Port-Team, que més a prop o més lluny, sempre són allà quan toca.

Finament, m'agradaria acabar aquests agraïments amb una menció especial a tots els contribuents i la gent que, amb la seva defensa dels serveis públics i els seus impostos, fan possible la nostra feina com a investigadors.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND NOTATION

ACRONYMS

Micro Canonical Ensemble MC Ensemble

Canonical Ensemble C Ensemble

Grand Canonical Ensemble GC Ensemble

Multi-Edge case ME case

Weighted case W case

Aggregated Binary case B case

NOTATION

$\vec{T}$	Matrix
$\vec{s}$	Vector
t_{ij}	Element ij of matrix $\vec{T}$
$\Theta(t_{ij}) = \begin{cases} 0 & t_{ij} = 0 \\ 1 & t_{ij} \geqslant 1. \end{cases}$	Binary projection of variable t_{ij}
s_i	Element of vector $\vec{s}$ at position i
t	Unidimensional (possibly random) independent variable
$s(t)$	Unidimensional dependent variable
$\hat{t}$	Fixed numerical value for independent variable t
$\hat{s} \equiv s(\hat{t})$	Fixed numerical value for dependent variable s
$\{\vec{T}\}$	Ordered set of all possible values of $\vec{T}$
$\langle t \rangle \equiv \sum_{\{t\}} p(t)t$	Ensemble average of variable t
$\sigma_t^2 \equiv \langle t^2 \rangle - \langle t \rangle^2$	Ensemble variance of variable t
$\mathcal{N}_i$	Cardinality of the sum over given index
$\bar{t} \equiv \frac{\sum_i t_i}{\mathcal{N}_i}$	Average of magnitude t over a single graph realisation
$\mathrm{std}[t] \equiv \frac{\sum_i (t_i - \bar{t})^2}{\mathcal{N}_i}$	Graph standard deviation of magnitude t
$\varepsilon_t \equiv \frac{\hat{t} - \langle t \rangle}{\langle t \rangle}$	Relative error between variable t and ensemble average
$\hat{\varepsilon}_t \equiv \frac{\langle t \rangle - \hat{t}}{\hat{t}}$	Relative error between ensemble average and variable t

Part I

INTRODUCTION: BIG DATA, COMPLEX NETWORKS, MOBILITY AND URBAN SCIENCE

Doing a Doctoral thesis is a long, laborious and exciting task. Often, it can be tortuous and include many detours in the way. This part tries to introduce the reader to the main motivations behind this work, relevant questions and objectives set at its start. It also serves to put into place the starting ingredients from which this thesis has emerged.

Finally, it helps to understand how this document is structured, and where to find the relevant information for each research question posed.

1

BRAHE, KEPLER AND NETWON OR THE SYMBIOTIC CIRCLE AMONG DATA, PHENOMENOLOGY AND THEORY: CHALLENGES FOR SCIENCE IN THE BIG DATA ERA

But [Computers] are useless. They can only give you answers

— *Attributed to* Pablo Picasso by William Fifield [76]

The world is becoming increasingly connected. Distances shorten at the same pace at which correlations among far events strengthen in a process which has been termed as Globalization. Such a process is happening in societies which are fast abandoning the analogic world to step into a digital one, where traces of most of our activities can be stored, analyzed and exploited. Yet our world has also become noisier: We live surrounded by a constant flow of inputs, and we are told that we live in the information age, however, it is becoming harder and harder to extract knowledge from the processing of such a magma of different (and often contradictory) elements.

Throughout the history of science, data (experiment) and theory have always evolved side by side in a symbiotic circle. Hence, the branch of statistics has emerged as a fundamental tool for experimentalists. As R. Fischer put it [77],

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

We are living a paradigm shift, which is altering many aspects of our life. The way we do science is not an exception, and we are witnessing important changes in the way problems are addressed. Many are claiming that what matters nowadays is only data, and even have predicted the "death of theory" (see the provocative piece by Anderson [19] and subsequent stir).

To respond to these claims, we can take a historical perspective going back in time to another paradigm shift, that of the abandon of geocentric theories for the solar system in favor of the heliocentric view. In an extremely coarse and overly simplified historical view[1]

[1] The tale of Brahe, Kepler and Newton presented here is simplistic and many concerns from a historical perspective can be raised regarding their contributions and those made by others, however, it remains close enough to actual facts as to deliver the intended message home. It is based in the blog entry [104].

we can identify three main characters in this story: Tycho Brahe, its successor Johannes Kepler and the genius of Isaac Newton.

Tycho Brahe was a prominent astronomer of his time, advocate of the geo-heliocentric theory of the solar system. Whilst he did propose theories, he was also a proficient observer, who accumulated a precise, documented and large dataset (for the time) of astronomical observations. In the last stages of his life, he collaborated with Kepler and shared his observations with him. Kepler, upon inspecting Brahe's observations to contrast his own theories on the movement of celestial bodies, noticed that he could reduce most of them to three basic phenomenological laws well reproduced in all observations known as *the laws of planetary motion*. While he did propose theories on the structure of the universe, his main legacy remains the three aforementioned laws. These, in addition to the work by other authors, helped Newton (and others) to establish the well known framework to describe not only celestial mechanics but a unified view on the physical laws of motion. Needless to say, the predictive power of these laws of motion has been validated over and over by experiments ever since.

And what can we learn from this story? That these three legs support the structure where scientific and societal advances are built. Data needs to be carefully and as objectively as possible analyzed in order to extract phenomenological observations, that later need to be explained by a unifying framework with predictive power, which leads again to the data for verification.

Getting back again to the present time, it is obvious that the problems we face today are different to those posed by the movement of the stars. The way to face them, however, should not be changed. The elegance and predictive power of such an approach has always attracted physicists and mathematicians to try and adapt the modeling methodologies developed for physical systems to a wide range of problems, including social and economical aspects of human life. These fields, however, require a multi-angled perspective. They cannot be studied in an isolated manner because human related activities are composed of many elements interacting at different scales (temporal, spatial, virtual or analogical) and from which we can only get partial (albeit rich) information. From a modeling point of view, this means that the focus needs to be put not only on the elements forming the systems under study, but also on the interactions among them.

The study of systems using this perspective has given rise to what is known as Complex Systems Science. For these types of systems, we need to use modeling structures that allow us a simple treatment of interactions among diverse elements, and for this reason, the particular subfield of complex networks has seen a spectacular rise in recent times.

1.1 COMPLEX NETWORKS AND STATISTICAL MECHANICS

A network or graph (as they are usually called in mathematics) is an apparently simple structure composed of nodes and relations among them (edges). These structures encode the interaction (or lack thereof) among elements of the system (nodes). It usually counts a large number of nodes, which can belong to a single or a variety of species (types of nodes).

There is a plethora of literature on the field of complex networks, but I refer the interested reader to the reference book by Newman [128].

The power of this approach resides precisely on its simplicity. Having roughly two types of elements, these structures are easy to visualize, manipulate computationally, use to store and represent data and also they are susceptible to be treated analytically in such a way that non trivial relations among their elements can be explored.

The fact that networks contain a large number of elements and pairwise interactions encoded in their edges, has inspired naturally the interest of theoretical physicists in this field. In particular, clear analogies have been stablished in a statistical treatment of these structures and the (extremely successful and powerful) tools developed by both equilibrium and out-of-equilibrium statistical mechanics branches used in the analysis of natural systems composed by a large number of elements (molecules/atoms) such as gases. The study of conventional dynamics deployed on top of non trivial interaction patterns like the ones represented by complex networks, has also enriched the field of statistical physics itself. Novel types of phase transitions and limitations to the effect of diffusion processes caused solely by the structure of interaction among nodes have been discovered among other interesting phenomena.

For a review of critical phenomena in complex networks see [68].

1.1.1 *Going beyond binary networks*

Basically, the field of complex networks can be subdivided in studies analyzing the structure (topology) of interaction among elements of a variety of systems (be them social, biological or others) and the non-trivial (in the sense of neither being additive nor linear) outcomes of dynamical processes studied on top of these topologies.

Even though these structures appear to be simple, in recent times, the need to attain higher levels of concision has provoked the rise of multiple types of complex networks: Binary, multi-layered, bipartite, temporal, interconnected networks and many more.

For a general view on the different types of structures considered in the field of complex networks, see [103].

However, extending and adding degrees of freedom or detail to these structures comes at a cost: The initial idea of simplicity can be progressively lost, with the obvious risk of losing control over the real influence of topology on the observed phenomena. For this, models with tunable features to be used as benchmark are needed. Yet, generating null models for increasingly complicated structures can be a difficult task [22]. And with a lack of solid theoretical foundations for these new structures, testing hypothesis on them is essentially impossible. It is my objective in the present work to contribute to this task by focusing specifically on the case of networks where connections among elements are numerically quantified.

For reasons provided later, I shall call this structures non-binary networks, but mostly in the literature those are known as *weighted networks*. Weighted networks are a particular type of graphs where the interactions among nodes are not dichotomic but graded in some way. They are useful for providing and additional level of detail beyond the mere *skeleton* of interactions among the elements of a given system. The tradeoff for this increase in information is that by adding new degrees of freedom, their mathematical and computational handling becomes more complex.

1.2 THE DAWN OF COMPUTATIONAL SOCIAL SCIENCE: MOBILITY AND URBAN STUDIES

As data-related policies and debate take a central role in our societies, research fields from the social sciences have turned their attention to it. Being these fields mainly speculative and based on partial observations for the majority of the cases through history, the scalability and availability of new data sources has fostered the appearance of what has been called Computational Social Science (CSS) [55].

In this aspect, studies centered in the structure of cities are no exception. Cities are a prominent example of a complex system: They are alive, interconnected, span multiple scales, contain extremely different actors and are scenario of a huge variety of activities. For this reasons, the need for a new *science of cities* [27] has been called out. Such a science, obviously, must be interdisciplinary by construction, and should incorporate the use of data, phenomenology and theory in a consistent manner.

Among the many aspects that may be studied in cities, an interesting one is that of mobility. Not only for obvious ecological, sociological and economical aspects (by 2050, it is forecasted that 60% of world population will live in urban areas [185]) but also as a testing benchmark of the limits on the capacity of this new big data paradigm to generate knowledge.

The study of mobility is not new, taking special importance in the field of transport theory and planning [60]. Traditionally, the study of transport needs is performed in two stages: Traffic generation forecasting and routing. The first stage consist in predicting the amount of traffic that will be observed among all the studied locations while the second one is focused on routing the predicted traffic along the existing infrastructure. Both stages are clearly related to the network approach, where locations can be considered as nodes and the connections among them, physical for the later stage (infrastructure) and in flow (number of people among locations) in the initial stage, can be associated to edges.

Focusing on the traffic generation aspect, one can easily associate the number of people travelling among nodes to the intensity of their link, hence giving rise to a non-binary, discrete network structure. Furthermore, such intensity is based on the decisions of the travelers, which may be related to a variety of factors. We are interested in developing tools that help in avoiding observation biases, and being able to clearly isolate the different factors that drive mobility. And to do so, once again, models are needed, as data by itself will never be able to explain whether an observed phenomena is caused (not correlated) by a given hypothesized ingredient. Hence, from a theoretical point of view, the connection of this phenomena to weighted complex networks and their theoretical foundations is natural.

But then, if theory is needed, and new statistics are called for, what can we, physicist, do about it, specially concerning the proposed *new science of cities*? I strongly believe that with intellectual humility, we can try and collaborate with the disciplines that have been studying these systems (albeit from radically different perspectives such as urban planning and sociology) for a long time with fruitful results. In particular, I feel that our role must be focused on providing them with the right tools for the analysis of the newly available sources of data, which allow them to separate as much as possible the diverse factors that may influence an observation and hence help in testing stablished theories in the different fields. Needless to say, in the process of adapting old methodologies to new problems, we must not renounce to gain knowledge and enrich in turn the field from which we borrow these tools, reframing questions and revisiting stablished theories, as we will try to do in the present case.

In consonance with the above, this thesis will try to develop tools for the analysis of non-binary network with a general applicability in mind, and exemplify their use in the particular case of urban mobility. It is not my objective to perform a complete dissertation of their social causes and consequences, nor its interaction with the city environment. That is not the job we physicists of complexity ought to do. We can provide interesting tools to extract knowledge from data or more precisely to uncover relations, treat data and answer questions, and we can even contribute in the framing of such questions. However, all this remains orphan if no additional insights from experts regarding phenomena are added to the analysis. Modern problem embrace a huge variety of fields, and thus must be faced accordingly by diverse and interdisciplinary teams.

To conclude, I would like to describe the process followed in the research behind the present work, as this serves to frame it around my personal motivations. It all started as a quest to explore the possibility of using weighted networks to study urban mobility from a mainly empirical point of view, due to the availability of a series of rich datasets. However, after many failed tries, it became obvious that attempting a phenomenological approach to these datasets without the proper tools was an impossible task. There were simply too many different variables involved, and too many metrics one could measure which were related in non obvious ways. Another issue was also made patent. Being urban mobility a highly cross-disciplinary subject, it had been studied by many points of view, yet at the same time it was hard to find appropriate references and relate their findings, which were expressed using different lexicons and approaches. Due to this fact, we resolved to go back to square one, developing appropriate theoretical tools which would allow to revisit our earlier analysis of the data and to guide the modelling approach to it. This in turn, opened many interesting questions related to weighted networks model generation, which has finally been the principal topic of the work. Hence, even if the results of this research are presented in a linear order, i. e., first the theory, then the data analysis, then their applications, that does not reflect its real chronological development. All the issues explained in this introduction are ingredients which have shaped the work presented ahead, where I have tried to convey my particular approach to combine data observations, phenomenology and theory to attack a complex and inter-disciplinary problem.

2 OBJECTIVES AND ORGANIZATION OF THE THESIS

May the force be with you.
— General Dodonna [109]

2.1 ORGANIZATION

The present thesis is structured in four distinct parts: An introduction, its two main parts and a final conclusion.

In the introduction (Part i) I have placed into context the area in which the work of this thesis has been developed. I have introduced some concepts about Complex Systems Science and its relation with the present information age we live in. I have further motivated the use of complex networks as structures that can help us in extracting knowledge from the enormous datasets available to us. While advocating for an interdisciplinary approach to current societal challenges, I have argued what is my view on the role physicist must play and I have tried to justify the approach taken in this respect by the present work.

In Part ii I develop the main mathematical tools that are needed to generate flexible null models for the analysis of non binary networks. Chapter 3 serves as introduction to this part of the document and presents previous work done in null model construction. The problem to be solved is explicitly defined in mathematical terms. Also the main mathematical framework of network ensembles used to solve it is presented. In the subsequent Chapter 4 and Chapter 5 details for the different Grand Canonical (GC) and Micro Canonical (MC) ensemble approaches are developed. In Chapter 6 practical issues on network generation and sampling are discussed and recipes for explicit simulation of network instances are given. The focus is then set on the null model that will mostly used for practical applications, the Multi Edge Configuration Model (MECM). Such a model is used to exemplify how analytical expectations for network observables can be obtained and to numerically test predictions made on earlier chapters concerning ensemble equivalence and other features.

While the Part ii is focused on theoretical aspects of non binary network ensembles, Part iii is devoted to their applications to the analysis of urban mobility data. In Chapter 7 I discuss and introduce some of the current challenges that urban mobility faces with respect to data analysis and modeling. The datasets used in this thesis are presented, their strengths and weaknesses analyzed and we study their main temporal, geographical and general features. In Chapter 8 we show how urban mobility can naturally be described in terms of one of the different types of non binary networks earlier studied. Taking this perspective we analyze their main topological and spatial features. Chapter 9 deals with the issue of modeling trip generation in cities. A review of existing models is performed and metrics to assess model accuracy are introduced and used to determine which of the present frameworks is more convenient for the forecasting of mobility generation. Finally, Chapter 10 presents two explicit applications of the theory developed in Part ii to solve data quality related problems: Under-sampling problems on the one hand and density problems (related with analysis and filtering of relevant data from urban mobility networks) on the other.

The concluding Part iv wraps up the work done by summarizing the contributions that this thesis represents. Last but not least, in this part, I discuss current challenges and thoughts, not strictly scientific, but which are related to the development of this thesis and the general scientific and social framework where it has been developed. Also possible criticism and future research perspectives are briefly hinted.

The appendices provide additional details on subjects mentioned along the main text, as well as information about datasets used.

A final note worth mentioning on language. Research is always performed in group, and the outputs presented in this thesis are not exception. For this reason, during this text, the pronouns "I" and "we" will be used indistinctively in general, except for the cases in which I would like to express an opinion or personal appreciation, where the "I" will be explicitly used.

2.2 WRAPPING UP: OBJECTIVES OF THIS THESIS

The original objectives of this work have obviously suffered many variations and detours along the way, but in a nutshell, can be grouped into mainly two different yet interconnected aspects. One places the focus on theoretical developments while the other is aimed at profiting from the theory to develop real applications for urban mobility problems.

A. **Theoretical aspects:** We aim to enunciate, develop and explore a complete mathematical framework for the study of non-binary complex networks with prescribed properties. Such an approach, rooted in the techniques and concepts developed by the field of classical statistical mechanics and by previous work done on binary networks, should allow for hypothesis testing and isolation of different factors that drive the formation of these structures. We want to uncover what can be achieved analytically and computationally, and to clarify the concepts and connections that can be stablished between the field of classical statistical mechanics and that of networks. Based on the above framework, yet focusing explicitly on urban mobility, we would like to develop tools allowing us to review existing mobility models and assess their limitations and strengths.

B. **Applications:** From the analysis above, the capacity to effectively generate network samples with prescribed constraints for a wide variety of cases should naturally emerge. Such structures, used as generative models, should help in understanding mobility processes encoded in terms of networks. We want to successfully apply network theory to study mobility problem by studying its temporal dimensions, spatial structure and topological features. By doing so, we want to expand also the field of non-binary network data analysis in general. We aim at developing data normalization and knowledge extraction strategies, identifying key indicators and metrics and detecting relevant features in empirical datasets represented in the form of non binary networks.

   Along these lines, all the work done should converge into openly accessible, ready to use, available tools for network practitioners willing to adopt the developed methodologies.

Part II

MAXIMUM ENTROPY ENSEMBLES OF NON-BINARY NETWORKS

Some people may find mathematical derivations overwhelming, even boring. However, these are necessary. Mathematics is the syntax in which we try to understand the world, hence, to correctly analyse what it tells us (the data), we first need to try and predict by writing in its own language what we expect from it.

In this part of the thesis, all the relevant mathematical and numerical aspects concerning null model generation of non binary complex networks are discussed. All later usage of analytical insights for data analysis refer here.

3 ENTROPY, RANDOMNESS AND NETWORK ENSEMBLES: GENERAL ASPECTS

It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so.

— *Adapted from* Josh Billings [36]

Science is interested in determining plausible causality relations between a hypothesis/theory and some observation. It does so by generating predictions (what one would expect given that the starting hypothesis were true) and then comparing the predictions to some observed result of an experiment.

The goal of this part of the thesis is to develop all the methodology needed to build models (*generate predictions*) that allow us to understand the effect that fixing a given structural property has on a network. This will allow us to later on quantify such effects and compare them to real data observations to assess whether any observed feature of a dataset is unexpected or can be explained by some plausible hypothesis. In short, my objective is to generate predictions of networks that have some pre-defined properties, being otherwise *as random as possible*.

Randomness, uncertainty and information are concepts that we use in our daily lives. They are however hard to define since they cannot be directly measured. In physics, entropy is a related notion that is widely used, yet again, one might find trouble upon trying to find a commonly accepted definition. To overcome these problems, I will follow the path set by previous researchers using an analogy to concepts of statistical mechanics (including entropy) to the study of complex networks.

The use of techniques from statistical mechanics to biology, sociology and other areas of human knowledge is not new. Although the field of *Complex networks* (which are called *graphs* in mathematics) has fostered a lot of interest in recent times, many applications related to entropy and statistics are much older. Exponential random graph models (ERG), also called *p-star models* in sociology, are a good example. For so-called *weighted* networks the earlier results using entropy and graphs can be found in transport analysis related areas[1]. Also in economics, trade networks have recently been a focus of attention, and many models using non-binary graph structures have been proposed [31, 158, 71, 59, 175, 80, 122].

See [147] and [13] for a review on ERG models.

[1] See the seminal work by Wilson, well summarized in [191], with abundant references of previous works within.

However, in my modest opinion (and up to my limited level of knowledge of the literature), maximum entropy models for non-binary graphs[2] (usually called *weighted* graphs) are not completely stablished, studied and specified in the literature[3]. The reason for this is that the weighted nature of a graph adds extra degrees of freedom that need to be handled carefully (and that can produce several views of *weighted* networks). For binary networks, in contrast, the presence (or not) of an edge between two nodes is a dichotomic variable and its analysis is considerably simpler.

Before reviewing the main mathematical aspects of null model construction using entropy maximizing techniques, recent works which heavily influenced this thesis deserve a special mention (besides the historical ones already mentioned). Bianconi and colleagues, and, specially Coolen[4] and colleagues devoted substantial work on calculation of entropies for binary graph structures under a wide variety of constraints, and in some cases, also for multiplex structures [33, 124]. This includes network generation [56, 20, 145], information quantification [35] and applications to the study of real-world datasets (specially biological networks [21, 153]).

3.1 PROBLEM DEFINITION

We consider a representation of a network of N nodes, based on an adjacency matrix $\vec{T}$ composed by positive integer valued entries $t_{ij} \in \mathbb{N}$ which we will call occupation numbers. Each of these entries accounts for the intensity of the interaction between any given pair of nodes i and j in the network, measured in terms of discrete events (which may be trips between locations in a mobility network or messages between users in social networks for instance). Throughout this thesis we will consider the case of networks with self-loops[5]. In general, the case of interest will be that of directed networks, the undirected case following in principle from the derivation[6]. Drawing inspiration from classical statistical mechanics [135] let's consider the set of all possible networks that one can build with a given adjacency matrix $\vec{T}$, consid-

Origin-Destination matrices or mobility networks, which will be used extensively in Part iii are in general directed and contain self-loops.

[2] I will use this term instead of the most accepted *Weighted* graph to avoid confusion, since as we will see later in this chapter several conceptions of what a weighted network is can exist.

[3] See for instance [134, 12, 32, 84, 176, 80, 122]. In this chapter I will not refer all the (immense) bibliography on analysis performed on weighted networks, only the cases related to maximum entropy models.

[4] Their papers, although highly technical in mathematical terms, attain the most complete and precise description of binary network sparse ensembles in graphs that I have found in the literature, hence I encourage the interested reader to look into it.

[5] In the cases where distinctive differences appear between the case with self-loops and the case without it, a point will be raised in the text. Otherwise its sufficient to alter any given summation over occupation numbers to exclude any terms with repeated indexes.

[6] This assertion is not general, but for the cases reviewed in this work, unless explicitly stated, the adaptation to undirected is straightforward and will not be carried out.

ering its entries $\{t_{ij}\}$ as random variables. From such a set, which we will call ensemble, we can decide to sample network instances with probabilities $\mathcal{P}(\vec{C}(\vec{T}))$, which in its most general form will depend on some macroscopic observables $\vec{C}(\vec{T})$. An observable is defined as any general property one can measure from a given realization of an adjacency matrix $\vec{T}$ as for example the number of binary connections a node has with others (which is called binary degree, as we will see). The most detailed observable one can measure is obviously $\vec{C} = \vec{T}$ and the minimal observable for an ensemble to make sense is the total number of events $C(\vec{T}) = T \equiv \sum_{ij} t_{ij}$.

Our objective is to obtain analytical expressions for the probabilities $\mathcal{P}(\vec{C}(\vec{T}))$ that would allow us to sample networks in such a way that some observable (or in general a set of Q observables) has some predefined statistical properties, the networks being otherwise as random as possible. We will call such set of of Q observables *constraints* $\vec{C} = \{C_q\}$, each element of which will be a scalar function of $\vec{T}$.

3.1.1 *A maximum entropy principle: Minimum information as a proxy for maximal randomness*

In order to proceed we need a mathematical characterization of what *constrained networks as random as possible* mean. In order to do so, a maximum entropy principle will be used: The maximally random ensemble of networks which are compatible with a set of given constraints will be that with associated sampling probabilities such that they use the minimum number of bits to be encoded. In other words, the ones which provide the minimal amount of information (besides that encoded in the constraints) per observation or conversely, the maximum uncertainty[7]. The measure of the amount of information encoded in each sampling process described by $\mathcal{P}(\vec{C}(\vec{T}))$ will be its associated entropy[8].

Several concepts of entropy for complex networks have been used in the literature [16] (including Von Neumann quantum-like entropies [17, 61]), in the present case, we will restrict ourselves to the most usual entropy definition, that of Shannon (and Gibbs)[9]. The reason for this is that it has a transparent meaning in information theoretic terms and also allows to continue with the analogy with clas-

[7] According to Wilson there are up to three different ways into which one can understand entropy: As a measure of probability (in a Boltzmann sense), as a measure of information (the one being used here) or from a Bayesian statistics point of view. As I will briefly mention during the development of this work, the three visions are closely related and can be unified under the present framework, which follows Jaynes [95]. More details can be found in [191].

[8] Using this definition its easy to see that the less entropic (more informative) network ensemble is one that always gives the same result per observation i. e. displays minimum variability.

[9] For more details on the properties, derivation and usage of entropy to quantify information, I refer the reader to the seminal paper by Shannon [163] or Jaynes [95].

sical statistical mechanics problems. It also extends the theory for tailored random (binary) graph ensembles developed by Coolen and others [34, 21, 153, 146] to some non-binary cases. The general formula for the Shannon-Gibbs entropy functional reads:

$$S^{\Gamma}[\mathcal{P}] = -\sum_{\Gamma} \mathcal{P}^{\Gamma}(\vec{C}(\vec{T})) \ln \mathcal{P}^{\Gamma}(\vec{C}(\vec{T})) = \left\langle \ln \frac{1}{\mathcal{P}^{\Gamma}(\vec{C}(\vec{T}))} \right\rangle, \qquad (3.1)$$

where $\mathcal{P}(\vec{C}(\vec{T}))$ has been earlier defined. The addition of superscript Γ is for the moment a matter of notation, but will be explained afterwards[10]. For the moment, suffice it to say that we take a *subjective view* on entropy, which means that entropy is a quantity related to a sampling process described in terms of a given probability $\mathcal{P}$: Changing the variables, the system under study or the definition of $\mathcal{P}$ will change also its associated entropy (we are closing following Jaynes here).

With the above definition of randomness, we can now specify the mathematical problem to be solved. It will be a maximization problem of finding the set of probabilities $\mathcal{P}(\vec{C}(\vec{T}))$ which maximize the functional form of the Shannon entropy (3.1) and fulfill a number of conditions or constraints. We will consider two types of constraints: *hard* and *soft* constraints. A hard constraint[11] C_H is any condition that any network belonging to the ensemble must obey exactly. In contrast, a soft constraint is a condition C_S which need only be fulfilled once it is *averaged* over the ensemble. Using this classification of constraints, two types of ensembles can be broadly distinguished. The ensembles where all the fixed observables are considered as hard constraints will be called Micro Canonical Ensembles (MC) and as soon as one (or more than one) soft constraint is added, we will talk about Grand Canonical Ensembles (GC)[12].

The problem to be solved when hard constraints are used and their numerical values $\hat{\vec{C}}$ are specified[13] is thus,

$$\max \left\{ S^{\Gamma}[\mathcal{P}] | \vec{C}(\vec{T}) = \hat{\vec{C}} \right\}. \qquad (3.2)$$

This problem leads to a uniform solution of the form,

$$\mathcal{P}_{H}^{\Gamma}(\vec{C}(\vec{T})) = \frac{\delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}}{\mathcal{Z}_{MC}} \qquad \mathcal{Z}_{MC} = \sum_{\Gamma} \delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}. \qquad (3.3)$$

Both ensembles are extensively discussed in chapter Chapter 4 (GC) and Chapter 5 (MC) respectively.

[10] Note that the set Γ over which the sum is performed (which I will call the phase or ensemble space) is for the moment still not completely defined, more ingredients will be needed for that.

[11] The subscripts H and S will be used to refer to hard or soft constraints respectively.

[12] In the literature, usually these ensembles are called Canonical, yet, in the present case, this term will be reserved to specify a specific ensemble within the Grand Canonical formalism that bears a direct analogy the usual Canonical ensemble studied in classical statistical mechanics.

[13] Whenever a random variable x is specified (or measured), its fixed numerical value will be represented as $\hat{x}$.

The summation is performed over all individual (possibly degenerated) realizations or configurations of $\vec{T}$, which we will call phase space Γ[14]. By introduction of (3.3) in (3.1) one obtains[15]

$$S^{\Gamma}[\mathcal{P}_H] = -\frac{1}{\sum_{\Gamma} \delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}} \sum_{\Gamma} \delta_{\vec{C}(\vec{T}),\hat{\vec{C}}} \ln \left(\frac{\delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}}{\sum_{\Gamma} \delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}} \right) = \ln \mathcal{Z}_{MC}. \tag{3.4}$$

The above formula corresponds to the logarithm of the volume or subset of the phase space Γ where the constraints are exactly fulfilled, which in the MC ensemble corresponds to the available phase-space (number of configurations with non-zero sampling probability). We shall call this particular function the Boltzmann (or Γ) entropy of any given ensemble and its interpretation is straightforward in the sense Wilson calls entropy *a measure of probability*: It strictly counts the number of graphs one can generate compatible with the given (hard) constraints.

For soft constraints, we have,

$$\max \left\{ S^{\Gamma}[\mathcal{P}] | \left\langle \vec{C}(\vec{T}) \right\rangle = \hat{\vec{C}} \right\}. \tag{3.5}$$

A discussion on the total number of constraints Q one can consider for the problem is performed in Chapter 5.

In this case, we impose that any of the considered soft constraints are fulfilled on average,

$$\left\langle C_q(\vec{T}) \right\rangle \equiv \sum_{\Gamma} \mathcal{P}^{\Gamma}(\vec{C}(\vec{T})) C_q(\vec{T}) = \hat{C}_q \qquad \forall\, q = 1, Q. \tag{3.6}$$

The solution to this problem gives an exponential family of graphs [13],

$$\mathcal{P}^{\Gamma}_{S}(\vec{C}(\vec{T})) = \frac{e^{\vec{\alpha}\cdot\vec{C}(\vec{T})}}{\mathcal{Z}_{GC}} \qquad \mathcal{Z}_{GC} \equiv \sum_{\Gamma} e^{\vec{\alpha}\cdot\vec{C}(\vec{T})}, \tag{3.7}$$

where I define $\vec{\alpha} \cdot \vec{C}(\vec{T}) \equiv \sum_q^Q \alpha_q C_q(\vec{T})$ for ease in notation (and we assume there are Q fixed constraints). The partition function $\mathcal{Z}_{GC}$ ensures the normalization of the probabilities and the vector of Lagrange multipliers $\vec{\alpha}$ is used to enforce the constraints and must be solved from the constraint equations,

$$\hat{C}_q = \sum_{\Gamma} \frac{e^{\vec{\alpha}\cdot\vec{C}(\vec{T})}}{\mathcal{Z}_{GC}} C_q(\vec{T}) = \frac{d}{d\alpha_q} \ln \mathcal{Z}_{GC}(\vec{\alpha}). \tag{3.8}$$

The GC entropy hence reads,

$$S^{\Gamma}[\mathcal{P}_S] = -\vec{\alpha} \cdot \left\langle \vec{C}(\vec{T}) \right\rangle + \ln \mathcal{Z}_{GC} \tag{3.9}$$

The r.h.s. in equation (3.8) shows the generating function nature of the partition function, for which more details and discussion is provided in Chapter 5.

[14] Equations (3.2) and (3.7) refer to the *probability of obtaining a particular network realization in the Γ space*, not to the *probability of obtaining a network with adjacency matrix* $\vec{T}$. As we shall see, this distinction is very important for non-binary networks.

[15] To simplify notation, along the thesis we will drop the superscript on the probability $\mathcal{P}^{\Gamma}$ when representing entropies $S^{\Gamma}[\mathcal{P}^{\Gamma}]$.

If the ensembles are well defined, one expects that both entropies in (3.4) and (3.9) converge when considering the *high sampling limit* (see below).

3.2 COARSE-GRAINED COUNTING OF CONFIGURATIONS: Γ-SPACE, Ω-SPACE AND DEGENERACY TERMS

The variables used to describe any network are the entries of its adjacency matrix. Yet, any network structure is the (simplified) representation of an underlying system. This relation has been up to now hidden in the general framework presented and will be made apparent through the concept of the Γ and Ω coarse grained ensemble representations. This concept exemplifies the view of Jaynes that entropy is fundamentally a *subjective* quantity: It capitally depends on what the observer chooses (or not) to measure, consider and observe in a system[16].

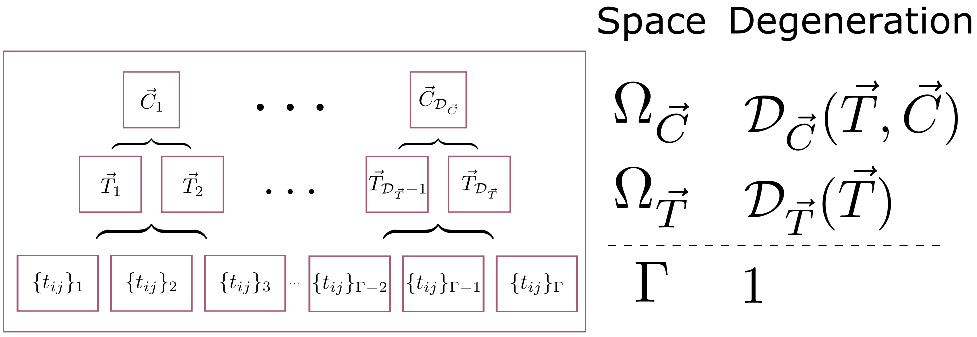

Figure 3.1: **Degeneracy, Γ and Ω representations explained.** Schema showing the different levels of description encoded in the definition of space used. In the bottom level or Γ space the description is made in terms of realizations of the adjacency matrix (which may give rise to the same form of adjacency matrix $\vec{T}$), while in subsequent Ω levels a coarse graining is applied with the corresponding degeneracy terms, which need to be considered when calculating the summation of the corresponding partition functions.

A general schema can be seen in Figure 3.1. On the one hand, the Γ space[17] of an ensemble is its representation in terms of individual realizations of network instances. Several elements of this set can correspond to the same value of $\vec{T}$. On the other, the Ω space is a

[16] To better grasp what this means and its relation to thermodynamic entropies, I encourage the reader to the short text [96].

[17] Throughout this thesis the concepts of set and space will be used as synonyms in an obvious abuse of language. Deviations from the use of the world *space* to refer to other instances such as metric spaces will be made explicit.

degenerate representation of an ensemble in terms of an observable $\vec{O}(\vec{T})$. The relation between both is straightforward,

$$\sum_{\Omega_O} \mathcal{D}_O(\vec{T}) = \sum_{\Gamma} 1 \tag{3.10}$$

where $D_O(\vec{T})$ is a degeneracy term counting how many non-degenerate states of a system give rise to the same observation of $O(\vec{T})$. Being the adjacency matrix the basic representation of a network, the first Ω_O space one can think of is $\Omega_{\vec{T}} \equiv \Omega$. Furthermore, recovering expressions (3.3) and (3.7), the probability to observe a given adjacency matrix $\vec{T}$ is then respectively,

$$\begin{aligned}
\mathcal{P}_H^{\Omega}(\vec{T}) &= \frac{\mathcal{D}(\vec{T})\delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}}{\mathcal{Z}_{MC}} \qquad & \mathcal{Z}_{MC} &\equiv \sum_{\Gamma} \delta_{\vec{C}(\vec{T}),\hat{\vec{C}}} = \sum_{\Omega} \mathcal{D}(\vec{T})\delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}. \\
\mathcal{P}_S^{\Omega}(\vec{T}) &= \frac{\mathcal{D}(\vec{T})e^{\vec{\alpha}\cdot\vec{C}(\vec{T})}}{\mathcal{Z}_{GC}} \qquad & \mathcal{Z}_{GC} &\equiv \sum_{\Gamma} e^{\vec{\alpha}\cdot\vec{C}(\vec{T})} = \sum_{\Omega} \mathcal{D}(\vec{T})e^{-\vec{\alpha}\cdot\vec{C}(\vec{T})}.
\end{aligned} \tag{3.11}$$

Note that we do however *pay a price* in terms of information with the change to a more coarse-grained description: If we consider the entropy of the ensembles (equations (3.4) and (3.9)) we have that,

$$\begin{aligned}
S^{\Omega}[\mathcal{P}_S^{\Omega}] &= S^{\Gamma}[\mathcal{P}_S] - \left\langle \ln \mathcal{D}(\vec{T}) \right\rangle_S \\
&= \ln \mathcal{Z}_{GC} - \vec{\alpha}\cdot\left\langle \vec{C}(\vec{T}) \right\rangle - \frac{1}{\mathcal{Z}_{GC}} \sum_{\Omega} \mathcal{D}(\vec{T})e^{\vec{\alpha}\cdot\vec{C}(\vec{T})} \ln \mathcal{D}(\vec{T}) \\
S^{\Omega}[\mathcal{P}_H^{\Omega}] &= S^{\Gamma}[\mathcal{P}_H] - \left\langle \ln \mathcal{D}(\vec{T}) \right\rangle_H \\
&= \ln \mathcal{Z}_{MC} - \frac{1}{\mathcal{Z}_{MC}} \sum_{\Omega} \mathcal{D}(\vec{T}) \ln \mathcal{D}(\vec{T})\delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}.
\end{aligned} \tag{3.12}$$

So we see an entropy reduction, $S^{\Omega} \leqslant S^{\Gamma}$, because the degrees of freedom are restricted, or conversely we have a lack of resolution. The addition of a degeneracy term $\left\langle \ln \mathcal{D}(\vec{T}) \right\rangle$ breaks the interpretation of entropies for the MC ensemble as a measure of volume of the phase space (the Boltzmann interpretation in physics): Even if all configurations of networks fulfilling the constraints appear with equal probability, not all configurations of $\vec{T}$ have the same statistical weight $\mathcal{D}(\vec{T})/\mathcal{Z}_{MC}$ and hence the entropy of in this space cannot be considered as a strict counting of compatible configurations only.

This poses a problem in terms of relating entropies between spaces: The additional *entropic* term is very difficult (or directly impossible) to measure in general cases for any ensemble (except for few exceptions) and its interpretation is difficult. In our analytical calculations, the quantity of interest will be the leading terms of Boltzmann entropies per event in the infinite sampling limit:

$$\lim_{\hat{T}\to\infty} \frac{S^{\Gamma}[\mathcal{P}_H]}{\hat{T}} \tag{3.13}$$

which is expected to be an specific (*sampling independent*) metric quantifying the amount of "freedom" (configurations meeting the considered constraints) encoded in each ensemble. Sadly, on the numerical side, the only one that can be evaluated directly while sampling networks is the Ω entropy as we will see.

In a nutshell, the difference among both is clear: The entropy in the Γ space is related to *the probability $\mathcal{P}^{\Gamma}$ of obtaining in our sampling process a given, specific, configuration of our system* while the entropy in the Ω space is related to *the probability $\mathcal{P}^{\Omega}$ of obtaining a (possibly degenerate) configuration of our system described uniquely by the variables chosen* (in our case, $\vec{T}$).

3.2.1 *Degeneracy terms for non-binary networks*

As we have seen, the degeneracy term is of capital importance whenever dealing with non-binary networks as it influences vastly the structure of the problems under study. The degeneracy term $\mathcal{D}_T(\vec{T}) \equiv \mathcal{D}(\vec{T})$ is in general subtle to compute and to the best of our knowledge, is seldom considered in the literature. It is entirely determined by the specifics of the system from which the adjacency matrix has been obtained [49].

One can construct adjacency matrices from systems in countless ways and in this work I will consider a specific, albeit quite general form usually found in real data: The aggregation of multiple layers of data [62]. In general, one can consider a multi layered system[18] composed by M levels, each described by an adjacency matrix $\vec{T}^m$, $m = 1, M$. Examples of this procedure range from aggregation of transportation layers [50], networks generated by accumulation of information over a certain time span such as Origin-Destination matrices [4], email communications [91], human contacts [85] or even an aggregation of trading activities in different sectors such as the World Trade Network [175].

Often in real situations, information is available about the aggregation of some of these layers, i. e. the overlay network, whose elements are the result of direct aggregation of all the single-level adjacency matrices,

$$t_{ij} = \sum_{m}^{M} t_{ij}^{m} \qquad \forall\, ij. \tag{3.14}$$

[18] The mono-layered case is recovered setting the number of layers to unity.

Each of these entries correspond to a set of events that quantify the interaction between node i and j. These events can be either distinguishable or indistinguishable, depending on the system under study and thus contribute in a different way to the degeneracy term $D(\vec{T})$.

The systems under consideration are an aggregation of M network layers containing the same type of events: They can be either a group of layers composed by distinguishable (which I will call Multi-Edge networks - ME) or indistinguishable (which will be called Weighted networks - W) events or even an aggregation of Binary (B) networks. Despite the multi-layered structure of the studied systems, we only have *access* to information about their accumulated value through all the layers, i.e. the aggregated occupation numbers $\vec{T} = \{t_{ij}\}$ (equation (3.14)).

The degeneracy term is the product of the multiplicity induced by the nature of the events times the nature of the layers (which in the only real possible scenario are always distinguishable) $D(\vec{T}) = D(\vec{T})_{\text{Events}} \times D(\vec{T})_{\text{Layers}}$. This last term is computed (for each pair of nodes or state ij) by counting the number of different groupings one can construct by splitting $t_{ij} = \sum_m t_{ij}^m$ (distinguishable or indistinguishable) aggregated events into M different layers respecting the occupation limitation of the considered events: Either only one event per layer (Binary network) or an unrestricted number (Weighted and Multi-Edge networks).

NETWORK TYPE	$D(\vec{T})_{\text{Events}}$	$D(\vec{T})_{\text{Layers}}$
Multi-Edge (ME)	$\frac{T!}{\prod_{ij}(\sum_m t_{ij}^m)!}$	$\prod_{ij}\sum_{\{t_{ij}^m\}}\frac{(\sum_m t_{ij}^m)!}{\prod_m t_{ij}^m!} = \prod_{ij} M^{\sum_m t_{ij}^m}$
Weighted (W)	1	$\prod_{ij}\binom{M+\sum_m t_{ij}^m-1}{\sum_m t_{ij}^m}$
Binary Dist. (BD)	$T!$	$\prod_{ij}\binom{M}{\sum_m t_{ij}^m}$
Binary Indist. (BI)	1	$\prod_{ij}\binom{M}{\sum_m t_{ij}^m}$

Table 3.1: **Degeneracy terms' formulas.** Degeneracy terms corresponding to the elements of the system and their layers for each case when considering ensembles with fixed total number of events $\hat{T}$.

The resulting degeneracy terms are shown[19] in Table 3.1, from which one can extract some preliminary conclusions. The degeneracy term on layers factorizes in all cases in ij independent terms while the degeneracy term (only interesting for distinguishable cases) on events factorizes in two parts: One (which itself factorizes) depending on the microstructure of $\vec{T}$, $\prod_{ij}\mathcal{D}_{ij}(t_{ij}) = \prod_{ij}(t_{ij}!)^{-1}$ and another depending on the total number of events $T = \sum_{ij} t_{ij}$, $T!$. For the binary case, both the distinguishable and indistinguishable scenarios will lead to the same statistics, since their degeneracy term on events does not depend on the micro-structure $\{t_{ij}\}$, see (3.11) (hence on the remainder of the thesis we will omit the case BD).

[19] For the layer degeneracy terms, one proceeds as follows: For each state ij out of the possible N^2 node-pairs ($N(N-1)$ if not accepting self-loops) one needs to consider the process of allocating t_{ij} events in M possible distinguishable levels. For the W case this corresponds to the *urn problem* of placing t_{ij} identical balls in M distinguishable urns. For the B case one faces the problem of selecting groups of $t_{ij} \leqslant M$ urns out of a set of M urns and finally for the ME case one must count how to place t_{ij} distinguishable balls in M distinguishable urns. These problems are well known and their solution leads to the second column in Table 3.1, with the product over ij representing the fact that the allocation among the layers for each node-pair is independent. The event related degeneracy terms are only relevant for the distinguishable case and are discussed in details in Section A.1.

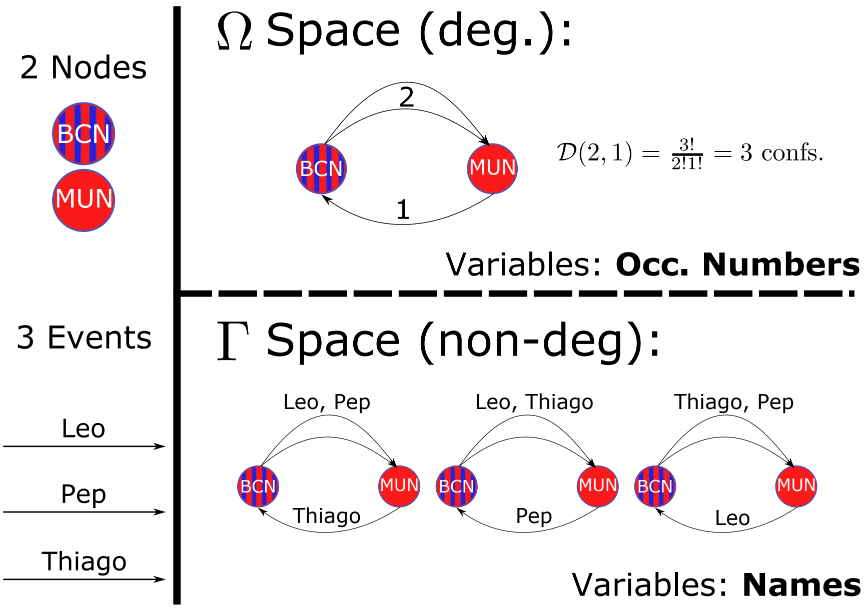

Figure 3.2: **The importance of degeneracy: ME case example.** Using a simple, football inspired, example representing transfer of players and coaches by a network, we show how three very different network configurations in the Γ space give rise to equivalent configurations of an adjacency matrix $\vec{T}$ in the Ω space (with 3 *different* events occupying each one state).

A correct understanding of the degeneracy terms involved in the network representation of each system is thus crucial to attempt any statistical analysis of a network. For the case of non-binary networks this issue is extremely important due to the additional degrees of freedom added to the topology by the integer adjacency matrix as exemplified in Figure 3.2. What representation to use in each case will depend on the specifics of the system under study and in our case, on the application stage I will focus our attention to the ME case. However, the complete framework in its most general form is here derived with the double objective of pinpointing the importance of degeneracy (and differences between obtained results) and completeness.

Modern literature on so-called Complex networks *(starting with [134] and many others such [12, 32, 84, 80]) seems to have overlooked the differences in non-binary network systems, mainly focusing on the single-layered Weighted network case.*

3.3 GRAPHS ARE NOT GASES: SCALING AND THERMODYNAMIC LIMITS

Some final definitions and concepts are needed to perform a comprehensive study of the problem to be solved. Since we have presented two broad types of ensembles from where to sample networks, the MC and GC ensembles, a clarification is in order about the relation between them and the types of study we want to perform on these ensembles.

Exact definition of thermodynamic limits is a wide subject in physics and I will only review minor aspects of it here. I refer the interested reader to [110] and references therein.

Returning once again to the (incomplete) analogy with classical statistical mechanics, we still need a definition of the concept of thermodynamic limit. In physics, the thermodynamic limit is an a useful concept relating to an *infinitely large* system. In this context, *infinitely large* is intended to mean a system where all secondary effects (incomplete sample, boundaries, etc.) related somehow to systematic measuring errors are mitigated and only the main ingredients of a system remain. Generally, this concept is defined as the system in which the specific quantities one can measure become independent of its scaling variable X while its extensive quantities become linear on X. A broad mathematical description would be to consider a system for which all macroscopic observables $O(X)$ that depend on the scaling variable of interest X (normally the number of particles) such as energy, or volume, evolve according to,

$$\lim_{X\to\infty} \frac{O(X)}{X} = \mathrm{Ctnt}. \tag{3.15}$$

As one can readily realize, *graphs are not gases* and hence a definition of a thermodynamic limit is not straightforward (as well as definitions for volume and energy), and using such term can be misleading. For this reason, I will avoid this term and use instead *high sampling limit*. We aim at a limit in the form of (3.15), yet a network has two simple elements to choose from when looking for a candidate for the scaling variable X: The number of nodes N or the number of events T.

3.3.1 *Scaling quantities: Number of events* T *and number of nodes* N

The ensembles of graphs that have been constructed for binary networks in the literature deal with networks that are *sparse* (and cleverly exploit their properties). For these graphs the scaling quantity of interest is the number of nodes N (so $X = N$ in this case). These graphs are normally defined using (3.15) as families of networks for which the density $\rho = T(N)/L(N)$ vanishes as the number of nodes tends to infinity, being $L(N)$ the number of available node-pairs where to allocate edges or binary events (in general[20] $L \sim \mathcal{O}(N^2)$).

[20] For directed graphs we have $L = N^2$ or $L = N(N-1)$ (no self-loops) and for undirected ones one has $L = N(N+1)/2$ and $L = N(N-1)/2$ respectively.

Mathematically, this means that the graph-average degree remains constant in this limit,

$$\lim_{N\to\infty} \bar{k} = \lim_{N\to\infty} \frac{1}{N} \sum_{ij} \Theta(t_{ij}) \equiv \lim_{N\to\infty} \frac{E}{N} = \text{Ctnt.} \tag{3.16}$$

We have used that for binary networks, the state occupation numbers are dichotomic variables, hence $t_{ij} = \Theta(t_{ij}) \in \{0, 1\}$ and we can identify the number of events $T = \sum_{ij} t_{ij}$ with the number of binary connections or edges $E = \sum_{ij} \Theta(t_{ij})$ ($\Theta(x)$ follows the usual definition of the Heaviside Theta function).

Using the above definition, one can work with constraints that are homogeneous functions of degree 1 in N (i. e. *extensive* in N). Two possible examples are the Erdos-Renyi ensemble [69], where the ensemble graph-average degree is fixed $\langle\bar{k}\rangle = E/N$ or the ensembles of graphs with a fixed distribution of degrees $p(\hat{k})$ [34] (so $\langle p(\hat{k})\rangle = \left\langle \sum_i \delta_{k_i,\hat{k}} \right\rangle /N$ is fixed).

If one wants to extend this idea to non-binary graphs, one encounters a problem by the addition of new degrees of freedom in the interaction between nodes. Now three ingredients come into play, the total number of events T, the number of binary events E and the number of nodes N. Additionally, T and E are not independent since by definition $T \geqslant E$. If one tries to use the same definition as in (3.16) for high sampling limit, one gets,

$$\begin{aligned} &\lim_{N\to\infty} \bar{s} = \lim_{N\to\infty} \frac{1}{N} \sum_{ij} t_{ij} = \lim_{N\to\infty} \frac{T}{N} = \text{Ctnt} \\ &\lim_{N\to\infty} \bar{t} = \lim_{N\to\infty} \frac{\sum_{ij} t_{ij}}{L(N)} = \lim_{N\to\infty} \frac{\bar{s}}{N} = 0 \\ &t_{ij} \sim \mathcal{O}(N^{-1}) \implies t_{ij} \ll 1 \, \forall ij \implies t_{ij} \to \Theta(t_{ij}) \implies \bar{s} \to \bar{k}. \end{aligned} \tag{3.17}$$

And one is lead to the binary sparse case again, because the occupation of each link asymptotically vanishes, so there is no difference among binary and non binary sparse graphs. So as we see, defining a large sampling limit is by no means trivial for the non-binary case and will depend again on the problem at hand.

3.3.2 High Sampling *limit for non-binary, non-sparse networks*

In the current case, we are developing our theory to study networks that represent mobility of users between locations, for this reason I shall consider N as a fixed number (the set of locations in a city/country can hardly change once specified by their coordinates). Obviously, the number of recorded trips depends on the observation time τ one spends gathering data (events). Hence, for *infinite time*, $t_{ij} \to \infty \, \forall \, ij$, so, how do we find an specific quantity to determine our high sampling limit? The workaround follows from what we would expect from a non-binary description of a graph: It must add something new, hence it must be sufficiently different from a binary graph. We define the high sampling limit (with fixed number of nodes), following what is proposed in Transport analysis [70]. It is expressed as the limiting case where the number of events T becomes *infinite* (and so do the occupation numbers), while the *fraction* of events allocated to each node-pair is constant. In other words,

$$\lim_{T\to\infty}\left\langle \frac{t_{ij}}{T}\right\rangle \equiv \lim_{T\to\infty}\langle p_{ij}\rangle = p_{ij}^{\infty}\,\forall\, ij \qquad \sum_{ij} p_{ij} = 1. \tag{3.18}$$

The obtained networks will no longer be sparse, since in this limit $\lim_{T\to\infty}\langle\bar{s}\rangle = \infty$. This can be seen as a nuisance, but also helps in clarifying the difference between a node-pair with no-events (for which p_{ij}^{∞} is strictly 0, and an occupied node-pair. It also helps in raising another issue recurrently appearing in non-binary networks and in general not widely discussed. Philosophically, a weighted network should be treated as a fully connected network, where optionally some states can take 0 occupation number values (or weight). The reason for that is what we pinpointed earlier relating to sparse graphs: All sparse graphs are by definition binary, hence all non-binary graphs with a well defined high sampling limit should be non-sparse, and in general display large values of occupation ($\bar{k} \sim \mathcal{O}(N)$). A complementary argument is that usually for non-binary networks, the distribution of existing (non-zero) occupation numbers on states displays a highly skewed nature. In this case, why should we treat differently weights with $t_{ij} = 0$, which are only one unit divergent of existing weights with $t_{ij} = 1$, while treating equally the observed weights (whose maximum value can exceed by far this distance of a unit $t_{ij}^{max} \gg 1$)?.

Under the present framework, a limiting situation we can consider is the *sparse* limit. In such a limit, $T \ll L(N)$ and hence states are weakly populated, $t_{ij} \ll 1$ and we deal with a binary network ($t_{ij} \to \Theta(t_{ij}) \in \{0,1\}$) where any non-binary statistics must converge since no traces of the non-binary nature of the networks can be observed[21].

A final comment must be made about sparse, non-binary, networks: Although this thesis will not deal with them, future work could involve considering separately the binary and non-binary structure of such graphs from an ensemble theoretical point of view taking N as a scaling variable. This would involve defining at least two limits of the form (3.15) to specify the dependence on it of T and E while keeping sparse properties (optionally an additional dependence should

[21] The analogy to the *classical* limit of equilibrium statistical mechanics is clear. What is sometimes called *high temperature* limit is only a particular case of the classical limit, whose strict definition is made in terms of the very sparse occupation of the energy levels of a system.

be added to account for the number of layers M). Doing so, however, requires a careful handling of the concept of sparseness due to the inherent correlations between binary events and events [160].

I discuss some examples of the problematic correlations between binary events and events using the GC ensemble in Chapter 4.

Finally, I want to make a note about the discrete nature of the studies that will be carried out. No continuous weights will be considered, since the extension of this results to the continuous case is by no means obvious. To my knowledge, this problem related to null model construction is rarely mentioned in the literature [49, 186, 22] and is still open today (despite the fact that many weighted networks are used with continuous weights such as correlation values in connectivity brain networks [173]). Usually it is *avoided* by setting all weights discretized according to a minimal unit, yet this minimal unit is somewhat always arbitrary, since obviously there is no analogue to Planck's constant ℏ for graph weights.

3.3.3 *Ensemble equivalence*

And one may wonder, why do we need a high sampling limit? The answer to this is related to network *normalization* on the practical side and to the concept of ensemble equivalence on the theoretical one.

Usually in real applications, several instances of *the same* network are never available to practitioners, however many networks sharing similar *statistical* properties are discovered (despite diverging in number of nodes, edges or events). Studying thus ensembles of networks sharing the same properties can give insights on how to proper rescale the variables of interest and study the significance of the observed common patterns.

In Part iii several examples on the use of analysis over different samples of the same ensemble is shown.

The second, purely theoretical reason is related with another aspect which is usually loosely defined in statistical physics: Ensemble equivalence. In the present case, following [174] we will say that two ensembles describing the same system and same constraints are equivalent if under an appropriate *sampling limit* their relative Shannon entropy or Kullback-Leibler divergence (KL)[22], normalized by the sampling, vanishes.

For extended discussion on ensemble equivalence for physical systems see [183].

$$\begin{cases} \lim_{X\to\infty} X^{-1} \sum_{\Gamma} \mathcal{P}^{\Gamma}_{H} \ln \frac{\mathcal{P}^{\Gamma}_{H}}{\mathcal{P}^{\Gamma}_{S}} = 0 \\ \lim_{X\to\infty} X^{-1} \sum_{\Gamma} \mathcal{P}^{\Gamma}_{S} \ln \frac{\mathcal{P}^{\Gamma}_{S}}{\mathcal{P}^{\Gamma}_{H}} = 0. \end{cases} \tag{3.19}$$

$$\lim_{\hat{T}\to\infty} \frac{1}{T} \sum_{\Gamma} \mathcal{P}^{\Gamma}_{H} \ln \frac{\mathcal{P}^{\Gamma}_{H}}{\mathcal{P}^{\Gamma}_{S}} = \lim_{\hat{T}\to\infty} \frac{1}{T} \ln \frac{\mathcal{P}^{\Gamma}_{H}}{\mathcal{P}^{\Gamma}_{S}(\vec{T}|\vec{C}(\vec{T}) = \hat{\vec{C}})}.$$

[22] KL is not a symmetric measure, but since we are only interested in the absence of divergence, we can use either formula in (3.19).

The prior expression can be worked, since by definition the MC ensemble is a subspace of the GC ensemble (some realizations which fulfill the constraints on average will do so also exactly), hence:

$$\lim_{\hat{T}\to\infty} \frac{1}{\hat{T}} \ln \mathcal{P}^{\Gamma}_{H} = -\lim_{\hat{T}\to\infty} \frac{\mathcal{Z}_{MC}}{\hat{T}} = \lim_{\hat{T}\to\infty} \frac{1}{\hat{T}} \ln \mathcal{P}^{\Gamma}_{S}(\vec{C}(\vec{T}) = \hat{\vec{C}}). \tag{3.20}$$

So it basically means that the logarithm of the probability of observing deviations from the hard-constraints in the GC ensemble scales sub-linearly with the total number of events $\hat{T}$, and thus the leading terms of the Boltzmann entropies are asymptotically equal for both ensembles. An alternative definition, used frequently in statistical physics, states that the two ensembles will be equivalent if relative fluctuations for the constraints defined in each case are equal, and hence vanish because they cannot fluctuate in the MC ensemble,

$$\lim_{\hat{T}\to\infty} \left.\frac{\sigma^2_{C_q}}{\langle C_q \rangle}\right|_{GC} = 0 \qquad \forall\, q = 1, Q. \tag{3.21}$$

In usual physical systems, all ensemble descriptions are equivalent but this is not always the case for networks. Thus studying constraint fluctuations is capital in the GC ensemble, since a statistical -fluctuating- description of an ensemble of networks will not be of much use once applied to real data if the relative fluctuations of its constraints are not zero or bounded, because that would mean that the information carried by the mean value of the constraints averaged over the whole ensemble will not be very informative, and hence the ensemble will not be of much use for null model generation. In Chapter 5 a discussion on the issue of ensemble equivalence for networks of the type discussed in this thesis is performed.

A good example of statistical distributions whose mean carries limited information are scale-free distributions, often found in a variety of social systems. See [129].

3.3.4 *An important note about the analogy to classical statistical mechanics*

A final discussion is important in order to conclude this introductory chapter. By now, it must be obvious to the reader the constant analogies that one can establish between non-binary network ensembles and the classical equilibrium statistical mechanics of physical systems. It is important to make now some points about the limits of this analogy, in order to gain some perspective and in order to be coherent, systematic and frame the conclusions to be learned from our study.

First and foremost of all, in physical systems, the constraints we define in the formulation of a problem are clearly related to the concept of equilibrium. For a GC ensemble description, this means that the Lagrange multipliers we obtain when solving the constraint equations for these problems have a clear physical interpretation (temperature and chemical potential mainly). This connection between statistics and thermodynamics can be used to make predictions on different systems, but this is clearly not the case for networks. Unless we find a way of defining equilibrium states for networks, and evidence that supports this view, the interpretation of given values of Lagrange multipliers $\vec{\alpha}$ for a particular problem (system with set of constraints) will not be possible to extend to other similar problems with different details (same constraints fixed at different numerical values). Hence, extreme care should be taken when trying to do *Network thermodynamics* and extracting general conclusions.

Secondly, in usual physical systems, it is obviously impossible to obtain detailed information about the occupation of each (distinguishable) energy state. Hence, a constraint such as $\vec{T}$ (where the occupation of each state ij is clearly observed) cannot be considered a macroscopic observable. In general, it will be very hard to define either local or state dependent constraints (in our case that would correspond to individual entries t_{ij} and also node related variables) and as a result it should not surprise us that no discussion is performed or even needed on the entropies associated to what we called the Ω space. Such a discussion, however, can be important also in these cases, and as we shall see in Chapter 4 (extended discussion in Section A.1) can lead to interesting and novel results with application also in the field of classical equilibrium systems.

The last obvious divergence between models concerns the total number of available states to be filled L. In our case, L is fixed and so as sampling is increased the occupation of each state increases and diverges in the high sampling limit. In contrast, for physical systems, L is extremely large and not fixed but rather grows for each particle added to the system, hence as N is enlarged to reach the thermodynamic limit, the occupation of the majority of states is very low. This however, does not affect the equivalence of ensembles because the vanishing of relative fluctuations we require is related to constraints, which we have justified that are global and not local, and so are not related to individual occupation of states. To conclude, this means that considering our approach in the high sampling limit to be straightforwardly applicable to a quantum system with discretized states can entail some difficulties, as it would lead to a system with infinite density of particles, which in addition should not interact among them (the majority of cases studied here deal with what one would call non-interacting, state ij separable Hamiltonians).

Despite all of the above, the insights gained from the present study can add value in understanding the structure of statistical mechanics problems in a novel way, as shown by the interpretation some authors have done exploiting the analogies between both cases [174, 74]. For more details, I encourage the interested reader in following the discussion in Chapter 4 (and most specially the extended part in Section A.1).

3.4 WRAPPING UP: INGREDIENTS OF AN ENSEMBLE

In this section, I have tried to review the main concepts that will be needed to address the problem of generating non-binary networks with prescribed constraints. The extended discussion is justified to clearly establish the minimal ingredients that *any* trial to study networks from an ensemble point of view, inspired by statistical mechanics, should cover. They are summarized below in the form of questions any practitioner should try to answer prior to starting such an endeavor:

A. **System details - Degeneracy:** What does the network represent? Are the events in it distinguishable or not? Is it a multilayer system? Do we have access to the whole structure?

B. **Shared properties - Constraints:** What system properties we expect to be invariant/constrained/conserved in each network instance and what not? The effects of fixing which observable do I want to separate from my observations using a null model?

C. **Fluctuations - Ensemble type:** Is the data fluctuating? Do I want to consider hard or soft constraints?

D. **Normalization - Sampling limit:** What will I consider is an appropriate sampling limit to test equivalence between ensembles/fluctuations of constraints? Which relevant scaling variable should I use?

The above questions need to be clearly solved, otherwise, an ensemble description of a system will not be unambiguous. Other relevant and deep mathematical questions, not always easy to solve, such as ergodicity, existence and uniqueness escape the scope of this work, which is intended primarily as a practical tool, however can be important in some cases and should be reviewed if mathematical inconsistencies were to be found.

In the next chapters I develop the theory for GC (Chapter 4) and MC (Chapter 5) ensembles of non-binary, non-sparse, aggregated multiplex networks for both distinguishable and indistinguishable events, using the total number of events T as a scaling variable and setting the number of nodes N and of layers M constant. To conclude this part, I discuss in Chapter 6 also details on how to generate networks in the GC ensemble, as well as a practical application to obtain expected values for network observables (to later compare to real data).

As a final comment, it should be noted that another possible way to obtain randomized versions of networks keeping magnitudes constant is algorithmic randomization (*rewiring*) of graphs[23]. This possibility however, will not be discussed here for several reasons: Firstly,

[23] The alternatives include edge-switching schemas [156] and stub pairing algorithms [30, 42, 193].

no analytical insight can be gained from it, which is useful (as we will see) with regards to understand the structure of the problem at hand and has many applications such as data normalization. Secondly, the properties of such algorithmic randomization are not always clear, and for example an unbiased sampling of graphs can be difficult to obtain [20] as well as be costly in terms of computational time [56]. Finally, our approach is more flexible since it allows to fix a very general type of constraints, which in the case of manual randomization would need to have algorithms designed on a case-by-case basis.

4 GRAND CANONICAL ENSEMBLES

By numberless examples it will evidently appear that human affairs are as subject to change and fluctuation as the waters of the sea agitated by the winds.

— Francesco Guicciardini [90]

In this chapter I will deal with all aspects regarding the Grand Canonical formalism approach to generate networks with prescribed constraints of two broad types: Those that can be written in the form of linear combinations of functions of the occupation numbers t_{ij} and those including also their binary projections $\Theta(t_{ij})$. For all types of constraints, the three cases of Multi-Edge, Weighted and Binary networks will be discussed together with the strengths, insights and limitations of this approach. Also, explicit examples linking to known models and usual situations will be explicitly developed. A very special case where the total number of events can be fixed as a hard constraint for the Multi-Edge scenario will also be discussed (and will be called Canonical Ensemble in analogy to classical Statistical Mechanics).

However, before starting, one needs to carefully review a crucial aspect related with the degeneracy terms mentioned in the earlier chapter, which needs to be taken into account when dealing with ensembles where the number of events is not fixed as a hard constraint and the events are distinguishable. Considering this particular case introduces notable complications because it demands to imagine a reservoir[1] of $F \gg \hat{T}$ (distinguishable) events from where to sample from. However, the distinguishability of all events present in the reservoir leads to a problem: The degeneracy of choosing T events out of a reservoir of F distinguishable particles to populate a given adjacency matrix $\vec{T}$ reads,

$$\mathcal{D}_{\text{Reservoir}} = \binom{F}{T}. \tag{4.1}$$

So the total degeneracy for events reads,

$$\mathcal{D}(F, \vec{T}) = \binom{F}{T} \frac{T!}{\prod_{ij} t_{ij}!} = \frac{1}{\prod_{ij} t_{ij}!} F(F-1)(F-2)...(F-T+1), \tag{4.2}$$

[1] The very same discussion given here can be performed imaging instead of a reservoir a set of infinite copies of the system, see Section A.1 and [3].

which is a quantity that diverges whenever considering the limit $F \gg T$, prior to considering the high sampling limit ($\hat{T} \to \infty$). As it can be seen, distinguishability introduces correlations between all the events present both in the networks and in the imaginary reservoir. However, one needs to take into account that even if the degeneracy term becomes infinite as the reservoir does, the probabilities to sample a given adjacency matrix $\vec{T}$ from the ensemble, given by (3.11), can still be well defined. The reason for this is that if the total number of events is not fixed as a hard constraint, one needs to consider a Lagrange multiplier θ accounting for a closing condition $\sum_{ij} \langle t_{ij} \rangle = \hat{T}$. If one considers the scaling of the Lagrange multiplier θ on F (only possible for certain types of constraints), one can establish an *effective* degeneracy $\tilde{\mathcal{D}} = \prod_{ij} (\hat{T}!)^{1/L}/(t_{ij}!)$ which is factorizable, and hence allows for a statistical description in terms of state ij independent probabilities (a complete justification of the effective degeneracy term and its consequences is given in Section A.1).

Once settled the degeneracy term, we can start by constructing ensembles of networks for all cases (W, ME, B) considering Q constraints of the following form,

As a reminder, W stands for Weighted graph (indistinguishable events), ME for Multi-Edge (distinguishable events) and B for accumulated binary graph (both types of events).

$$C_q = \sum_{ij} f_q^{ij}(t_{ij}) \quad \forall\, q \in Q. \tag{4.3}$$

This form only imposes that the considered constraints can be written as sums of individual functions $f_q(t_{ij})$ of the occupation numbers for each node pair. The minimal constraint which will be used in all cases is that of the total number of events T, which will need to be fixed on average over the ensemble and equal to $\hat{T}$ and will always be explicitly separated from the other Q constraints. The Lagrange multiplier identified with $\hat{T}$ will be identified by θ.

Once the constraints are defined, we can note that the term $e^{\vec{\alpha}\cdot\vec{C}(\vec{T})}$ appearing in expression (3.11) will factorize, i. e. ,

$$e^{\vec{\alpha}\cdot\vec{C}(\vec{T})} = e^{\theta T} \prod_{ij} e^{\sum_q \alpha_q f_q^{ij}(t_{ij})} \equiv z_T^T \prod_{ij} z_{ij}(t_{ij}). \tag{4.4}$$

Where we identify $z_T = e^{\theta}$, $z_{ij} = \prod_q e^{\alpha_q f_q(t_{ij})}$. This circumstance, combined with the factorization of the effective degeneracy terms $\tilde{\mathcal{D}}(\vec{T}) \propto \mathcal{D}_{ij}(t_{ij})$ into state ij independent terms, leads to[2]

$$\mathcal{P}(\vec{T}) = \frac{\prod_{ij} \mathcal{D}_{ij}(t_{ij}) z_T^{t_{ij}} z_{ij}(t_{ij})}{\mathcal{Z}}. \tag{4.5}$$

Thanks to the *soft* constraints, summing such a function becomes simple and can be done individually for each node-pair ij and thus

[2] For ease in notation, through this chapter we shall skip the subindex GC referring to the grand-canonical partition function and the tilde in the effective degeneracy terms.

the partition function factorizes $\mathcal{Z} = \sum_{\Omega} \prod_{ij} \mathcal{D}_{ij}(t_{ij}) z_T^{t_{ij}} z_{ij}(t_{ij}) = \prod_{ij} \mathcal{Z}_{ij}$ and we reach state independent probabilities of occupation $q_{ij}(t_{ij})$ for each node-pair ij,

$$\begin{aligned} \mathcal{P}(\vec{T}) &= \prod_{ij} \frac{\mathcal{D}_{ij}(t_{ij}) z_T^{t_{ij}} z_{ij}(t_{ij})}{\mathcal{Z}_{ij}} \equiv \prod_{ij} q_{ij}(t_{ij}) \\ \mathcal{Z}_{ij} &= \sum_{t_{ij}=0}^{\infty} \mathcal{D}_{ij}(t_{ij}) z_T^{t_{ij}} z_{ij}(t_{ij}). \end{aligned} \tag{4.6}$$

The importance of the degeneracy term becomes now apparent: In combination with the considered constraints, it determines the obtained statistics (and most notably its independency).

Concerning the Shannon entropies in the Ω space, S^{Ω}, and using the formula for the GC ensemble (3.12), we have,

$$\begin{aligned} S^{\Omega}[\mathcal{P}_S^{\Omega}] &= -\sum_{ij} \left(\sum_{q} \alpha_q \left\langle f_q^{ij}(t_{ij}) \right\rangle + \theta \ln t_{ij} + \left\langle \ln \mathcal{D}_{ij}(t_{ij}) \right\rangle - \ln \mathcal{Z}_{ij} \right) \\ &= \sum_{ij} S_{ij}^{\Omega}[q_{ij}] = \sum_{ij} \left(S_{ij}^{\Gamma}[q_{ij}] - \left\langle \ln \mathcal{D}_{ij}(t_{ij}) \right\rangle \right). \end{aligned} \tag{4.7}$$

As expected, being the statistics of the occupation terms independent, one has additivity for the entropy terms. Note that the surprise $-\ln \mathcal{P}_S^{\Omega}(\vec{T})$ is as a function of random variables for each realization of a network belonging to the ensemble. In this GC case, being a sum of independent variables with non-diverging first and second cumulant, its distribution will be gaussian and no outliers are expected, being its average informative about the number of bits contained in the ensemble. This means that when averaging numerically over network realizations, expression (4.7) allows to easily obtain an histogram of $-\ln \mathcal{P}_S^{\Omega}$, whose average value is the GC entropy in the Ω space.

Sadly, the prior expression depends on the Lagrange multipliers $\vec{\alpha}$ through the constraint equations. Once the values of $\vec{\alpha}$ are known, the values for the entropy can be computed, provided one is able to sum the individual partition functions: In some cases they can be approximated by analytical expressions for the large sampling limit or otherwise be computed by numerical simulation. In the next chapter, we will deal with the calculation of entropies for each case in the MC ensemble, which allows for a more transparent treatment. This will also lead us to some considerations about the number of constraints to be considered.

On the following, we specify two generic forms for the functions $f_q^{ij}(t_{ij})$ and perform explicit calculations for each case.

4.1 LINEAR CONSTRAINTS

Let's consider firstly $f_q^{ij}(t_{ij}) = a_q^{ij} t_{ij}$ with $a_q^{ij} \in \mathbb{R} \geqslant 0$ so the constraints take the form,

$$C_q(\vec{T}) = \sum_{ij} a_q^{ij} t_{ij} \quad \forall\, q \in Q. \tag{4.8}$$

There are many examples of models in the literature with these types of constraints, and we will review some examples explicitly. In this section, it will be useful to redefine[3] $z_{ij}(t_{ij}) \equiv z_{ij}^{t_{ij}} = \left(e^{\theta} \prod_q e^{\alpha_q a_q^{ij}}\right)^{t_{ij}}$.

4.1.1 *Summing the partition function*

Summing the partition function for the three cases considered is simple and straightforward. Inserting the values in Table 3.1 (see previous chapter) into (4.9) one can analytically and directly perform the summation,

As a reminder, M *stands for the number of layers from which the occupation numbers are aggregated.*

$$\begin{aligned}\mathcal{Z}_{ij} &= \sum_{t_{ij}} D_{ij}(t_{ij}) z_{ij}^{t_{ij}} \\ &= \begin{cases} \text{ME:} & (\hat{T}!)^{1/L} \sum_{t_{ij}=0}^{\infty} \frac{(Mz_{ij})^{t_{ij}}}{t_{ij}!} = (\hat{T}!)^{1/L} e^{Mz_{ij}} \\ \text{W:} & \sum_{t_{ij}=0}^{\infty} \binom{M+t_{ij}-1}{t_{ij}} z_{ij}^{t_{ij}} = (1-z_{ij})^{-M}; \qquad z_{ij} < 1 \\ \text{B:} & \sum_{t_{ij}=0}^{M} \binom{M}{t_{ij}} z_{ij}^{t_{ij}} = (1+z_{ij})^{M}; \qquad t_{ij} \leqslant M \end{cases}\end{aligned} \tag{4.9}$$

A minor comment is in order here: The sum for the B case cannot be carried up to $t_{ij} \to \infty$, since by construction the occupation of a state is limited up to the number of layers M (it is an aggregation of binary layers).

4.1.1.1 *Canonical subcase for ME networks*

For the specific case of ME networks, one can distinguish a *Canonical* ensemble where both T and N are fixed as hard constraints, and other constraints are introduced as soft. In this case, the partition function to be summed is (note that degeneracy term to be used is no longer *effective*),

The Canonical *name is used in analogy to statistical equilibrium physical systems.*

$$\mathcal{Z}_C^{ME} = \sum_{\{\vec{T}\}} \delta_{\hat{T}, \sum_{ij} t_{ij}} \frac{T!}{\prod_{ij} t_{ij}!} \prod_{ij} (Mz_{ij})^{t_{ij}} = \left(\sum_{ij} Mz_{ij}\right)^{\hat{T}} \tag{4.10}$$

[3] Note that now, as opposed to (4.4), the definition of z_{ij} no longer depends on the occupation numbers t_{ij}.

and hence the explicit probabilities to obtain a given adjacency matrix are (not state independent),

$$\mathcal{P}(\vec{T}) = \frac{\hat{T}!}{\prod_{ij} t_{ij}!} \prod_{ij} \left(\frac{M z_{ij}}{\sum_{ij} M z_{ij}} \right)^{t_{ij}} \equiv \frac{\hat{T}!}{\prod_{ij} t_{ij}!} \prod_{ij} (p_{ij}^{\infty})^{t_{ij}}. \tag{4.11}$$

The obtained statistics correspond to a multinomial distribution with probabilities $\{p_{ij}^{\infty} = z_{ij} / \sum_{ij} z_{ij} = \langle t_{ij} \rangle / \sum_{ij} \langle t_{ij} \rangle\}$ and sampling $\hat{T}$. Note that such magnitudes ought to be specific (sampling independent) in the high sampling limit (see (3.18)). In this chapter we will also consider this case, since it completes the analysis of ME networks, which will be our main interest in Part iii of this work[4].

4.1.2 *Explicit statistics*

We recover well known probability distributions: Poisson distribution for the Multi-Edge case [2] (independent of the number of layers M), Negative Binomial for the Weighted case (being the geometric distribution [176] a special case when $M = 1$) and Binomial distribution for the aggregated Binary case (being the Bernoulli distribution [134] a special case for $M = 1$).

$$\begin{aligned} q_{ij}^{ME}(t_{ij}) &= e^{-M z_{ij}} \frac{(M z_{ij})^{t_{ij}}}{t_{ij}!} \\ q_{ij}^{W}(t_{ij}) &= \binom{M + t_{ij} - 1}{t_{ij}} z_{ij}^{t_{ij}} (1 - z_{ij})^{M} \\ q_{ij}^{B}(t_{ij}) &= \binom{M}{t_{ij}} \left(\frac{z_{ij}}{1 + z_{ij}} \right)^{t_{ij}} (1 + z_{ij})^{-(M - t_{ij})}. \end{aligned} \tag{4.12}$$

The resulting statistics show some important features: On the one hand, one sees that albeit the degeneracy term changes for Multi-Edge networks for either case of a monolayer or a multilayer, the form of the obtained statistics does not. This means that *it is not possible to distinguish a Multi-Edge mono-layered network from an aggregation of multiple Multi-Edge layers belonging to an ensemble with the same constraints*. On the other hand, the situation for the other cases changes: For multiplexes the resulting occupation numbers will have different statistics from the monoplex case. This has the implication than one could in principle *discern* the aggregated nature of a network by inspection of their accumulated edge statistics $\{t_{ij}\}$, provided that one had access to enough realizations of a system and that it belongs to

[4] It can be shown that the result of $L(N)$ independent Poisson process with expected average $\{\langle t_{ij} \rangle\}$ is equivalent to a multinomial process over $L(N)$ sites with probabilities $\{p_{ij}^{\infty} = \langle t_{ij} \rangle / \sum_{ij} \langle t_{ij} \rangle\}$ and total sampling T drawn from a Poisson distribution with average $\langle T \rangle$, which as we will see, are the occupation number statistics for the GC, ME case.

NETWORK TYPE	$\langle t_{ij} \rangle$	$\sigma^2_{t_{ij}}$	$\frac{\sigma^2_{t_{ij}}}{\langle t_i \rangle^2}$	DOMAIN z_{ij}
ME	Mz_{ij}	Mz_{ij}	$(Mz_{ij})^{-1}$	$[0, \infty)$
W	$M\frac{z_{ij}}{1-z_{ij}}$	$M\frac{z_{ij}}{(1-z_{ij})^2}$	$(Mz_{ij})^{-1}$	$[0, 1)$
B	$M\frac{z_{ij}}{1+z_{ij}}$	$M\frac{z_{ij}}{(1+z_{ij})^2}$	$(Mz_{ij})^{-1}$	$[0, \infty)$

Table 4.1: **Relevant moments of occupation number statistics.** First and second moment of the considered distributions, together with the relative fluctuations and domain of distribution parameters.

the same ensemble (i.e. the system evolves according to some given, even if unknown, linear constraints of the form in equation (4.8)).

An example of possible multiplex detection is shown in Section D.5.

Finally, we can see that for a large number of layers, the ensembles converge to the ME, as the degeneracy term on (distinguishable) layers dominates the phase space of the ensembles. In this case, the total number of layers exceeds by far the typical occupation of a state $M \gg \bar{t} = \hat{T}/L(N)$, hence $z_{ij} \sim \langle t_{ij} \rangle /M \ll 1$ (see Table 4.1), and one can confirm the convergence of partition functions to the ME case taking the limit $M \to \infty$ while $z_{ij} \sim \langle t_{ij} \rangle /M \ll 1$ in (4.9).

4.1.3 *Interpretation of Lagrange multipliers*

Another important implication of the obtained statistics is the very different interpretations encoded in the values z_{ij}. This collection of values are related to the constraints originally imposed to the network ensemble through the set of Lagrange multipliers $\{\alpha_q\}$ (equations (3.6) and (4.5)) and can be understood as *a posteriori* measures related to the intensity of each node-pair ij. These measures encode the correlations between nodes imposed by the constrained topology (note that for local constraints only at the level of nodes we obtain a factorization $z_{ij} = Mx_i y_j$). Table 4.1 reports the two first central moments of each distribution. For the Multi-Edge case z_{ij} is both directly mapped to the average occupation of the considered link ij, $\langle t_{ij} \rangle$ and to its (relative) importance in the network (see (4.11)). In all the other cases, however, z_{ij} relates to a probability of a set of events emerging from a given node, to be allocated to its local links. Obviously, as we approach the high sampling limit, z_{ij} grows in all cases, but not in the same linear way (in the W case, for instance, z_{ij} is bounded to a maximum value of 1). This means that while in all cases z_{ij} is related to the importance of a given link with respect to the others, the dependency in all non ME cases is highly non-linear.

A second important feature related to the obtained statistics and the interpretation of z_{ij} concerns the use of *entropic* arguments to justify patterns detected in real networks [97, 140] and the relation of entropy to *hidden variable models* [40]. As one can see, not only the obtained statistics but also the form of its parameters z_{ij} is not *free* and depend on both the degeneracy terms and the considered constraints. If the constraints are non-factorizable for instance, no state independent variables will be reached.

This means that there is only one maximally entropic hidden variable model for each unambiguously defined ensemble, because the constraints and the degeneracy (fixed by the underlying systems one is representing) determine the obtained statistics and their parameters. Obviously any probabilistic model can always be tuned so that the constraint equations match those of real data [15], yet the fact that other observables of these models (beyond the ones already fixed as constraints) reproduce *well* real data cannot be explained invoking a maximum entropy principle.

In other words, one cannot choose a statistic and a predefined form of the parameters z_{ij}, and then maximize its associated entropy by fitting the constraint equations *en passant* [97]. The way to use entropic arguments is first to solve the complete problem of specifying all the constraints and details of the system, obtain predictions (analytically or by simulation) and then check if these predictions match those of real data sufficiently well. When this is the case, on can say that the reason to observe some network property (not fixed in the ensemble considered) can be possibly explained by an entropic origin.

4.1.4 *Constraint equations and fluctuations in high sampling limit*

The main difficulty of the *soft-constrained* maximum entropy framework hereby presented for null model generation is the problem of solving the constraint equations (3.6) associated to each ensemble. Under the constraints in (4.8) they read,

$$\hat{C}_q = \langle C_q \rangle = \sum_{ij} a_q^{ij} \langle t_{ij} \rangle = \begin{cases} \text{ME} & M\sum_{ij} a_q^{ij} z_{ij} \\ \text{W} & M\sum_{ij} a_q^{ij} \frac{z_{ij}}{1-z_{ij}} \\ \text{B} & M\sum_{ij} a_q^{ij} \frac{z_{ij}}{1+z_{ij}} \end{cases} . \tag{4.13}$$

The hindrance of solving the constraint equations is one of the main weaknesses of GC network ensembles, and often is overlooked in the literature.

With the exception of some particular cases, these equations do not have an analytical solution and must be obtained numerically. In this case, the best approach is to maximize the associated log-likelihood of each model to a set of observations (constraints), yet the difficulty of each problem increases with the number of constraints since each fixed magnitude has an associated variable to be solved. Considering the different statistics obtained, the most difficult case by far is the Weighted one (W), since the condition that $0 \leqslant z_{ij} < 1$ imposes a non-convex condition in the domain of the log-likelihood function to maximize, while the others are in general easily solved using iterative balancing algorithms (more will be discussed in Chapter 6).

Concerning the constraint fluctuations, in each case we have,

$$\begin{aligned}
\frac{\sigma^2_{C_q}}{\langle C_q\rangle^2} &= \frac{\sum_{ij}(a_q^{ij})^2\langle t_{ij}\rangle}{\left(\sum_{ij}a_q^{ij}\langle t_{ij}\rangle\right)^2} + \frac{a}{M}\frac{\sum_{ij}(a_q^{ij})^2\langle t_{ij}\rangle^2}{\left(\sum_{ij}a_q^{ij}\langle t_{ij}\rangle\right)^2} \\
&= \frac{\sum_{ij}(a_q^{ij})^2\langle t_{ij}\rangle}{\left(\sum_{ij}a_q^{ij}\langle t_{ij}\rangle\right)^2} + \frac{a}{M}\frac{1}{1+r_q} \qquad (4.14) \\
r_q &\equiv \frac{\sum_{ij,kl|k\neq i,l\neq j} a_q^{ij}a_q^{kl}\langle t_{ij}\rangle\langle t_{kl}\rangle}{\sum_{ij}(a_q^{ij})^2\langle t_{ij}\rangle^2}.
\end{aligned}$$

where $a = 0$ for ME case, $a = 1$ for W case and $a = -1$ for B case. We thus see that the fluctuations only disappear for large sampling for the ME description (by construction, the constraints are extensive in the occupation numbers t_{ij} and $\langle t_{ij}\rangle \propto Tp^{\infty}_{ij}$, p^{∞}_{ij} being an specific quantity). The maximally random allocation of events will be made as homogeneous as possible among the states while preserving the constraints, hence $\{r_q\}$ will in general be large numbers (the denominator in the sums has L terms while the numerator has $L(L-1)$, being L the number of available node pairs for the allocation) and relative fluctuations will be bounded and $\mathcal{O}(M^{-1})$. For very large number of layers, then the ensembles become equivalent to the ME case, and fluctuations vanish in the large sampling limit.

The non-vanishing nature of constraint fluctuations in the B and W cases indicates problems with this limiting behavior which will be addressed in depth in the next chapter. For the time being, note that this limit becomes ill defined for the B case, since the total sampling is limited to $\max(T) = ML(N)$, where we obtain a trivial fully connected network with all node pairs carrying M events. However in such a case, the considered constraints will be in general violated and the equalities $\{\hat{C}_q = \langle C_q\rangle, q = 1, Q\}$ will no longer hold.

We see that even if the GC ensemble will be a useful way of generating models mimicking properties of real networks, from an ensemble point of view with fixed number of nodes (and thus sites) $L(N)$ where to allocate events, it will not be completely well defined for arbitrary sampling in all cases.

4.1.5 *Explicit examples*

In the following we provide several examples to illustrate the possible use of GC ensembles with linear constraints. The schemas to solve the associated constraint equations are provided in Section 6.1. We consider here three types of (possibly overlapping) constraints, those depending on global observables of the networks (total number of events $\hat{T}$ and total cost $\hat{C}$), those depending on its modular mesoscopic structure (partition of the nodes of the network into tightly

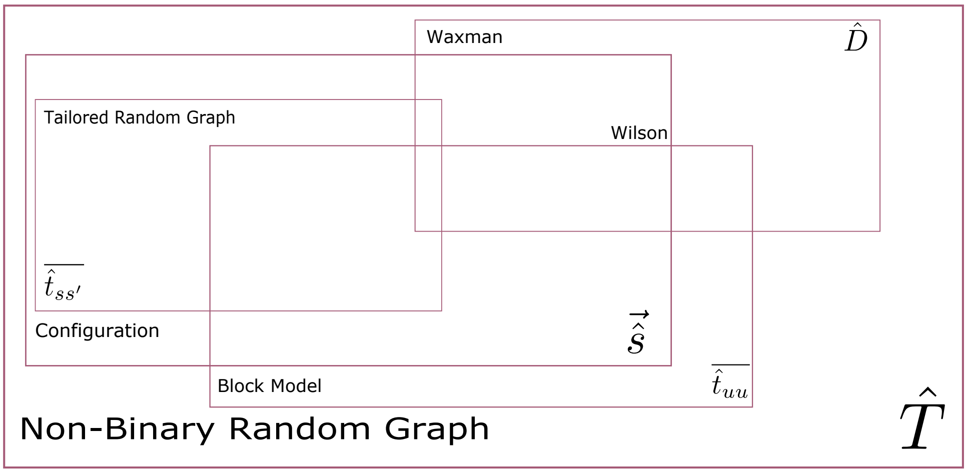

Figure 4.1: **Topology of the phase space of non-binary ensembles with linear constraints represented.** Schematic example of the volume of phase space for different overlapping examples of ensembles with linear constraints. The names represent known models in the literature and the formulas explicit the constraint being enforced, which in the case of overlapping squares represent the combination of different constraints.

connected communities) and those depending on its local node structure (node strength and strength correlations).

For each example, we present the constraint equations in terms of $\langle t_{ij} \rangle$, which are related to the coefficients z_{ij} via Table 4.1 according to each case ME, B and W. Note that any of the considered constraints can be combined to obtain partially overlapping ensembles, whose degrees of freedom are more and more constrained as the number of prefixed properties are increased, as schematically presented in Figure 4.1. In here, we only present a few examples which are related to known models in the literature, but others could be developed depending on the needs for null model construction.

4.1.5.1 *Global constraints*

One can study the effect that fixing global properties of the network can have on its observables, which is the simplest types of constraints one can imagine.

NON-BINARY RANDOM GRAPH: FIXED $\hat{T}$ This would be the direct analogy to the Erdos-Renyi case [69] for binary graphs. In this case all nodes and node-pairs are statistically equivalent and we only have one constraint equation (4.13),

$$\hat{T} = \sum_{ij} \langle t_{ij} \rangle = L \langle t \rangle \implies \hat{t} \equiv \frac{\hat{T}}{L} = \langle t \rangle . \tag{4.15}$$

The prior equation yields different values for z applying Table 4.1 in each case. All the node and graph statistics are known since the sum of independent Poisson distributed variables is a Poisson, the sum of Binomials is a Binomial and the sum of Negative Binomials with identical parameters is also a Negative Binomial. For the W case, one recovers the Weighted random graph model [82]. Once again, it must be noted that some problems arise in this case when considering the high sampling limit. We have that for all the edges $\lim_{\hat{T}\to\infty} z = 1$ and hence a condensation with all events being allocated to a single, yet different for every realization, node-pair, is reached.

In this case we can calculate the relative fluctuation of the constraint exactly,

$$\frac{\sigma_T^2}{\langle T\rangle^2} = \hat{T}^{-1} + \frac{a}{LM}\begin{cases} \mathrm{ME} & a = 0 \\ W & a = +1 \cdot \\ \mathrm{B} & a = -1 \end{cases} \tag{4.16}$$

As expected, we observe how the ME and B cases lead to vanishing relative fluctuations for large sampling (the B case is a fully connected graph with every node pair carrying M events) while this is not the case for the W case.

NON-BINARY WAXMAN GRAPH: FIXED $\hat{T}$, $\hat{C}$ If one considers networks where each node pair has an associated cost d_{ij} per allocated event, then an additional global cost constraint can be added. An example are graphs existing on top of metric spaces defined by an inter-nodal cost matrix $\vec{D} = \{d_{ij}\}$ such as those representing mobility between locations. We can require that the total cost is fixed, hence, $a_d^{ij} = d_{ij}$ and the constraint equations look (4.13),

$$z_{ij} = e^{\theta} e^{-\gamma d_{ij}} \qquad \hat{T} = \sum_{ij} \langle t_{ij} \rangle \qquad \hat{D} = \sum_{ij} d_{ij} \langle t_{ij} \rangle . \tag{4.17}$$

The model cannot be solved analytically, yet some insights can be gained. By graphical arguments we can identify $z_{ij} = \mu e^{-\gamma d_{ij}}$ and we are then lead to the non-binary maximum entropy versions of the Waxman graph [188][5].

For OD matrices $d_{ij} = f(r_{ij})$ where $f(r_{ij})$ is a function representing the cost perception of the users with distance r_{ij} which is usually set as $f(r) = \ln r$ or $f(r) \propto r$. It is called deterrence function *[70] and will be used in Part iii.*

This reasoning can be extended to study the interesting case where the distribution of costs is also fixed[6], which is of particular interest in the field of OD matrices used to analyze mobility, where the mobility of users using certain types of transports is assumed to follow particular statistical forms [28, 116]. One can add recursively more

[5] Note that for single layer binary graphs, the Waxman graph (a graph whose binary connections are stablished according to $\langle \Theta(t_{ij}) \rangle \propto e^{-\beta d_{ij}}$) does not have a strictly maximally entropically form unless we consider it as sparse $\mu e^{-\gamma d_{ij}} \ll 1$ so $\langle t_{ij} \rangle = \langle \Theta(t_{ij}) \rangle = \mu e^{-\beta d_{ij}} / (1 + \mu e^{-\beta d_{ij}}) \simeq \mu e^{-\beta d_{ij}}$.

[6] As done in [35] for the binary case.

and more constraints concerning the moments of the cost distribution $\hat{D^n} \equiv \sum_{ij} \hat{t}_{ij} d_{ij}^n$ (or the elements of a binned histogram), which in all cases will have the linear form (4.8). The solution of the $Q + 1$ constraint equations becomes, however, increasingly complicated.

4.1.5.2 *Mesoscopic constraints: Non-binary Block-model graph. Fixed $\hat{T}, \overline{\hat{t}_{uu}}$*

The next level of detail one can obtain is to introduce some heterogeneity in the mesoscopic structure of the network. The considered linear constraints also include this possibility by predefining a certain community structure into which network nodes can be classified. One can define a set of $\mathcal{N}_u$ communities to which the nodes can belong (specified with a set of labels $\{u_i\}$) and fix the average number of connections within nodes of the same community, $\overline{t_{uu}} = \sum_{ij} t_{ij} \delta_{uu_i} \delta_{uu_j} / \sum_{ij} \delta_{uu_i} \delta_{uu_j}$. In this fashion, we obtain the non-binary counterpart to a model enforcing a block structure [93]. In this case, we have $z_{ij} = e^{\theta} e^{\alpha_u \delta_{u_i u} \delta_{u_j u}}$ and all nodes belonging to the same community are statistically equivalent. The solution is analytical and yields,

See the work (and references therein) by Peixoto et altr. [136, 137] for extended analysis for the binary version of stochastic block models.

$$\langle t_{uu'} \rangle = \overline{t_{uu}} \delta_{uu} + \frac{\hat{T} - \sum_u \overline{t_{uu}} N_u^2}{L - \sum_u N_u^2} (1 - \delta_{uu}). \tag{4.18}$$

A more complicated variant of the Block-model where the strengths of the nodes are also fixed (see below) is discussed in Section B.1.

4.1.5.3 *Local constraints: Non-binary tailored random graphs*

The final level of detail, prior to *freezing* entirely the adjacency matrix, consists in determining node related quantities. In this case, one can mirror the theory for Tailored Random Network Ensembles proposed by Coolen et altr. [21, 153, 146] adapted to the present case.

NON-BINARY CONFIGURATION MODEL GRAPH: FIXED $\hat{\vec{s}}$ We can consider fixing also the strength sequence[7] $\hat{\vec{s}}$ to a predefined value. For each node, we have a pair value of incoming and outgoing strength counting the number of incoming and outgoing events respectively allocated to it, $s^{out} = \sum_j t_{ij}$ and $s^{in} = \sum_i t_{ij}$, and we have $Q = 2N$ constraints. The resulting equations are,

$$\begin{aligned} z_{ij} &= e^{\theta} e^{\alpha_{q_i^{out}}} e^{\alpha_{q_j^{in}}} \equiv x_i y_j \\ \hat{s}_i^{out} &= \sum_j \langle t_{ij} \rangle \qquad \hat{s}_j^{in} = \sum_i \langle t_{ij} \rangle \qquad i = 1, N \quad j = 1, N. \end{aligned} \tag{4.19}$$

[7] Naming this graph as configuration model is an abuse of language inspired by the famous stub-matching algorithm [30, 42, 193]. Throughout this thesis we shall use it to refer to ensembles with a fixed strength sequence.

And thus we have an uncorrelated structure for z_{ij}, where the N pairs $\{x_i, y_i\}$ are solved from (4.13). In this case, its easy to see that for the ME scenario with self-loops an analytical solution can be obtained [159] $\langle t_{ij} \rangle = \hat{s}_i^{out} \hat{s}_j^{in} / \hat{T}$. Since usually in real non-binary networks the distribution of strengths is highly heterogeneous, this can be considered as a fundamental null model where to test and check network metrics and it is fully reviewed in Section 6.4.

Specifically for the ME case, an example of these networks are the newly proposed activity driven networks aggregated over time [138], which are composed by a uniform incoming strength distribution and a heterogeneous outgoing distribution. A complete demonstration of the equivalence between the models is provided in Section A.2.

FIXED $\hat{\vec{s}}, \overline{\hat{t}_{ss'}}$ An extension of the above model can be considered by not only fixing the basic topological structure imposed by the nodes through their strength values but also fixing second order properties. In particular, we can fix the graph-average number of events $\overline{t_{ss'}}$ joining nodes with given strength pairs $\vec{s}$ and $\vec{s}'$, retaining the statistical equivalence of nodes with equal strength pairs. We then have $2N$ constraints imposed by the strength sequence plus $\mathcal{N}_{s^{out}} \times \mathcal{N}_{s^{in}}$ imposed by the correlation structure (where $\mathcal{N}_s$ is the cardinality of each strength sequence). The constraint equations thus look (4.13),

$$\begin{aligned}
z_{ij} &= x_i y_j e^{\alpha_{ss'}} \\
\hat{s}_i^{out} &= \sum_j \langle t_{ij} \rangle \qquad \hat{s}_j^{in} = \sum_i \langle t_{ij} \rangle \qquad i = 1, N \quad j = 1, N \\
\overline{\hat{t}_{\hat{s}\hat{s}'}} &= \frac{\sum_{ij} \langle t_{ij} \rangle \, \delta_{\hat{s}_i^{out} s^{out}} \delta_{\hat{s}_j^{in} s^{in'}}}{\mathcal{N}_{\hat{s}} \mathcal{N}_{\hat{s'}}} \qquad s = 1, \mathcal{N}_s \quad s' = 1, \mathcal{N}_s .
\end{aligned} \tag{4.20}$$

In this special case the constraint equations can be solved. Being our main variables the occupation numbers and the sole constraints related to their strengths, we have that each node-pair joining nodes with the same pair of $\hat{s}_i^{out}, \hat{s}_j^{in}$ will be linked by the same average value $\langle t_{ij} \rangle (z_{ij}(s_i^{out}, s_j^{in'}))$. So the equations are automatically satisfied when $\overline{\hat{t}_{\hat{s}\hat{s}'}} = \langle t_{ij} \rangle (z_{ij}(s, s'))$ because the enumeration imposed by the constraint equations is complete and all z_{ij} share the same form.

4.1.5.4 *Mixed constraints: Wilson gravity model graph. Fixed $\hat{C}, \hat{\vec{s}}$*

A final model to be considered is the one originally proposed by Wilson [190] to unify gravity-like mobility models [70] under a maximum entropy framework. It merges the cost-constrained model with the configuration model and is very relevant for the study of urban mobility, as we shall see in Part iii. We have $Q = 2N + 1$ constraints corresponding to the ones imposed by the node-strength sequence and a total average cost constraints. The constraint equation thus read,

$$\begin{aligned} z_{ij} &= x_i y_j e^{-\gamma d_{ij}} \\ \hat{s}_i^{out} &= \sum_j \left\langle t_{ij} \right\rangle \qquad \hat{s}_j^{in} = \sum_i \left\langle t_{ij} \right\rangle \qquad i = 1, N \quad j = 1, N \\ \hat{D} &= \sum_{ij} d_{ij} \left\langle t_{ij} \right\rangle . \end{aligned} \tag{4.21}$$

It is important to note, however, that in this case the statistically equivalence of nodes is broken, since for two nodes to be equal we require that they share the exact same distances to other node classes and same strength value pairs. This has important implications with regards to the interpretation of the obtained collection of Lagrange parameters $\vec{x}, \vec{y}, \gamma$ and will be discussed in Section 9.2.2.1.

4.2 LINEAR AND BINARY CONSTRAINTS

A second interesting generic form for the individual functions of occupation numbers we can consider is $f_q^{ij}(t_i) = a_q^{ij} t_{ij} + \tilde{a}_q^{ij} \Theta(t_{ij})$. This form allows to control both the non-binary topological structure and its binary projection. The constraints then look,

$$C_q = \sum_{ij} (a_q^{ij} t_{ij} + \tilde{a}_q^{ij} \Theta(t_{ij})) \quad \forall q \in Q. \tag{4.22}$$

In this section it will be useful to redefine the quantities $z_{ij} = e^{\theta} \prod_q e^{\alpha_q a_q^{ij}}$ and $\tilde{z}_{ij} = \prod_q e^{\alpha_q \tilde{a}_q^{ij}}$ which allow to control separately the binary and non-binary structure.

4.2.1 *Summing the partition function*

In this case, the partition function can again be computed directly,

$$\begin{aligned} \mathcal{Z}_{ij} &= \sum_{t_{ij}} D_{ij}(t_{ij}) z_{ij}^{t_{ij}} \tilde{z}_{ij}^{\Theta(t_{ij})} \\ &= \begin{cases} \text{ME:} & \sum_{t_{ij}=0}^{\infty} \frac{M^{t_{ij}} z_{ij}^{t_{ij}} \tilde{z}_{ij}^{\Theta(t_{ij})}}{t_{ij}!} = \left(\tilde{z}_{ij} \left(e^{M z_{ij}} - 1\right) + 1\right) (\hat{T}!)^{1/L} \\ \text{W:} & \sum_{t_{ij}=0}^{\infty} \binom{M+t_{ij}-1}{t_{ij}} z_{ij}^{t_{ij}} \tilde{z}_{ij}^{\Theta(t_{ij})} = \tilde{z}_{ij} \left((1 - z_{ij})^{-M} - 1\right) + 1; \quad z_{ij} < 1 \\ \text{B:} & \sum_{t_{ij}=0}^{M} \binom{M}{t_{ij}} z_{ij}^{t_{ij}} \tilde{z}_{ij}^{\Theta(t_{ij})} = \tilde{z}_{ij} \left((1 + z_{ij})^{M} - 1\right) + 1; \quad t_{ij} \leqslant M \end{cases} \end{aligned} \tag{4.23}$$

4.2.2 *Explicit statistics*

The obtained statistics are simply Zero Inflated [111] versions (ZI) of the previous statistics encountered. A zero inflated random variable is generated by adding an additional parameter that controls separately the binary probability of occurrence of the random variable (which in our case will be related to the collection $\{\tilde{z}_{ij}\}$).

Elsewhere, ZI statistics have been called generalized Fermi-Bose statistics *[84] for networks yet their use is widespread in the field of statistics and modeling.*

Our coarse-grained description in terms of independent occupation numbers implies that for each state, two outcomes can be considered: Either the edge does not exist (obviously with 0 occupation) or it does exist, in which case the resulting (conditioned) statistics will have mean value $\langle t_{ij} | t_{ij} \geqslant 1 \rangle$. We thus obtain Zero-Inflated Poisson for ME case, Zero-Inflated Negative Binomial for W case and Zero-Inflated Binomial for the B case.

$$\begin{aligned}
q_{ij}^{ME}(t_{ij}) &= \frac{(Mz_{ij})^{t_{ij}}}{t_{ij}!} \frac{\tilde{z}_{ij}^{\Theta(t_{ij})}}{\tilde{z}_{ij}\left(e^{Mz_{ij}}-1\right)+1} \\
q_{ij}^{W}(t_{ij}) &= \binom{M+t_{ij}-1}{t_{ij}} z_{ij}^{t_{ij}} \frac{\tilde{z}_{ij}^{\Theta(t_{ij})}}{\tilde{z}_{ij}\left((1-z_{ij})^{-M}-1\right)+1} \\
q_{ij}^{B}(t_{ij}) &= \binom{M}{t_{ij}} z_{ij}^{t_{ij}} \frac{\tilde{z}_{ij}^{\Theta(t_{ij})}}{\tilde{z}_{ij}\left((1+z_{ij})^{M}-1\right)+1}.
\end{aligned} \tag{4.24}$$

Note how the binary projection in all cases corresponds to Bernoulli statistics, $\sigma^2_{\Theta(t_{ij})} = \left\langle \Theta(t_{ij}) \right\rangle (1 - \left\langle \Theta(t_{ij}) \right\rangle)$ with,

$$\left\langle \Theta_{ij} \right\rangle = \begin{cases} \text{ME} & \frac{\tilde{z}_{ij}(e^{Mz_{ij}}-1)}{1+\tilde{z}_{ij}(e^{Mz_{ij}}-1)} \\ \text{W} & \frac{\tilde{z}_{ij}\left((1-z_{ij})^{-M}-1\right)}{1+\tilde{z}_{ij}\left((1-z_{ij})^{-M}-1\right)} \\ \text{B} & \frac{\tilde{z}_{ij}\left((1+z_{ij})^{M}-1\right)}{1+\tilde{z}_{ij}\left((1+z_{ij})^{M}-1\right)} \end{cases} . \tag{4.25}$$

The statistics conditioned on the non-zero value of each occupation number retain their original form (Poisson, Negative Binomial and Binomial), and a useful quantity to analyze is its conditioned average $\langle t_{ij} | t_{ij} > 0 \rangle \equiv \left\langle t_{ij}^{+} \right\rangle$,

$$\left\langle t_{ij}^{+} \right\rangle = \begin{cases} \text{ME} & M\frac{z_{ij}}{1-e^{-Mz_{ij}}} \\ \text{W} & M\frac{z_{ij}}{1-z_{ij}} \frac{1}{\left(1-(1-z_{ij})^{M}\right)} \\ \text{B} & M\frac{z_{ij}}{1+z_{ij}} \frac{1}{\left(1-(1+z_{ij})^{-M}\right)} \end{cases} . \tag{4.26}$$

Note how $\left\langle t_{ij}^{+} \right\rangle$ does not depend on $\tilde{z}_{ij}$, hence the prior expression can be inverted to obtain the relation $z_{ij}(\left\langle t_{ij}^{+} \right\rangle)$ (which will be useful to analyze the high sampling limit and associated entropies afterwards). Sadly, only the ME case yields an analytical expression for arbitrary M,

$$Mz_{ij}^{ME} = W\left(-\left\langle t_{ij}^{+} \right\rangle e^{-\langle t_{ij}^{+} \rangle}\right) + \left\langle t_{ij}^{+} \right\rangle . \tag{4.27}$$

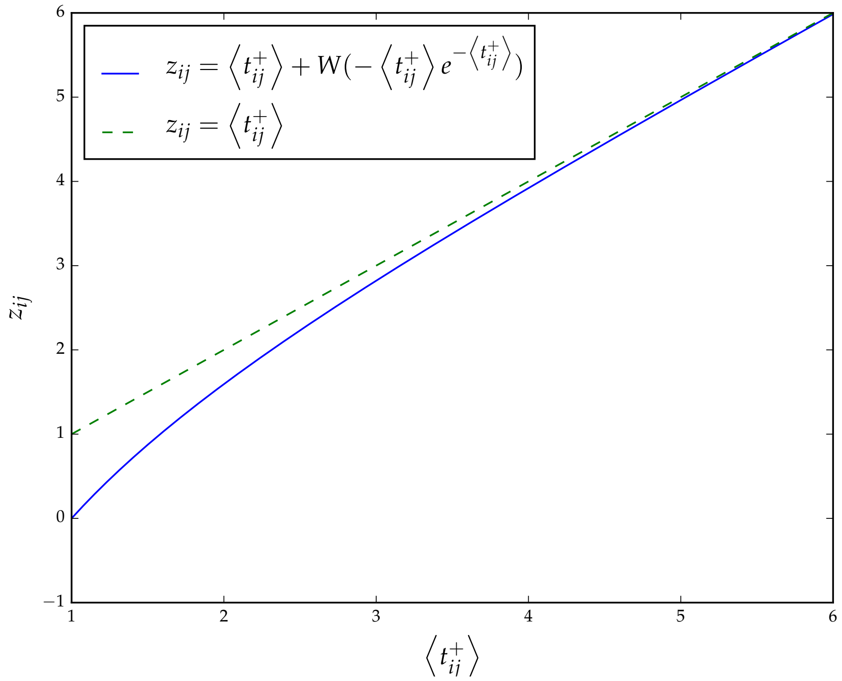

Figure 4.2: **Convergence $z_{ij}(\left\langle t_{ij}^{+} \right\rangle)$ for the ME case with binary constraints.** The result of equation (4.27) is shown together with the equality line $z_{ij} = \left\langle t_{ij}^{+} \right\rangle$. One can see the rapid convergence: For $\left\langle t_{ij}^{+} \right\rangle = 2.31$ we obtain $\frac{z_{ij}}{\left\langle t_{ij}^{+} \right\rangle} = 0.9$ and for $\left\langle t_{ij}^{+} \right\rangle = 4.615$ one finds $\frac{z_{ij}}{\left\langle t_{ij}^{+} \right\rangle} = 0.99$.

Where $W(x)$ corresponds to the Lambert W function [57]. Figure 4.2 shows a plot of equation (4.27) stressing the rapid asymptotical convergence $z_{ij} \to \left\langle t_{ij}^{+} \right\rangle$ as $\left\langle t_{ij}^{+} \right\rangle \to \infty$ (in fact, the approximation is clearly good as soon as $\left\langle t_{ij}^{+} \right\rangle \simeq 5$).

Observe also that

$$\langle t_{ij}^{n} \rangle = \langle \Theta_{ij} \rangle (z_{ij}, \tilde{z}_{ij}) \left\langle (t_{ij}^{+})^{n} \right\rangle, \tag{4.28}$$

so unless $\left\langle t_{ij}^{+} \right\rangle$ is constant, we will encounter important correlations between the binary and non-binary topology. The introduction of $\{\tilde{z}_{ij}\}$ allows us to control separately the binary and non-binary structure of the networks belonging to the ensemble, however, a non-trivial highly non-linear relation will be expected in general between both (they are coupled through the constraint equations). Correlations between occupation numbers and binary topology are often observed in real networks too [160] but one should be careful to consider them as statistically relevant since they may not be a trace of any unexpected feature of the data.

A final note must be made about the macroscopic observation of binary magnitudes at the level of nodes. In general, it is widely accepted that for non-binary networks, a linear relation between strengths[8] $s_i^{\text{out,in}} = \sum_x t_{ij}$ and degrees $k_i^{\text{out,in}} = \sum_x \Theta(t_{ij})$ signals the absence of correlations between non-binary and binary topology. However, this is not exactly true: A condition $\langle s \rangle \propto \langle k \rangle$ is a necessary but not sufficient condition for $\left\langle t_{ij} \right\rangle \propto \left\langle \Theta(t_{ij}) \right\rangle$ and we will see a case where this is made explicit in Section 6.4.

Finally, the average, variance and relative fluctuations of the zero-inflated statistics are readily calculated.

$$\begin{aligned}
\left\langle t_{ij} \right\rangle &= \left\langle \Theta(t_{ij}) \right\rangle \left\langle t_{ij}^{+} \right\rangle \\
&= \begin{cases}
\text{ME} & Mz_{ij} \frac{\tilde{z}_{ij} e^{Mz_{ij}}}{1+\tilde{z}_{ij}(e^{Mz_{ij}}-1)} \\
\text{W} & M\frac{z_{ij}}{1-z_{ij}} \frac{\tilde{z}_{ij}}{(1-z_{ij})^M+\tilde{z}_{ij}(1-(1-z_{ij})^M)} \\
\text{B} & M\frac{z_{ij}}{1+z_{ij}} \frac{\tilde{z}_{ij}}{(1+z_{ij})^{-M}+\tilde{z}_{ij}\left(1-(1+z_{ij})^{-M}\right)}
\end{cases} .
\end{aligned} \tag{4.29}$$

In short notation, considering $a = 0$ for ME, $a = -1$ for B and $a = 1$ for W,

$$\begin{aligned}
\sigma^2_{t_{ij}} &= \left\langle \Theta(t_{ij}) \right\rangle \left\langle t_{ij}^{+} \right\rangle \left\{ 1 + \left\langle t_{ij}^{+} \right\rangle \left(1 - \left\langle \Theta(t_{ij}) \right\rangle + \frac{a}{M} \right) \right\} \\
\frac{\sigma^2_{t_{ij}}}{\left\langle t_{ij} \right\rangle^2} &= \frac{1}{\left\langle t_{ij}^{+} \right\rangle \left\langle \Theta(t_{ij}) \right\rangle} + \frac{1 - \left\langle \Theta(t_{ij}) \right\rangle + \frac{a}{M}}{\left\langle \Theta(t_{ij}) \right\rangle} .
\end{aligned} \tag{4.30}$$

Notice how on the large sampling limit (which implies also $\left\langle t_{ij}^{+} \right\rangle \to \infty$), the *surviving* relative fluctuations are those related with the Bernoulli statistics of the binary structure. These expressions reflect the bimodal structure of the state statistics and helps understanding the non-vanishing relative fluctuations: The relative variance of the occupation numbers has a maximum for $\sigma^2_{t_{ij}} / \left\langle t_{ij} \right\rangle^2 |_{\max} = 1 + 2 \left(\left\langle t_{ij}^{+} \right\rangle^{-1} + a/M \right)$, vanishes for the absence ($\left\langle \Theta(t_{ij}) \right\rangle \to 0$) of an edge and converges to the non-inflated statistics for edges that always exist ($\left\langle \Theta(t_{ij}) \right\rangle \to 1$). The existence of an edge is a binary event, hence the maximum variability correspond to the draw situation (50% chance). In such a case, approximately half of the times a graph is created the considered edge will have (on average) occupation $\left\langle t_{ij}^{+} \right\rangle$ and the other half occupation 0, generating important fluctuations on the overall statistics which are caused by the constrained binary structure of the graph.

[8] Here x will be j for the outgoing case and i for the incoming one.

4.2.3 *Constraint equations and fluctuations in high sampling limit*

The constraint equations in this case become

$$\hat{C}_q = \langle C_q \rangle = \sum_{ij} \left(a_q^{ij} \left\langle t_{ij} \right\rangle (z_{ij}) + \tilde{a}_q^{ij} \left\langle \Theta(t_{ij}) \right\rangle (\tilde{z}_{ij}, z_{ij}) \right) \tag{4.31}$$

And the remarks of the analogous section for the linear case are essentially aggravated: The constraint equations are even more complicated to solve and for the W case, the non-convex maximization domain is still present ($z_{ij} < 1 \, \forall \, ij$).

Analyzing the high sampling limit in this case becomes sketchier than in the previous section. The reason is that the considered constraints are no longer extensive on the sampling $\hat{T}$ so their relative fluctuations will never vanish in the high sampling limit, even for the ME case. Since the analytical formulas become lengthier but no remarkable different insights from the linear case can be gained, we will address specifically these applied directly to the explicit examples considered below.

In the following we provide several examples to illustrate the possible use of GC ensembles with linear and binary constraints and only purely binary constraints. We also relate these ensembles to known models in the literature. The schema to solve the associated constraint equations is provided in Section 6.1.

In all cases, $\tilde{C}_q(\Theta(t_{ij})$ will refer to binary constraints (for which $\{\tilde{a}_q^{ij} = 0\}$) while $C_q(t_{ij})$ will be reserved for linear constraints for which $\{a_q^{ij} = 0\}$, also θ will be the Lagrange multipliers related to the total number of events.

4.2.4 *Explicit examples with binary constraints*

We start by considering only binary constraints and a single non-binary constraint, that of total number of events $\hat{T}$. So $z_{ij} = e^{\theta} = z = \text{Ctnt}$ and $\tilde{z}_{ij} = \prod_q e^{\alpha_q \tilde{a}_q^{ij}}$ (the constraint that the total number of events is fixed needs to always be considered). Note from equation (4.26) that the conditioned averages become independent of each state ij.

$$\left\langle t_{ij}^{+} \right\rangle = \left\langle t^{+} \right\rangle \simeq \frac{\langle T \rangle}{\langle E \rangle} + \mathcal{O}(\langle (T - \langle T \rangle)^3 \rangle) = \text{Ctnt} \implies \left\langle t_{ij} \right\rangle \propto \left\langle \Theta(t_{ij}) \right\rangle . \tag{4.32}$$

This fact has important implications: We obtain proportionality between average occupation and probability of occupation per state. All *occupied* states become then statistically equivalent and hence the distribution of *existing* occupation numbers has a Poisson, Negative Binomial or Binomial form respectively for the ME, W and B cases with average $\langle t^{+} \rangle$, so no heterogeneity in the existing occupation numbers is to be expected in these kind of ensembles. Usually in empirical data [24], the distribution of existing occupation number values is highly skewed, hence the applicability of these models to reproduce these situations is limited.

Concerning the relative fluctuations for the constraint associated to the number of events,

$$\frac{\sigma_T^2}{\langle T\rangle^2} = \hat{T}^{-1} + \langle E\rangle^{-1}\left(1 + \frac{a}{M} - \frac{\sum_{ij}\langle\Theta(t_{ij})\rangle^2}{\langle E\rangle}\right) \tag{4.33}$$

we see how again the fluctuations will not vanish due to the binary nature of the considered phase space.

4.2.4.1 *Global constraints: Non-binary Erdos-Renyi graph. Fixed $\hat{E}, \hat{T}$.*

The simplest ensemble we can consider would be the non-binary analogy to the Erdos-Renyi graph with fixed number of events T and binary connections $E \equiv \sum_{ij}\Theta(t_{ij})$, where all node-pairs are statistically equivalent. In this case $\tilde{z}_{ij} = \tilde{z} = \mathrm{Ctnt}\,\forall\, ij$ and we have two associated equations,

$$\begin{aligned}\hat{T} &= \sum_{ij}\langle t_{ij}\rangle = L\langle t^+\rangle\langle\Theta(t)\rangle \\ \hat{E} &= L\langle\Theta(t)\rangle\,.\end{aligned} \tag{4.34}$$

Which are trivially solved to obtain $\langle\Theta(t)\rangle = \hat{E}/L$ and $\langle t^+\rangle = \hat{T}/\hat{E}$ without need to consider $z, \tilde{z}$. In this case we can evaluate the constraint fluctuations using directly (4.30),

$$\begin{aligned}\frac{\sigma_T^2}{\langle T\rangle^2} &= \hat{T}^{-1} + \hat{E}^{-1}\left(1 + \frac{a}{M}\right) - L^{-1} \\ \frac{\sigma_E^2}{\langle E\rangle^2} &= \hat{E}^{-1} - L^{-1}.\end{aligned} \tag{4.35}$$

And we observe how the relative fluctuations do not vanish and are caused by the binary constraints in all cases.

4.2.4.2 *Local constraints. Soft configuration model graph $\hat{\vec{k}}, \hat{T}$.*

Since all nodes with same degree pair are statistically equivalent, the maximum entropy formula allows to explain the observed natural correlations caused by the configuration model [47].

Another interesting example of these kind of ensembles is that where the node binary topology is enforced through the degree sequence $\vec{k}$. In this case, z is constant and $\tilde{z}_{ij} = e^{\alpha_i+\alpha_j} \equiv \nu_i w_j$. For the probability of existence of each link we have,

$$\langle\Theta(t_{ij})\rangle = \frac{\mu_c \nu_i w_j}{1+\mu_c \nu_i w_j} \qquad \mu_c = \begin{cases}\text{ME:} & e^{Mz}-1 \\ \text{W:} & (1-z)^{-M}-1 \\ \text{B:} & (1+z)^{M}-1\end{cases} \tag{4.36}$$

and we obtain the well-known *Soft Configuration Model* [134, 125]. Concerning the constraint fluctuations we have,

$$\begin{aligned}
\frac{\sigma_T^2}{\langle T\rangle^2} &= \hat{T}^{-1} + \hat{E}^{-1}\left(1+\frac{a}{M}\right) - \frac{\sum_{ij}(\langle\Theta(t_{ij})\rangle)^2}{\hat{E}^2} \\
\frac{\sigma^2_{k_i^{\text{out,in}}}}{\left\langle k_i^{\text{out,in}}\right\rangle^2} &= (\hat{k}_i^{\text{out,in}})^{-1} - \frac{\sum_x \langle\Theta(t_{ij})\rangle^2}{(\hat{k}_i^{\text{out,in}})^2}.
\end{aligned} \tag{4.37}$$

with $x = i, j$ depending on the direction considered for the degrees. Note that in this case the relative fluctuations can become large ($\sim \mathcal{O}(1)$) for lightly connected nodes.

4.2.5 *Explicit examples with linear and binary constraints*

For this last set of examples, we consider networks where both types of constraints are considered. Two relevant cases are reviewed: That where the node non-binary topology is fixed through the strength sequence $\vec{s}$ with an additional requirement that the total number of events in a graph E is fixed and the case where the node topology is determined through the strength and degree sequence pairs $\vec{s}, \vec{k}$. Fluctuations of the constraints will not be discussed since they do not vanish in any case and the expressions can be easily derived from earlier examples with a bit of algebra.

4.2.5.1 *Mixed constraints: Binary constrained non-binary configuration model graph. Fixed $\hat{\vec{s}}, \hat{E}$.*

In this case we fix the total number of binary connections E and the strength sequence $\vec{s}$. Hence, all nodes with the same strength pairs are statistically equivalent and we have $2N + 1$ constraints. Note that the only binary constraint considered is general to the whole graph, hence $\tilde{z}_{ij} = \tilde{z} = \text{Ctnt}$. Yet, this does not mean that either the binary connection probability or the existing average occupation lose their site-dependency ij. We have,

$$\begin{aligned}
z_{ij} &= e^{\theta} e^{\alpha_{q_i^{\text{out}}}} e^{\alpha_{q_j^{\text{in}}}} \equiv x_i y_j \qquad \tilde{z}_{ij} \equiv \tilde{z} \\
\hat{s}_i^{\text{out}} &= \sum_j \langle\Theta(t_{ij})\rangle(\tilde{z}, x_i, y_j)\left\langle t_{ij}^{+}\right\rangle(x_i, y_j) \qquad i = 1, N \\
\hat{s}_j^{\text{in}} &= \sum_i \langle\Theta(t_{ij})\rangle(\tilde{z}, x_i, y_j)\left\langle t_{ij}^{+}\right\rangle(x_i, y_j) \qquad j = 1, N \\
\hat{E} &= \sum_{ij} \langle\Theta(t_{ij})\rangle(\tilde{z}, x_i, y_j).
\end{aligned} \tag{4.38}$$

4.2.5.2 *Local constraints. Enhanced soft configuration model graph. Fixed* $\hat{\vec{k}}, \hat{\vec{s}}$.

A final case to consider is that where both the node topology at the binary and non-binary level is fixed: We control both the degree $\vec{k}$ and strength $\vec{s}$ pair of each node. Such an ensemble has been termed for the W case as *enhanced configuration model* [122]. For the general example, we have 4N constraints and only nodes with equal strength and degree pairs are statistically equivalent.

$$
\begin{aligned}
z_{ij} &= e^{\theta} e^{\alpha_{q_i^{out}}} e^{\alpha_{q_j^{in}}} \equiv x_i y_j \qquad \tilde{z}_{ij} = e^{\tilde{\theta}} e^{\tilde{\alpha}_{\tilde{q}_i^{out}}} e^{\tilde{\alpha}_{\tilde{q}_j^{in}}} \equiv v_i w_j \\
\hat{s}_i^{out} &= \sum_j \langle \Theta(t_{ij}) \rangle (\tilde{z}_{ij}, z_{ij}) \left\langle t_{ij}^{+} \right\rangle (z_{ij}) \qquad i = 1, N \\
\hat{s}_j^{in} &= \sum_i \langle \Theta(t_{ij}) \rangle (\tilde{z}_{ij}, z_{ij}) \left\langle t_{ij}^{+} \right\rangle (z_{ij}) \qquad j = 1, N \\
\hat{k}_i^{out} &= \sum_j \langle \Theta(t_{ij}) \rangle (\tilde{z}_{ij}, z_{ij}) \qquad i = 1, N \\
\hat{k}_j^{in} &= \sum_i \langle \Theta(t_{ij}) \rangle (\tilde{z}_{ij}, z_{ij}) \qquad j = 1, N
\end{aligned}
\tag{4.39}
$$

This is the most complicated case we consider in this thesis. Note a general correlation between $\langle t_{ij} \rangle$ and $\langle \Theta(t_{ij}) \rangle$, which cannot be simplified to obtain approximated analytical solutions in any of the considered cases.

4.3 MAXIMUM LIKELIHOOD SOLUTION FOR DUAL PROBLEM

All the obtained statistics are derived from a maximum entropy methodology, and hence the values of $\vec{\alpha}$ obtained from (3.6) fulfill in all cases the *maximum likelihood principle* for model selection for networks [83]. Having a set of candidate statistical models depending on some parameters $\vec{\alpha}$ we can choose the set of parameters $\vec{\alpha}^*$ for each model which maximize the probability in each model to observe $\hat{\vec{T}}$ in the Ω space (note that the degeneracy term does not depend on $\vec{\alpha}$, hence we may equivalently maximize the probability to observe the fixed constraints $\hat{\vec{C}}$ in the Γ space probabilities),

$$
\begin{aligned}
&\mathcal{P}^{\Omega}(\hat{\vec{T}}|\vec{\alpha}) = \frac{\mathcal{D}(\hat{\vec{T}}) e^{\vec{\alpha} \cdot \vec{C}(\hat{\vec{T}})}}{\mathcal{Z}_{GC}(\vec{\alpha})} \qquad \mathcal{P}^{\Gamma}(\hat{\vec{C}}|\vec{\alpha}) = \frac{e^{\vec{\alpha} \cdot \vec{C}(\hat{\vec{T}})}}{\mathcal{Z}_{GC}(\vec{\alpha})} \\
&\mathcal{L}^{\Omega}(\hat{\vec{T}}|\vec{\alpha}) = \ln \mathcal{P}^{\Omega}(\hat{\vec{T}}|\vec{\alpha}) = \ln \mathcal{D}(\hat{\vec{T}}) + \mathcal{L}^{\Gamma}(\hat{\vec{C}}|\vec{\alpha}).
\end{aligned}
\tag{4.40}
$$

Upon maximization of the prior expression for each variable $\{\alpha_q\}$ we get,

$$\begin{aligned} \partial_{\alpha_q}\mathcal{L}^{\Omega}(\hat{\vec{T}}|\vec{\alpha})\Big|_{\vec{\alpha}^*} = 0 \implies \partial_{\alpha_q}\mathcal{L}^{\Gamma}(\hat{\vec{C}}|\vec{\alpha})\Big|_{\vec{\alpha}^*} = 0 \\ \vec{C}_q(\hat{\vec{T}}) \equiv \hat{C}_q = \partial_{\alpha_q}\ln \mathcal{Z}_{GC}(\vec{\alpha}) = \langle C_q\rangle\,(\vec{\alpha}^*)\,\forall\, q = 1, Q. \end{aligned} \tag{4.41}$$

We see then that the form of the associated equations for the log-likelihood maximization problem is no other than the constraint equations for the statistical models with a maximum entropic form, given the form of the constraints and of the degeneracy term. This fact provides us with a useful numerical way of solving the saddle point equations by maximization of the scalar function $\mathcal{L}^{\Gamma}(\hat{\vec{C}}|\vec{\alpha})$, as will be discussed in Section 6.1.

4.4 WRAPPING UP: GRAND CANONICAL ENSEMBLE MAIN FEATURES

In this chapter we have developed the Grand-Canonical ensemble formalism to generate maximally random networks which fulfill some prescribed constraints $\hat{\vec{C}}$ on average which may be written as linear functions of the individual occupation numbers $\{t_{ij}\}$ and/or their binary projections $\{\Theta(t_{ij})\}$. The main strengths of this formalism have become apparent:

A. **Specificity:** Poisson (ME), Negative Binomial (W) and Binomial (B) explicit statistics and their zero-inflated version have been obtained in each case. These recover previously studied ensembles in the literature. Moreover, they are distinctively different and display important differences in the observed macroscopic network features, as will be shown in Chapter 6.

B. **Flexibility:** We have presented a range of examples where the models provided can be useful relating them to existing models in the literature. This gives a glimpse on the generality and flexibility of the approach to modeling given here.

C. **Equivalence:** We have also seen how the obtained statistics become equivalent in the limit where the total number of layers becomes extremely large, which leads to the ME case.

D. **Maximum likelihood equivalence:** I have discussed how the maximum entropy problem can also be related to a maximum likelihood problem, which *en passant* provides a useful way to attempt the numerical solving of the constraint equations and to prove several interesting properties of the maximization problem.

E. **Effective degeneracy for distinguishable case:** The need to add an effective degeneracy term, taking into account the infinite degeneracy of configurations with equivalent $\vec{T}$ while considering an infinite reservoir (or number of system copies) of distinguishable events has been discussed. This fact is capitally important and its derivation constitutes a renewed view on the derivation of Maxwell-Boltzmann statistics for the Grand Canonical ensemble of equilibrium classical statistical mechanics from a purely statistical view (an extended discussion is provided in Section A.1).

However, also some weaknesses have been detected:

A. **Parameter dependence:** All the treatment provided depends capitally on the ability to solve the associated constraint equations which allow to obtain $\vec{\alpha}$. This will be discussed in detail in Chapter 6.

B. **Non-vanishing fluctuations:** We have seen that all the cases dependent on the binary projection of the occupation numbers, and even the ones dependent on the occupation numbers for the W and B case, lead to non-vanishing relative fluctuations in the infinite sampling limit, indicating some problems in its definition for some of the ensembles. This feature will be discussed at length in the next Chapter 5.

C. **Explicit entropy calculations:** Explicit calculations of entropy have not been developed and are left for Chapter 5.

From the practitioner point of view, all the results obtained in this chapter in conjunction with the recipes given in Chapter 6 will be useful for model generation. One may choose from a wide variety of situations and with the ingredients provided (most of them implemented in [10]) generate null models in a *relatively simple* way.

From the theoretical point of view, the obtained results can also help in relating known network models to maximum entropy models and/or justify repeatedly observed features in empirical data.

5 MICRO CANONICAL ENSEMBLES

Only entropy comes easy.

— Anton Chekov

The last ensemble we have left to review is the Micro Canonical one (MC). This ensemble does not accept fluctuations of the constraints and hence calculations are consideratly more complicated than the previously studied cases.

The use of this ensemble, however, has some important advantages: First it allows a transparent computation of entropies and second, it can be used to understand the connection between the GC ensemble and the present one in a transparent way.

It also allows to obtain well defined entropies for the probability of obtaining a given network configuration in the Γ space without resorting to an effective degeneracy term for the case where events are distinguishable.

Our main interest will be to compute the entropy per event in this ensemble. More specifically, we are interested in studying the asymptotic behaviour of the micro canonical entropy as we approach the high sampling limit, which counts exactly the logarithm of the number of network configurations strictly compatible with the chosen constraints.

$$S^{\Gamma}[\mathcal{P}_H] = \ln \mathcal{Z}_{MC}. \tag{5.1}$$

While doing so, we will also discuss the relation between the statistics of the GC ensemble and the conditions under which both can be regarded as equivalent.

5.1 UNCOVERING THE RELATION BETWEEN ENSEMBLES: CALCULATING THE MICRO CANONICAL PARTITION FUNCTION

The procedure to obtain explicit statistics is two-fold: Firstly, we will use integral representations of the Kronecker deltas to enforce the hard constraints of the MC ensemble which will lead us to path-dependent integrals in the complex plane. We will then try to solve or approximate these integral forms of the partition function using a *steepest descent method*.

5.1.1 *Integral form representation*

Taking expression (3.3) for the (uniform) probability to sample a graph in the MC ensemble and considering the partition function $\mathcal{Z}$ one has,

$$\mathcal{Z}_{MC} = \sum_{\Omega} \mathcal{D}(\vec{T})\delta_{\vec{C}(\vec{T}),\hat{\vec{C}}}\,. \tag{5.2}$$

Where $\hat{\vec{C}}$ are the exact values to which one wants to fix the constraints for each network realization sampled from the ensemble. To do so, we can express the Kroenecker deltas in its integral form,

$$\delta_{x\hat{x}} = \oint D\omega_x \omega_x^{\hat{x}-x-1}, \tag{5.3}$$

considering a counter-clockwise contour integral around the origin in the complex plane. Introducing them into the prior expression[1], we obtain

$$\mathcal{Z}_{MC} = \oint D\vec{\omega} \frac{1}{\prod_q \omega_q^{\hat{C}_q+1}} \sum_{\Omega} \prod_q \omega_q^{C_q(\vec{T})}. \tag{5.4}$$

We see that the prior expression is related with the grand canonical partition function of the ensemble considering the transformation $\omega_q = e^{\alpha_q}\,\forall q = 1, Q$. We may thus write

$$\mathcal{Z}_{MC} = \oint D\vec{\alpha} \exp\left(-\vec{\alpha}\cdot\hat{\vec{C}} + \ln \mathcal{Z}_{GC}(\vec{\alpha})\right) \equiv \oint D\vec{\alpha} e^{G(\vec{\alpha})}. \tag{5.5}$$

Furthermore, for the cases studied in this thesis, the partition functions can be summed and are holomorphic in the domains where they are defined. Under these conditions and assuming that $G(\vec{\alpha})$ has a unique minimum at $\vec{\alpha}^*$ along the real axis[2] , we may represent it as a Taylor expansion around this point[3],

$$\begin{aligned} \mathcal{Z}_{MC} &= e^{G(\vec{\alpha}^*)} \oint D\vec{\alpha} \exp\left(\sum_{n=2}^{\infty} \frac{1}{n!} \sum_{\{q\}_n} \kappa_{C_q, C_{q'}, \ldots, C_{q^n}}|_{\vec{\alpha}^*} \prod_i^n (\alpha_i - \alpha_i^*)\right) \\ &\equiv e^{G(\vec{\alpha}^*)} e^{\Delta S^{\Gamma}} \\ &\partial_{\alpha_q, \alpha_{q'}, \ldots, \alpha_{q^n}} G(\vec{\alpha})|_{\vec{\alpha}^*} \equiv \kappa^n_{C_q C_{q'} C_{q''} \ldots C_{q^n}}|_{GC}. \end{aligned} \tag{5.6}$$

Taking logarithms, we reach:

$$S^{\Gamma}[\mathcal{P}_H] = \ln \mathcal{Z}_{MC} = G(\vec{\alpha}^*) + \Delta S^{\Gamma}. \tag{5.7}$$

[1] On the following, $D\vec{\omega}$ will refer to an appropriately normalized differential.

[2] In the next chapter we justify that this indeed holds for the ME and B cases, while is likely for the W case under sufficient plausible conditions and finite sampling.

[3] To make the handling of the forthcoming equations easier, we use the following convention for indexes of the sums $\{q\}_n$: This means all the unique groupings of elements of $\{q\}$ $(q = 1, Q)$ taken in groups of n, where q can be repeated and the order is unimportant.

At this point, it is useful to recall the work done for the GC ensemble. In particular, one can see from (4.40) that $G(\vec{\alpha})$ is in fact minus the log-likelihood function of obtaining a network with given constraints $\hat{\vec{C}}$ under the model given by the GC ensemble and parameters $\vec{\alpha}$. Moreover, its derivatives with respect to $\vec{\alpha}$ give rise to the cumulants of the constraints generated by the GC version of the ensembles with parameters $\vec{\alpha}$, which, at the extreme point $\vec{\alpha}^*$ fulfil the constraint equations:

$$\begin{aligned} G(\vec{\alpha}) &\equiv -\mathcal{L}^{\Gamma}(\hat{C}|\vec{\alpha}) \\ \partial_{\alpha_q} G(\vec{\alpha})|_{\vec{\alpha}^*} &= \partial_{\alpha_q} \ln \mathcal{Z}_{GC}|_{\vec{\alpha}^*} - \hat{C}_q = \langle C_q \rangle_{GC} - \hat{C}_q = 0 \\ \partial_{\alpha_q,\alpha_{q'}} G(\vec{\alpha})|_{\vec{\alpha}^*} &= \partial_{\alpha_q \alpha_{q'}} \ln \mathcal{Z}_{GC}|_{\vec{\alpha}^*} = -\sigma^2_{C_q C'_q}|_{GC} \\ &\dots \\ \partial_{\alpha_q,\alpha_{q'},\dots,\alpha_{q^n}} G(\vec{\alpha})|_{\vec{\alpha}^*} &= \kappa^n_{C_q C_{q'} C_{q''} \dots C_{q^n}}|_{GC}. \end{aligned} \tag{5.8}$$

We can now observe how the entropies of the both ensembles are related, noting that $G(\vec{\alpha}^*) = -\mathcal{L}^{\Gamma}(\hat{C}|\vec{\alpha}^*) = -\left\langle \ln \mathcal{P}^{\Gamma}_{S}(\vec{C}(\vec{T})) \right\rangle_{GC} = S^{\Gamma}[\mathcal{P}_S]$ is in fact, the entropy of the soft-constrained ensemble.

$$\ln \mathcal{Z}_{MC} = S^{\Gamma}[\mathcal{P}_H] = S^{\Gamma}[\mathcal{P}_H] + \Delta S^{\Gamma}. \tag{5.9}$$

We are thus left with the analysis of the object ΔS^{Γ}, which we will call *entropy excess* of the ensemble and is the contribution to the integral outside the point $\vec{\alpha} = \vec{\alpha}^*$. A specially indicated contour for the complex plane integration to analyze it will be that passing through the point $\vec{\alpha}^*$. Let's consider such a circuit $q(\vec{\alpha})$, which crosses the point $\vec{\alpha}^*$ in the direction of the imaginary axis. By virtue of Cauchy's theorem $\vec{\alpha}^*$, has to be a saddle point given that it is a minimum in the direction along the real axis, so, given that the extremum point is unique, as a first approximation, we may consider that the majority of the contribution of the integral along the path will come from the immediate neighborhood of this point. Proceeding in this way, we are using Laplace's argument (or conversely a steepest descent approximation)[4].

5.1.1.1 *Conditions for ensemble equivalence*

In a strict sense, the above reasoning will be exact if we can deform the integration path of the complex integrals in such a way that all the terms in the expansion (5.6) are negligible in the high sampling limit. By negligible, we mean that the contribution of the maxima overwhelms the rest of the contour, i. e.,

$$\lim_{\hat{T}\to\infty} \frac{\Delta S^{\Gamma}}{G(\vec{\alpha}^*)} = 0. \tag{5.10}$$

[4] The description given here of the steepest descent approximation is crude and aims at exploring the relation between ensembles. For a more detailed and technical discussion one may consult classical books, as for instance [37, 72].

If the maximum $G(\vec{\alpha}^*)$ grows with our scaling variable $\hat{T}$, in such a way that in the high sampling limit it becomes *very steep*, then, we can assume that the contribution of the immediate neighborhood of the point will dominate the rest of the (unspecified) contour. To look further into this possibility, let's consider for the moment a given (unspecified) transformation $g(\hat{T})\vec{\beta} = \vec{\alpha}^* - \vec{\alpha}$ on the integral in (5.6) which leads to,

$$e^{\Delta S^\Gamma} = \frac{1}{(g(\hat{T}))^Q} \oint D\vec{\beta} \exp\left(\sum_{n=2}^{\infty} \frac{1}{n!} \sum_{\{q\}_n} \frac{\kappa^n_{C_q, C_{q'}, C_{q''}, \dots, C_{q^n}}|_{GC}}{\left(g(\hat{T})\right)^n} \prod_i^n \beta_i \right). \tag{5.11}$$

Thus we observe that the main matter of discussion will be the scaling of the ratio between the joint cumulants of $\vec{C}$, $\kappa^n_{C_q, C_{q'} \dots C_{q^n}}|_{GC}$ and $\left(g(\hat{T})\right)^n$ inside the integral. Yet, by general properties of the cumulants,

$$\frac{\kappa^n_{C_q, C_{q'} \dots C_{q^n}}|_{GC}}{\left(g(\hat{T})\right)^n} = \kappa^n_{c_q, c_{q'} \dots c_{q^n}}|_{GC}. \tag{5.12}$$

so we observe that the saddle point approximation will only be accurate if the cumulants of the scaled constraints $\vec{c} = \vec{C}(\hat{T})/g(\hat{T})$ of order higher than one vanish with increased sampling. Obviously, any joint cumulant related with binary constraints will vanish in such a limit, as the binary constraints (assuming they are graphical) do not scale with the number of events. Hence, the main matter of concern will be the linear constraints, which being extensive, require a scaling $g(\hat{T}) = \hat{T}$. If the first term of the series, which is the largest contribution to the integral in the neighborhood of the maxima, does not vanish asymptotically, then the series will not do so either. Hence we require:

$$\lim_{\hat{T} \to \infty} \frac{\kappa^2_{C_q, C_{q'}}|_{GC}}{\hat{T}^2} = 0 \, \forall \, q, q' = 1, Q. \tag{5.13}$$

Note that by general properties of the cumulants, if this term vanishes, so will all the others, (and if it does not, the others will not)[5]. Hence, if this condition is met, we may thus truncate the expansion up to a second term to obtain a closed integral (assuming a contour parallel to the imaginary axis passing through $\vec{\alpha}^*$ joined by an arc at either side, whose contribution vanishes):

$$e^{\Delta S^\Gamma} = (2\pi)^{Q/2} \left(\det \Sigma^2(\vec{\alpha}^*)) \right)^{-1/2}. \tag{5.14}$$

[5] Note that assuming that the cumulant series can be truncated amounts to consider the joint distribution of $\mathcal{P}(\vec{C})$ are Gaussian, since these distributions are the only non-trivial ones for which the cumulant generating function is a finite-order polynomial (of degree two) [120].

Where $\det \Sigma^2(\vec{\alpha}^*)$ refers to minus the determinant of the Hessian matrix of $G(\vec{\alpha})$ evaluated at the saddle point. Each of the elements of this matrix are known from (5.8) and equate to $\sigma^2_{C_q, C_{q'}}$ for each entry q, q' of the $Q \times Q$ constraint correlation matrix Σ^2.

Note, however, that condition (5.10) is weaker than condition (5.13). For this last case, when evaluating the asymptotic entropy per event (5.1), we will obtain equivalent entropies for the ensembles if both the total number of constraints Q and the constraint fluctuations scale sub-linearly with $G(\vec{\alpha}^*)$. Even if relative fluctuations do not disappear, in particular cases we might still be able to find an appropriate path such that the condition of asymptotic equivalence of entropies is met, yet these cases must be reviewed on a one to one basis. The ME case with linear constraints in either its Grand Canonical or Canonical version is the only one for which condition (5.13) is fulfilled, so the rest of cases must be addressed in a different manner.

5.1.2 *Steepest descent approximation: Evaluation of Grand Canonical entropies*

We can still try to approximate the integral using a steepest descent argument for the W and B cases, and for the ME case with binary constraints, which do not display vanishing relative occupation number fluctuations. To do so, we need informally that the contribution to the integral is concentrated around the maximum value of $G(\vec{\alpha})$ in the chosen direction, so, in the following, we review on a case by case basis the asymptotical behaviour of the maximum $G(\vec{\alpha}^*)$, which is the entropy of the GC ensemble.

5.1.2.1 *Linear constraints:*

For cases where linear constraints only are considered, by our definition of large sampling limit we have $\langle t_{ij} \rangle (z(\vec{\alpha}^*)_{ij}) \sim \mathcal{O}(\hat{T})$. Moreover, at $\vec{\alpha}^*$, $\left\langle \vec{C} \right\rangle = \hat{\vec{C}}$ and recovering (4.7),

$$z^*_{ij} = \prod_q e^{\alpha^*_q a^{ij}_q} = \begin{cases} \text{ME:} & \frac{\langle t_{ij} \rangle}{M} \\ \text{W:} & \frac{\frac{\langle t_{ij} \rangle}{M}}{1+\frac{\langle t_{ij} \rangle}{M}} \\ \text{B:} & \frac{\frac{\langle t_{ij} \rangle}{M}}{1-\frac{\langle t_{ij} \rangle}{M}} \end{cases} \tag{5.15}$$

$$-\vec{\alpha}^* \cdot \hat{C} = -\sum_{ij} \langle t_{ij} \rangle \ln z^*_{ij}(\langle t_{ij} \rangle),$$

so we can evaluate asymptotically the limiting behaviour[6],

$$G(\vec{\alpha}^*) = -\sum_{ij} \left(\langle t_{ij}\rangle \ln z^*_{ij}(\langle t_{ij}\rangle) - \ln \mathcal{Z}_{ij}(z^*_{ij}(\langle t_{ij}\rangle))\right)$$

$$\begin{cases} \text{ME:} & -\sum_{ij}\left(\langle t_{ij}\rangle \ln \langle t_{ij}\rangle - \langle t_{ij}\rangle \ln M - \langle t_{ij}\rangle \ln \hat{T}\right) + \mathcal{O}(\frac{1}{2}\ln \hat{T}) \\ & \sim \mathcal{O}(\hat{T}) \\ \text{W:} & -M\sum_{ij}\left(\frac{\langle t_{ij}\rangle}{M} \ln \frac{\frac{\langle t_{ij}\rangle}{M}}{1+\frac{\langle t_{ij}\rangle}{M}} - \ln\left(1+\frac{\langle t_{ij}\rangle}{M}\right)\right) \\ & \sim \mathcal{O}(\ln \hat{T}) \\ \text{B:} & -M\sum_{ij}\left(\frac{\langle t_{ij}\rangle}{M} \ln \frac{\frac{\langle t_{ij}\rangle}{M}}{1-\frac{\langle t_{ij}\rangle}{M}} + \ln\left(1-\frac{\langle t_{ij}\rangle}{M}\right)\right) + d\ln(\hat{T}!) \\ & = d\ln \hat{T}! \end{cases} . \tag{5.16}$$

Where $d = 1$ if the events are distinguishable or $d = 0$ otherwise for the B case. We observe how while the ME case is well behaved, this is not the case of the other two. For the W case, the maximum scales slowly with $\hat{T}$, while for the B case it does not scale at all. This indicates that the approximation of truncating the first terms in the expansion (5.6) is not sufficient.

Note also that the above expression coincides with our discussion in the previous chapter on scaling with the number of layers. If for the W and B cases, we take $M \sim \mathcal{O}(\hat{T})$, then the approximation is exact for all models (they all converge to the ME case).

Considering the leading terms in $\hat{T}$ of the GC entropy per event we obtain,

$$\lim_{\hat{T}\to\infty} \frac{S^{\Gamma}[\mathcal{P}_S]}{\hat{T}} =$$

$$\begin{cases} \text{ME (all):} & -\lim_{\hat{T}\to\infty}\sum_{ij}\frac{\langle t_{ij}\rangle}{\hat{T}} \ln \frac{\langle t_{ij}\rangle}{\hat{T}} + \ln M = -\sum_{ij} p^{\infty}_{ij} \ln p^{\infty}_{ij} + \ln M \\ \text{W:} & -\lim_{\hat{T}\to\infty}\frac{M}{\hat{T}}\sum_{ij}\left(\ln \frac{M}{\langle t_{ij}\rangle} - 1\right) = 0 \\ \text{B:} & -\lim_{\hat{T}\to ML, \langle t_{ij}\rangle \to M}\frac{1}{\hat{T}M}\sum_{ij}\left(\langle t_{ij}\rangle (M-\langle t_{ij}\rangle)\right)\left(\ln \frac{M-\langle t_{ij}\rangle}{M} - 1\right) \\ & +d\frac{\ln T!}{\hat{T}} = d\frac{\ln(LM)!}{LM} \end{cases} . \tag{5.17}$$

We observe how only the ME case takes a Shannon, state-specific form (with a constant offset $\ln M$ generated by the possible ways to allocate an event into the M state layers) while the other cases lead to vanishing asymptotic entropies per event.

[6] We do not cover here the Canonical case which leads to multinomial statistics, treated in the earlier chapter, since it is obvious that by construction it fulfills the requirements for the saddle point approximation to be exact.

For the Binary case, the limit corresponds to a fully connected network with all occupation numbers equal to M, because by construction $\hat{T} \leqslant LM$, leading to a single configuration in the indistinguishable case, while in the distinguishable case one needs to add a factor accounting for the permutation of (different) events among the L states and M layers. In this case, the fluctuations of the ensemble disappear and only one configuration is available in the MC ensemble, so $\mathcal{Z}_{MC} = 1$ and thus the entropy vanishes (or is constant in the case of distinguishable events). For the GC ensemble, in general, as soon as $\hat{T} \geqslant ML$ no feasible solutions to the saddle point equations will exist. This is a clear indication that for this case, the high sampling limit is not well defined and thus the ensembles in general will not display adequate asymptotic properties.

For the Weighted case, even if the high sampling limit is correctly defined ($\hat{T}$ can grow indefinitely) we also reach vanishing asymptotic entropies per event. Taking into account (5.15), we have that $z^*_{ij} \to 1$ as $\langle t_{ij} \rangle \to \infty$ and thus $\max \left\{ z^*_{ij} \right\}$ reaches this point in the first place for the most occupied state, and this generates a concentration of all events on this single state. So, asymptotically, as more events are added to the system, they are allocated to the same state (only possible configuration) in a process akin to a Bose-Einstein condensation. In this situation, again, the saddle point equations in general will not heave feasible solutions and thus the constraints will in general be violated.

5.1.2.2 *Linear and binary constraints:*

For binary constraints, the above reasoning must be adapted. If the saddle point equations have solutions, then one expects the binary probability of occupation to be well defined, hence $0 \leqslant \langle \Theta(t_{ij}) \rangle < 1$, so for all the cases considered we will be able to separate $G(\vec{\alpha}^*)$ into two separate contributions, one of which will not scale with $\hat{T}$. Noting,

$$\ln \mathcal{Z}_{ij} \; = d\frac{\ln \hat{T}!}{L} - \ln(1 - \langle \Theta(t_{ij}) \rangle). \tag{5.18}$$

and proceeding in a similar way as previously,

$$\begin{aligned}G(\vec{\alpha}^*) &= -\sum_{ij}\left(\langle t_{ij}\rangle \ln z^*_{ij} + \sum_{ij}\langle\Theta(t_{ij})\rangle \ln \tilde{z}^*_{ij} - \ln \mathcal{Z}_{ij}\right) = \\ &= -\sum_{ij}\left(\langle\Theta(t_{ij})\rangle \ln\langle\Theta(t_{ij})\rangle + (1-\langle\Theta(t_{ij})\rangle)\ln(1-\langle\Theta(t_{ij})\rangle)\right) - \\ &- \sum_{ij}\left(\langle t_{ij}\rangle \ln z^*_{ij} - \langle\Theta(t_{ij})\rangle \ln \Delta^*_{ij}\right) + d\ln T! \\ &= S^{\Gamma}_{\text{bin}}[\mathcal{P}_S] + S^{\Gamma}_{\text{non-bin}}[\mathcal{P}_S] \\ d &= \begin{cases}\text{Dist:} & 1 \\ \text{Indist:} & 0\end{cases} \qquad \Delta^*_{ij} = \begin{cases}\text{ME:} & e^{Mz^*_{ij}} - 1 \\ \text{W:} & (1-z^*_{ij})^{-M} - 1 \\ \text{B:} & (1+z^*_{ij})^{M} - 1\end{cases} .\end{aligned} \tag{5.19}$$

We observe two contributions to the maximum, one coming from the binary form of the statistics (commonly found also when dealing with ensemble descriptions of binary networks), which is not expected to scale[7] with $\hat{T}$, and the other from the non-binary statistics. In an explicit form, we have,

$$\begin{aligned}&S^{\Gamma}_{\text{non-bin}}[\mathcal{P}_S] = \\ &\begin{cases}\text{ME:} -\sum_{ij}\left(\langle t_{ij}\rangle \ln z^*_{ij} - \langle\Theta(t_{ij})\rangle \ln(e^{Mz^*_{ij}}-1)\right) + \ln\hat{T}! \\ \text{W:} -\sum_{ij}\left(\langle t_{ij}\rangle \ln z^*_{ij} - \langle\Theta(t_{ij})\rangle \ln((1-z^*_{ij})^{-M}-1)\right) \\ \text{B:} -\sum_{ij}\left(\langle t_{ij}\rangle \ln z^*_{ij} - \langle\Theta(t_{ij})\rangle \ln((1+z^*_{ij})^{M}-1)\right) + d\ln\hat{T}!\end{cases}\end{aligned} \tag{5.20}$$

The previous expressions are hard to interpret due to the *hidden* dependency between z^*_{ij} and $\left\langle t^+_{ij}\right\rangle$, but in some cases we can study limiting behaviours. For the ME case, we are able to invert the relation $\left\langle t^+_{ij}\right\rangle(z^*_{ij})$ (expression (4.27)) for the general case, so $\left\langle t^+_{ij}\right\rangle_{ME} \simeq Mz^*_{ij}$ and $e^{\langle t^+_{ij}\rangle} - 1 \simeq e^{\langle t^+_{ij}\rangle}$. For the W and B case we shall consider the

[7] If the binary structure is not completely determined by the binary constraints, then for some values ij we will have in the high sampling limit that $\langle\Theta(t_{ij})\rangle \to 1$, yet, in this case and given the factorization of the partition function, we can always extract this contribution and place it with the part of $G(\vec{\alpha}^*)$ that scales with $\hat{T}$.

case $M = 1$,which is easily invertible ($\left\langle t^+_{ij} \right\rangle_B = 1$ and $\left\langle t^+_{ij} \right\rangle_W = 1/(1-z^*_{ij})$).

$$S^{\Gamma}_{\text{non-bin}} \simeq \begin{cases} \text{ME:} & \hat{T}\left(\ln M - \sum_{ij} p^{\infty}_{ij} \ln p^{\infty}_{ij} + \sum_{ij} p^{\infty}_{ij} \ln \left\langle \Theta(t_{ij}) \right\rangle\right) + \mathcal{O}(\frac{1}{2}\ln \hat{T}) \\ & \sim \mathcal{O}(\hat{T}) \\ \text{W:} & \sum_{ij} \left\langle \Theta(t_{ij}) \right\rangle \left(\ln \left\langle t^+_{ij} \right\rangle - 1\right) \sim \mathcal{O}(\ln \hat{T}) \\ \text{B:} & -\sum_{ij} \left(\left\langle \Theta(t_{ij}) \right\rangle \ln z^*_{ij} - \left\langle \Theta(t_{ij}) \right\rangle \ln((1+z^*_{ij})^M - 1)\right) + d \ln \hat{T}! \\ & = d \ln \hat{T}! \end{cases} . \tag{5.21}$$

Basically the same result encountered earlier are repeated. We have again vanishing entropies for the W and B case.

For the ME case, we see that the usual form for linear constraints is recovered with a decrease in entropy induced by limited number of events that can be occupied once the binary structure has been set.

$$\begin{aligned} \lim_{\hat{T}\to\infty} \frac{S^{\Gamma,ME}_{\text{non-bin}}}{\hat{T}} &= \ln M - \sum_{ij} p^{\infty}_{ij} \ln \frac{p^{\infty}_{ij}}{\left\langle \Theta(t_{ij}) \right\rangle} \\ &= \ln M - \sum_{ij} p^{\infty}_{ij} \ln p^{\infty}_{ij} + \sum_{ij} p^{\infty}_{ij} \ln \left\langle \Theta(t_{ij}) \right\rangle . \end{aligned} \tag{5.22}$$

Again, we see an approximate linear scaling of the maximum (despite having non relative fluctuations of the individual occupation numbers), so while the quality of the approximation will depend in the general case, we can expect the truncation of the series in (5.6) not to have dramatic effects (but one would need to analyze the behaviour of each particular case to be sure).

In these cases, the non-vanishing relative fluctuations of the linear constraints are clearly caused by an ergodicity breaking in our sampling process: Being the existence of an edge an event drawn from a Bernoulli distribution, if some states are not occupied when the binary structure is drawn, then no matter how much sampling we obtain (observation time we wait), there will always be configurations of the Γ space where these states are occupied and which are compatible with the binary constraints which will be not observed.

As a final note, one can check that as expected, for both linear and binary constraints, all the obtained entropies converge if instead of the high sampling limit we consider the sparse case ($\left\langle t_{ij} \right\rangle \to \left\langle \Theta(t_{ij}) \right\rangle$). In particular, $\left\langle t^+_{ij} \right\rangle \to 1 \implies Mz^*_{ij} \to 0, \hat{T} \to \langle E \rangle$, with $\langle E \rangle$ being the expected number of occupied binary edges. All what is left of the non-binary entropy contribution besides the effect of the multi-layered structure is the logarithm of the permutation of events on (single occupied) states for the distinguishable case. For the indistinguishable cases the non-binary entropy is zero as again a single configuration can be built once the binary structure is fixed.

$$\lim_{\hat{T}\to\hat{E}} \frac{S^{\Gamma}_{\text{non-bin}}}{\hat{T}} - \ln M = \\ -\lim_{z^*_{ij}\to 0} \sum_{ij} \frac{\langle \Theta(t_{ij}) \rangle \left(\frac{1}{2}(M \pm a) z^*_{ij} (\ln z^*_{ij} - 1) \right)}{\hat{E}} + d \ln \hat{E}! = d \ln \hat{E}! \tag{5.23}$$

where a corresponds to the usual notation for the different cases ($a = 0$ ME, $a = 1$ W and $a = -1$ B).

5.2 EXPLICIT COMPARISON OF ENSEMBLE ENTROPIES

To conclude this chapter, we proceed to compare Γ entropies of the hard and soft constrained versions of two examples which are simple enough to provide a complete analytical overview of the problems of ensemble equivalence. In these cases, the MC partition functions can be exactly computed while the GC entropy can be approximated in the high sampling limit by use of (5.16) and (5.19), so we can analyze the asymptotic behaviour of the entropy excess ΔS^{Γ}. This will allow us to compare the behaviour of the contributions to the integral around its maximum and at the maximum $\vec{\alpha}^*$ itself.

5.2.1 *Linear constraints: Fixed* $\hat{T}$

The model with fixed number of events is the simplest case with linear constraints studied in this thesis, but already provides interesting insights about the behaviour of the different cases. The MC entropies can be exactly evaluated using the Residue Theorem in (5.4)[8],

$$\mathcal{Z}_{MC} = \oint \frac{D\omega}{\omega^{\hat{T}+1}} \mathcal{Z}_{GC}(\omega) = \frac{1}{\hat{T}!} \lim_{\omega\to 0} \frac{d^{\hat{T}}}{d\omega^{\hat{T}}} \mathcal{Z}_{GC}(\omega) \\ = \begin{cases} \text{ME:} & \lim_{\omega\to 0} \frac{d^{\hat{T}}}{d\omega^{\hat{T}}} (e^{Mz})^L = (ML)^{\hat{T}} \\ \text{W:} & \frac{1}{\hat{T}!} \lim_{\omega\to 0} \frac{d^{\hat{T}}}{d\omega^{\hat{T}}} (1-z)^{-ML} = \binom{ML-1+\hat{T}}{\hat{T}} \\ \text{B:} & \frac{(\hat{T}!)^d}{\hat{T}!} \lim_{\omega\to 0} \frac{d^{\hat{T}}}{d\omega^{\hat{T}}} (1+z)^{ML} = (\hat{T}!)^d \binom{ML}{\hat{T}} \Theta(ML - \hat{T}). \end{cases} \tag{5.24}$$

The above cases can also be obtained by simple combinatorial arguments. Let's now compare the behaviour of the entropy excess ΔS^{Γ} to that of the asymptotical GC entropy, which is the contribution to the integral in the saddle point, $G(\vec{\alpha}^*)$.

[8] Note that for the W case the contour does not include the point $z = 1$ as it does not lie within the radius of convergence of $\mathcal{Z}_{GC}$.

5.2.1.1 *Multi-Edge case:*

In this scenario, we have $z^* = \hat{T}/ML$ to use in (5.16) and so,

$$\begin{aligned} G(\vec{\alpha}^*) &= S^\Gamma[\mathcal{P}_S] \simeq \hat{T}\ln(ML) + \mathcal{O}(\frac{1}{2}\ln\hat{T}) \\ \Delta S^\Gamma &\sim -\mathcal{O}(\frac{1}{2}\ln\hat{T}) \\ \lim_{\hat{T}\to\infty} \frac{\Delta S^\Gamma}{G(\vec{\alpha}^*)} &= 0. \end{aligned} \tag{5.25}$$

And the ensembles are fully equivalent as expected because they fulfill both conditions in (5.10) and (5.13). An additional interesting observation to be made for this case is that if one considers the additional condition that the relative strength sequence must be fixed, for which (for the case of accepting self-loops) $p_{ij}^\infty = \frac{\hat{s}_i^{out}\hat{s}_j^{in}}{\hat{T}^2}$, one obtains:

$$\begin{aligned} \lim_{\hat{T}\to\infty} \frac{S^\Gamma[\mathcal{P}_S]}{\hat{T}} - \ln M &= -\sum_{ij} p_{ij}^\infty \ln p_{ij}^\infty \\ = \ln L - \sum_i p_{s_i^{out}} \ln \frac{p_{s_i^{out}}}{p^{1/2}} &- \sum_j p_{s_j^{in}} \ln \frac{p_{s_j^{in}}}{p^{1/2}}. \end{aligned} \tag{5.26}$$

Where $p = 1/L$. This is nothing more than a decrease in the entropy since the ensembles with fixed strength sequence are clearly a subset of the ensembles with fixed number of events. Such a decrease is introduced in terms of K-L divergence between the "uniform" distribution of events per node $p_s = p^{1/2} = 1/N$ and the constrained distribution in s, $p_{s_i} \equiv \hat{s}_i/\hat{T}$ in both incoming and outgoing direction[9].

5.2.1.2 *Weighted case:*

For this case, we have $z = \frac{\hat{T}/LM}{1+\hat{T}/LM}$. We can then compute the asymptotical entropy,

$$G(\vec{\alpha}^*) = S^\Gamma[\mathcal{P}_S] = -\hat{T}\ln\frac{\frac{\hat{T}}{LM}}{1+\frac{\hat{T}}{LM}} + LM\ln\left(1+\frac{\hat{T}}{LM}\right). \tag{5.27}$$

So by comparison,

$$\begin{aligned} \Delta S^\Gamma &\sim -\mathcal{O}(\ln T) \\ \lim_{\hat{T}\to\infty} \frac{\Delta S^\Gamma}{G(\vec{\alpha}^*)} &= -\frac{1}{ML}. \end{aligned} \tag{5.28}$$

And we clearly see how despite having vanishing entropies, even in this simple case, the two contributions are of the same order. So, in general, the two descriptions will be highly different, even for finite sampling[10].

[9] If we wished, by similar, yet more involved arguments we could try to obtain a formula considering the further case where correlations between $\hat{s}, \hat{s}'$ are considered in the same fashion as in [153].

[10] In this particular case, by the Central Limit Theorem, $\mathcal{P}(T)$ is a Gaussian distribution and thus the previous result can be obtained also by using (5.14). However, as soon

5.2.1.3 *Binary case:*

Finally, for this example $z = \frac{\hat{T}/LM}{1-\hat{T}/LM}$. And analogously to the earlier cases,

$$\begin{aligned} G(\vec{\alpha}^*) = S^{\Gamma}[\mathcal{P}_S] &= -LM \ln\left(\frac{LM-\hat{T}}{LM}\right) + \hat{T}\ln\left(\frac{\hat{T}}{LM-\hat{T}}\right) \\ \Delta S^{\Gamma} &\sim -\mathcal{O}(\frac{1}{2}\ln\left(\frac{LM}{2\pi(LM-\hat{T})\hat{T}}\right)) \\ \lim_{\hat{T}\to LM}\frac{\Delta S^{\Gamma}}{G(\vec{\alpha}^*)} &\sim \mathcal{O}(1). \end{aligned} \tag{5.29}$$

And we see again both contributions to the entropy are of the same order. Note that, in this particular case, the entropies are still well defined (even being 0) because the high sampling limit can be reached, yet this would not be the case as soon as we impose more complicated constraints, since the high sampling limit cannot be reached for the MC ensemble as we have already discussed (in general a fully connected topology will not be compatible with the chosen constraints and hence $\mathcal{Z}_{MC} = 0$).

5.2.2 *Linear and binary constraints: Fixed* $\hat{T}, \hat{E}$

In this case, even if we cannot obtain a closed expression for the MC partition function, many analytical work can be done. Again using

as we enforce increasingly complicated constraints, this will no longer be assured and even this approximation is highly likely to fail.

(5.4), (4.23) and considering the binomial formula and $\Delta(\omega)$ from (5.19),

$$\begin{aligned}
\mathcal{Z}_{MC} &= \oint \frac{D\omega}{\omega^{\hat{T}+1}} \oint \frac{D\lambda}{\lambda^{\hat{E}+1}} \mathcal{Z}_{GC}(\omega,\lambda) \\
&= (\hat{T}!)^d \oint \frac{D\omega}{\omega^{\hat{T}+1}} \sum_{q=0}^{L} \binom{L}{q} (\Delta(\omega))^q \oint \frac{D\lambda}{\lambda^{\hat{E}+1}} \lambda^q \\
&= (\hat{T}!)^d \oint \frac{D\omega}{\omega^{\hat{T}+1}} \sum_{q=0}^{L} \binom{L}{q} (\Delta(\omega))^q \delta_{q,\hat{E}} \\
&= (\hat{T}!)^d \binom{L}{\hat{E}} \sum_{q}^{\hat{E}} \binom{\hat{E}}{q} (-1)^{\hat{E}-q} \oint \frac{D\omega}{\omega^{\hat{T}+1}} (\Delta(\omega)+1)^q \\
&= (\hat{T}!)^d \binom{L}{\hat{E}} \sum_{q}^{\hat{E}} \binom{\hat{E}}{q} (-1)^{\hat{E}-q} \frac{1}{\hat{T}!} \lim_{\omega\to 0} \frac{d^{\hat{T}}}{d\omega^{\hat{T}}} (\Delta(\omega)+1)^q \\
&= \begin{cases}
\text{ME:} & \binom{L}{\hat{E}} (M\hat{E})^{\hat{T}} \left(1 + \sum_q^{\hat{E}-1} (-1)^{\hat{E}-q} \binom{\hat{E}}{q} \left(\frac{q}{\hat{E}}\right)^{\hat{T}}\right) \\
\text{W:} & \binom{L}{\hat{E}} \binom{M\hat{E}-1+\hat{T}}{\hat{T}} \left(1 + \sum_q^{\hat{E}-1} (-1)^{\hat{E}-q} \binom{\hat{E}}{q} \frac{\binom{Mq-1+\hat{T}}{\hat{T}}}{\binom{M\hat{E}-1+\hat{T}}{\hat{T}}}\right) \\
\text{B:} & \binom{L}{\hat{E}} (\hat{T}!)^d \binom{M\hat{E}}{\hat{T}} \left(1 + \sum_q^{\hat{E}-1} (-1)^{\hat{E}-q} \binom{\hat{E}}{q} \frac{\binom{Mq}{\hat{T}}}{\binom{M\hat{E}}{\hat{T}}}\right) .
\end{cases}
\end{aligned} \tag{5.30}$$

The above formulas are highly interesting. On the one hand, we observe the contribution of the binary entropy given by $\mathcal{Z}_{MC,\text{bin}} = \binom{L}{\hat{E}}$ which does not depend on $\hat{T}$, and for which ($\langle\Theta(t)\rangle = \hat{E}/L$):

$$\Delta S^{\Gamma}_{\text{bin}} \sim \mathcal{O}\left(\frac{1}{2}\ln\left(\frac{L}{\hat{E}(L-\hat{E})}\right)\right). \tag{5.31}$$

Besides this contribution, we observe a combinatorial factor similar to the case where we only fixed the total number of events $\hat{T}$ (see (5.24)), but with less available states to be allocated, $\hat{E}$ instead of the total L states. The final term are the corrections taking into account the correlation among the binary and non-binary structure. Let's review each of the cases separately.

5.2.2.1 *Multi-Edge case*

In this case, following the asymptotical solution for the saddle point equations,

$$\begin{aligned}
G(\vec{\alpha}^*) = S^{\Gamma}[\mathcal{P}^{\Gamma}_S] &\simeq \hat{T}\ln(M\hat{E}) + \mathcal{O}(\frac{1}{2}\ln(2\pi\hat{T})) \\
\Delta S^{\Gamma}_{\text{non-bin}} &\simeq -\frac{1}{2}\ln(2\pi\hat{T}) \\
\lim_{\hat{T}\to\infty} \frac{\Delta S}{G(\vec{\alpha}^*)} &= 0.
\end{aligned} \tag{5.32}$$

So, the entropy corrections belonging to the non-binary structure vanish in the high sampling limit, leading to equivalent asymptotic entropies for this ensemble despite the fact that this ensemble clearly violates (5.13) by having non-vanishing asymptotic relative fluctuations. This result amounts to consider that once the $\hat{E}$ binary events have been fixed in each round, we apply a multinomial process of the $\hat{T}$ events on the $\hat{E}$ available states. As we approach the high sampling limit, the combinatorics of the allocation of events on occupied states overwhelmingly contribute to the phase space of the ensemble, compared to the entropy generated by the binary structure. Hence, the ensembles display equivalent entropies per event, despite yielding non-vanishing relative constraint fluctuations. Obviously this is a particular result whose applicability should be reviewed in each case, yet, the scaling of $G(\vec{\alpha}^*) \sim \mathcal{O}(\hat{T})$ (see (5.21)) suggests that this result is likely to hold for a wide variety of cases.

5.2.2.2 *Weighted and Binary cases:*

These cases are not interesting to analyze, since the earlier identification of $\Delta S^{\Gamma}_{\text{non-bin}}$ in (5.30) cannot be performed because one cannot identify the term in front of the sum with the asymptotic entropy of the GC and one must hence compute it by comparison with (5.21). Then, for the case $M = 1$ (where the relation $\langle t^+ \rangle (z)$ can be inverted) a similar thing to the earlier considered linear case happens, which is already bad-behaved.

For the B case and $M = 1$ the non-binary entropy is trivially zero, while for the W case we have again that,

$$\lim_{\hat{T}\to\infty} \frac{S^{\Gamma}[\mathcal{P}_H]}{\hat{T}} = \lim_{\hat{T}\to\infty} \frac{S^{\Gamma}[\mathcal{P}_S]}{\hat{T}} = \lim_{\hat{T}\to\infty} \frac{\Delta S^{\Gamma}}{\hat{T}} = 0 \tag{5.33}$$

$$\lim_{\hat{T}\to\infty} \frac{\Delta S^{\Gamma}}{G(\vec{\alpha}^*)} = -\mathcal{O}(1/\hat{E}) \neq 0 \tag{5.34}$$

and hence the contribution of the maxima does not overwhelm the rest of the integral, and the saddle point argument is essentially not usable in the general case.

5.3 WRAPPING UP: MICRO CANONICAL ENSEMBLE MAIN FEATURES

In this chapter we have reviewed the main aspects concerning the MC ensemble of graphs introduced in Chapter 3. We have represented the MC partition function using an integral representation of the Kronecker deltas that enforce the hard constraints in this ensemble and we have related it to the GC partition function, thus connecting both statistics. Furthermore, we have discussed the conditions under which both ensembles can be considered equivalent. Using a steepest descent approach we have been able to provide explicit entropic expressions for each case considered in this thesis and their asymptotic difference in the high sampling limit. The quality of the steepest descent approximation has been discussed and important limitation for the B and W cases have been highlighted. Two final models considering simple constraints where the micro-canonical entropies can be computed have been reviewed to exemplify the aspects of the micro-canonical formalism discussed in this chapter.

In a nutshell, those are:

A. **Number of constraints:** The MC treatment has allowed us to consider the conditions on the number of constraints Q that can be accepted in relation to $\hat{T}$. It must either be constant or scale sub-linearly with $\hat{T}$.

B. **Relative fluctuations:** The integral treatment has also allowed us to connect the conditions on relative constraint fluctuations to those relating to asymptotic equivalence of entropies between ensembles. We have seen that the former is a stricter condition than the latter, which has been exemplified for the ME case with binary constraints.

C. **Ensemble equivalence for ME linear case:** Furthermore, we have observed how for the ME case with linear constraints, the ensembles are equivalent in the large sampling limit because the relative constraint fluctuations vanish in such a limit. In contrast, for the case where binary constraints are also considered, the ensembles display asymptotically equivalent entropies despite having non-vanishing relative fluctuations. The reason for this is the non-ergodicity of the considered sampling process once the binary structure of the graph has been fixed. This discussion is important as it highlights how requiring the vanishing of constraint relative fluctuations to the scaling variable is a stronger condition for ensemble equivalence than asymptotic equivalence of entropies.

D. **Ill defined high sampling limit for B case:** We have observed that the ill defined nature of the high sampling limit for the

B case, hinted in earlier chapters, is expressed here in zero asymptotical entropies per event. For such a case, the limiting behaviour is a fully connected topology which cannot display any variations. So the GC and MC description will only equivalent for the case where solely the total number of events $\hat{T}$ is fixed and the high sampling limit is reached (because such a limit can be reached with the saddle point equations having graphical solutions). Otherwise, due to the constraints, the limit cannot be reached and the GC will always display non negligible fluctuations compared to the MC one.

E. **Inadequacy of saddle point approximation for the general W case:** For the W case, the scaling of the maximum around where the saddle point approximation is performed is so slow that the contribution outside of its neighborhood cannot be ignored, and hence, even if the relative entropies per event for both ensembles are zero, the entropy corrections cannot be shown to be negligible compared to the main entropic term,

$$\begin{aligned}&\lim_{\hat{T}\to\infty}\frac{S^{\Gamma}[\mathcal{P}_S]}{\hat{T}}=\lim_{\hat{T}\to\infty}\frac{\Delta S^{\Gamma}}{\hat{T}}=0\\&\lim_{\hat{T}\to\infty}\frac{\Delta S^{\Gamma}}{S^{\Gamma}[\mathcal{P}_S]}\neq 0.\end{aligned} \tag{5.35}$$

hence for any sampling, both ensembles will display relevant, non-vanishing, differences.

As conclusion, we emphasize that the only case which does not display any type of problems is the Multi Edge case with linear constraints.

In a second term we could consider the Multi Edge case with both linear and non linear constraints, which, despite yielding non vanishing fluctuations on the constraints, may lead to equivalent entropies per event once the high sampling limit is taken (but this fact needs to be reviewed for each particular case). This is a remarkable result, as it shows that vanishing relative fluctuations is a stricter condition for ensemble equivalence than equality of leading terms in the scaling variable $\hat{T}$ of the entropies. To exemplify this fact, for fixed $\hat{E}$ and $\hat{T}$ and ME case, we have presented and example where the latter is true while the GC ensemble displays non vanishing constraint fluctuations on the scaling variable $\hat{T}$.

Finally, I have shown that the rest of cases display pathologies that cannot be avoided due not only to the ill defined nature of the high sampling limit considered but also the forms of the statistics obtained for the GC ensemble. Even if these cases are still useful for null model construction of networks with prescribed constraints, they present some fundamental theoretical drawbacks that are specially apparent when treated under the present, transparent, MC ensemble. Specially for the W case, we have seen how even considering the simplest case where we only fix the total number of events $\hat{T}$, important differences among ensembles appear, casting doubts about its utility to analyze empirical datasets when more complicated constraints are considered.

6 NETWORK GENERATION AND COMPUTATION OF MACROSCOPIC OBSERVABLES

Simple things should be simple, complex things should be possible.

— Alan Kay

In prior chapters we have developed the framework to build graphs with prescribed constraints. Such a framework, however, is useless unless one is able to obtain explicit expressions for the coefficients $\{\alpha_q\}$ that are solutions of the constraint equations[1] in the GC ensemble or equivalently the saddle-point equations of the MC ensemble. Also, we have still yet to propose practical applications for the analytical models proposed earlier and to find ways in which to generate network instances to test our analytical predictions.

In this chapter I provide all necessary details to solve the saddle point equations for a wide variety of the earlier proposed examples of networks with different linear and binary constraints. I also provide details on how to sample networks[2] once these coefficients are obtained. Then, any network practitioner can average whichever network magnitude she/he wants to test against a null model to see its statistical importance. Another option to avoid simulation is to analytically derive entire predictions for network observables, and in this chapter we also provide all the elements to perform this calculation.

Since in the upcoming chapters the non-binary Multi-Edge configuration model (MECM) will be used extensively as a null-model to assess relevance of features for Origin-Destination matrices, we close this chapter by performing a detailed analysis of this model. Firstly we highlight how considering the same constraints in the different treated cases (ME, B and W) leads to distinct features for the network macroscopic features. Secondly, we focus on the ME case, which is the one we will be using in the second part of this thesis, and the only one for which fully analytical solutions to the saddle point equations can be obtained.

We develop closed expressions for a wide variety of network observable for this model. Furthermore, we computationally test the prediction of equivalence between Multi-Edge ensembles with linear

[1] Saddle point equations and constraint equations are two names relating to the same sets of equations and will be used indistinctively in this thesis.

[2] Generating networks in the MC ensemble can be very tricky depending on the constraints, and we only provide in this chapter the ways to generate all proposed GC models and an algorithm to generate the non-binary Multi-Edge configuration model in the MC ensemble.

constraints in this case, since it allows both a very accurate approximate generation of network configurations using different methods for the MC and the GC ensemble.

6.1 NETWORK GENERATION I: SOLVING THE SADDLE POINT EQUATIONS

The main drawback of the type of formalism presented in this thesis is related with the solving of saddle point equations. One has one equation to solve per each added constraint, and in many situations constraints scale linearly with the number of nodes, hence very large network will present computational issues for the solving of these equations. Additionally, as it already has been pointed out, from the three considered cases (ME, W and B) the W case has to be treated apart since it includes strong conditions $\{0 \leqslant z_{ij} < 1 \, \forall ij\}$ which make the solving of the equations a complicated problem.

As it also has been said, the strategy to use when solving the equations is to consider an optimization problem of maximizing the likelihood associated to the observation of a network with given constraints $\hat{\vec{C}}$ under each model (see Section 4.3). This way, we only have to consider the maximization of a scalar function,

$$\begin{aligned}\mathcal{L}^{\Gamma}(\vec{\alpha}) = \ln \mathcal{P}^{\Gamma}(\hat{\vec{C}}|\vec{\alpha}) &= \sum_{ij} \left(\ln q_{ij}(\hat{t}_{ij}|z_{ij}(\vec{\alpha})) - \ln \mathcal{D}_{ij}(\hat{t}_{ij})\right) \\ &= \sum_{ij} \left(\hat{t}_{ij} \ln z_{ij}(\vec{\alpha}) + \Theta(\hat{t}_{ij}) \ln \tilde{z}_{ij}(\vec{\alpha}) - \ln \mathcal{Z}_{ij}(\vec{\alpha})\right) \\ &= \sum_{ij} \mathcal{L}^{\Gamma}_{ij}(\vec{\alpha}),\end{aligned} \tag{6.1}$$

whose partial derivatives equated to zero correspond to the considered equations to solve.

$$\begin{aligned}\partial_{\alpha_q} \mathcal{L}^{\Gamma}(\vec{\alpha}) &= \sum_{ij} \left(\hat{t}_{ij} \frac{\partial_{\alpha_q} z_{ij}}{z_{ij}} + \Theta(\hat{t}_{ij}) \frac{\partial_{\alpha_q} \tilde{z}_{ij}}{\tilde{z}_{ij}} - \partial_{\alpha_q} \ln \mathcal{Z}_{ij}\right) \\ &= \sum_{ij} a_q^{ij}(\hat{t}_{ij} - \langle t_{ij} \rangle) + \tilde{a}_q^{ij}(\Theta(\hat{t}_{ij}) - \langle \Theta(t_{ij}) \rangle) = \hat{C}_q - \langle C \rangle_q (\vec{\alpha}) \\ \partial_{\alpha_q} \mathcal{L}(\vec{\alpha})\big|_{\vec{\alpha}^*} = 0 &\implies \hat{\vec{C}} = \left\langle \vec{C} \right\rangle (\vec{\alpha}^*)\end{aligned} \tag{6.2}$$

The optimization problem to be solved is shown concave for all considered problems in this thesis. Considering the function in (6.2) and computing the Hessian of $\mathcal{L}(\vec{\alpha})$, one has (assuming $\tilde{a}_q^{ij}, a_q^{ij} \geqslant 0 \, \forall ij, q$),

$$\begin{aligned}&\partial^2_{\alpha_q \alpha'_q} \mathcal{L}(\vec{\alpha}) = -\sigma^2_{C_q C'_q} = \\ &= -\sum_{ij} \left(a_q^{ij} a_{q'}^{ij} \sigma^2_{t_{ij}} + \tilde{a}_q^{ij} \tilde{a}_{q'}^{ij} \sigma^2_{\Theta(t_{ij})} + 2 a_q^{ij} \tilde{a}_q^{ij} \langle t_{ij} \rangle (1 - \langle \Theta(t_{ij}) \rangle)\right) \leqslant 0 \forall \vec{\alpha}\end{aligned}$$

(6.3)

And hence the problem is concave (so any critical points fulfilling (6.2) will be maxima). For the maximum to exist, we need in the first place that the considered constraints correspond to a graphical solution (there must exist at least one network that fulfills the considered conditions). This is usually not a problem, since normally we want to match a set of constraints to the values obtained from some real network, which must be necessarily graphical.

For the ME and B cases (both with linear and binary constraints), the bounded functions $-\infty < \mathcal{L}_{ij}(\vec{\alpha}) \leqslant 0$ to be optimized are continuous for all the domain $\vec{\alpha} \in \mathbb{R}^Q$, which is a convex open set, and at the boundaries have equal limiting values $\lim_{\vec{\alpha} \to \pm\vec{\infty}} \mathcal{L}_{ij}(\vec{\alpha}) = -\infty$. Given the concavity of the problem, we see thus that if the constraints are graphical, a maximum value must exist and be unique.

It must be noted, however, that for the W case this concavity is lost as soon as the enforcement $\{0 \leqslant z_{ij} < 1 \, \forall ij\}$ is imposed on the domain of the likelihood function. Such a domain is a non-convex set and in this case, proving the unicity of the maximum is not simple. Assuming that the domain $\mathcal{D} = \{\vec{\alpha} \in \mathbb{R}^Q | 0 \leqslant \prod_q^Q e^{\alpha_q a_q^{ij}} < 1\}$ is path connected, we may assume using an informal argument that if the constraints are graphical, again the maximum will exist and be unique for finite sampling[3]. In any case, even if the maximum exists and is unique, finding it is not a simple task, as no general algorithms exist for this kind of problem, so each problem must be examined on a one-to-one basis.

We deal hence with a large-scale, bounded (in all cases $\{z_{ij} \geqslant 0\}$) maximization problem. In the following subsections we review how to solve the different proposed cases separating between ME and B cases and the W case. The first two can be tackled using balancing algorithms, which allow for fast a robust solving of the equations whereas the latter case needs to be faced with a brute-force maximization approach with no guarantee of convergence.

Balancing algorithms are very close to what is known as Iterative Fitting Procedures (IPFP) in economics literature. and their mathematical properties have been extensively studied (see [141] and references therein).

[3] The argument proceeds as follows. Given that the function is concave everywhere where it is defined (and infinitely differentiable), which is in the interior of the set given by $\mathcal{D}$, either it has no critical points, has one, or more than one. If it has no critical points, the function has to be monotonic, but this cannot be since the limiting bounds are equal in all directions (this is the same argument as for the earlier ME and B cases). If it has more than one critical points, and they cannot be in the boundaries, then they must be maxima. However, since the domain is path connected, we can always join those points along a continuous and differentiable path, and necessarily along this path some minima should be observed, but this is prohibited by the concavity of the function. Hence, the maximum needs to be unique on the path connected domain.

EXPLICIT CONSTRAINT	ME	W	B
Linear			
$\hat{T}$	A	A	A
$\hat{T}, \hat{C}$	1D	2D	2D
$\hat{\vec{s}}$	A	F	B
$\hat{\vec{s}}, \overline{\hat{t}_{ss'}}$	A	A	A
$\hat{\vec{s}}, \hat{C}$	$B-2S$	F	$B-2S$
$\hat{T}, \overline{\hat{t}_{uu}}$	A	A	A
$\hat{\vec{s}}, \overline{\hat{t}_{uu}}$	B	F	$B-2S$
Binary			
$\hat{T}, \hat{E}$	A	A	A
$\hat{T}, \hat{\vec{k}}$	$A+B$	$1D+B$	$1D+B$
Linear and Binary			
$\hat{\vec{s}}, \hat{E}$	$B-2S$	F	$B-2S$
$\hat{\vec{s}}, \hat{\vec{k}}$	B	F	B

Table 6.1: **Solving the saddle point equations.** Strategies for solving the saddle point equations of each explicit example and case. A stands for analytical calculation, B for balancing Algorithm 1, $B-2S$ for a two step balancing Algorithm 2, F for brute force maximization and 1D(2D) for uni and bi-dimensional maximization respectively. All explicit details to solve each case can be found in Section B.1.

Table 6.1 depicts all explicit examples presented in this thesis. The exact methodology and procedure to solve each case are detailed in Section B.1.

In any of the proposed algorithms, some numerical precision problems can arise due to the presence of exponentials, or large powers, but I leave to the reader the details on how to avoid them. These problems often arise when dealing with the combined linear and binary constraints, since from equations (4.25) and (4.26) we have the presence of terms $e^{Mz_{ij}}$, $(1-z_{ij})^{-M}$ and $(1+z_{ij})^{M}$ which can complicate the numerical handling.

Here we explain the main ideas behind the balancing algorithm approach to solve the saddle point equations, which are widely applicable to the ME and B cases.

Algorithm 1: Balancing algorithm to solve saddle point equations for ME and B cases with node-defined constraints.

```
Input: Node-constraints $\hat{\vec{C}}$ (float vector length 2N), maximum
       tolerance on node-constraints $\varepsilon_N$ (float), tolerance on
       node-variables $\varepsilon_\alpha$ (float) and initial guess $\vec{\alpha}^{ini}$ (float
       vector length 2N).
Output: List of node-Lagrange multipliers $\{\alpha_q \, \forall q = 1, 2N\}$
begin Initialization
    Set $\vec{\alpha}^{(0)} = \vec{\alpha}^{ini}$. Set $n = 0$.;
end
begin Node-balancing
    while $\max |\hat{C}_q - \langle C_q \rangle (\vec{\alpha}^{(n)})| > \varepsilon_N$ or
    $\max |\alpha_q^{(n+1)} - \alpha_q^{(n)}| > \varepsilon_\alpha$ do
            $\vec{\alpha}^{(n+1)} = \vec{\alpha}(\vec{\alpha}^{(n)}, \hat{\vec{C}})$;
            $n = n + 1$
    end
    Set $\vec{\alpha} = \vec{\alpha}^{n+1}$;
end
return $\vec{\alpha}$
```

6.1.1 *ME and B cases: Balancing algorithms*

The most basic form of a balancing algorithm is explained in Algorithm 1. The values $z_{ij} \equiv \prod_q e^{\alpha_q(a_q^{ij})}$ can take any positive value for the ME and B cases, and this means that such a balancing algorithm will never fall outside of the domains of the function $\mathcal{L}^\Gamma(\vec{\alpha})$ and no additional conditions need to be imposed. A variant of this algorithm (see Algorithm 2) can be applied when the problem is separable: Any problem involving constraints at the node level and a single general constraint involving all the network (such as cost $\hat{C}$, total events $\hat{T}$ or total binary events $\hat{E}$) will be of this type. In this case, one separates the problem in two parts: One explores the phase space of the Lagrange multiplier associated to the global constraints while balancing (enforcing) the node-local constraints at each step. In doing so, we reach effectively an almost $1-D$ maximization problem which is concave, hence by simple exploration (gradient descent) the problem can be solved.

The Algorithms presented schematically in this section can obviously been perfected but the core idea remains unchanged.

6.1.2 *W case: Brute force optimization*

As commented before, the W cases are complicated to solve and need to be examined on a one to one basis. No general algorithm exists for optimization problems for non-convex problems, and hence it is difficult to give general guidelines to solve these problems. During this thesis, the only (approximated) implementation to find solutions

Algorithm 2: Balancing algorithm to solve saddle point equations for ME and B cases with node-defined constraints and an additional graph-general constraint.

```
Input: Node-constraints $\hat{\vec{C}}$ (float vector length 2N), General
       constraint $\hat{C}_{gen}$ (float), maximum tolerance on
       node-constraints $\varepsilon_N$ (float), maximum tolerance on
       general constraints $\varepsilon_{gen}$ (float), tolerance on
       node-variables $\varepsilon_\alpha$ (float), tolerance on general variable
       $\varepsilon_\theta$ and node-variables initial guess $\vec{\alpha}^{ini}$ (float vector
       length 2N) and general variable initial guess $\theta_{ini}$.
       Exploration step $\Delta\theta$.
Output: List of node-Lagrange multipliers $\{\alpha_q \; \forall q = 1, 2N\}$,
        general lagrange multiplier $\theta$.
begin Initialization
    Set $\vec{\alpha} = \vec{\alpha}^{ini}$, $\theta^{(0)} = \theta_{ini}$. Set q=1.;
end
begin Loglikelihood maximization
    while $|\hat{C}_{gen} - \langle C_{gen} \rangle (\vec{\alpha}, \theta)| > \varepsilon_{gen}$ or $|\theta - \theta^{(new)}| > \varepsilon_\theta$ do
        begin General constraint search
            Set $\theta^{(new)} = \theta + q\Delta\theta$.;
            if $\mathcal{L}^\Gamma(\hat{\vec{C}}|\vec{\alpha}, \theta^{(new)}) \geqslant \mathcal{L}^\Gamma(\hat{\vec{C}}|\vec{\alpha}, \theta)$ then
                $\theta = \theta^{(new)}$
            else
                $q = -q$
            end
        end
        begin Node-balancing
            Set $n = 0$. Set $\vec{\alpha}^{(0)} = \vec{\alpha}^{ini}$;
            while $\max |\hat{C}_q - \langle C_q \rangle (\vec{\alpha}^{(n)})| > \varepsilon_N$ or
            $\max |\alpha_q^{(n+1)} - \alpha_q^{(n)}| > \varepsilon_\alpha$ do
                $\vec{\alpha}^{(n+1)} = \vec{\alpha}(\vec{\alpha}^{(n)}, \hat{\vec{C}}, \hat{C}_{gen}, \theta)$;
                $n = n + 1$
            end
            Set $\vec{\alpha} = \vec{\alpha}^{n+1}$. Set $\vec{\alpha}^{ini} = \vec{\alpha}^{n+1}$;
        end
    end
    return $\vec{\alpha}, \theta$
end
```

to the problem has been concerning the non-binary configuration model (fixed strength sequence $\hat{s}$) using a mixed approach including pre-conditioning with a gradient descent method and a precision search using interior point methods. This case serves as example for the problem of the non-convexity of the domain and is discussed at length in Section B.2.

6.2 NETWORK GENERATION II: SAMPLING NETWORKS

Once the saddle point equations have been solved, one can proceed to generate networks and average network observables to obtain benchmark values for whichever model considered. Alternatively, it is also possible to perform analytical approximations to network magnitudes as we shall see in Section 6.3.

In the following we provide the technical details related to generating networks in the different ensembles.

6.2.1 *Canonical ensemble*

For this ensemble it has only been possible to perform calculations in the Multi-Edge case with linear constraints. In such a case, the method of generation is simple: For each network configuration, one must generate a set of multinomial distributed variables with probabilities $\{p_{ij}^{\infty} = z_{ij}/\sum_{ij} z_{ij}\}$. This can be achieved individually with a simple rejection method. The only drawback of this method is that the generation cannot be made independent for each edge, and this limits the amount of memory one can use to generate large networks. For this reason and due to the fact that different occupation numbers are correlated, this ensemble is not very useful to work on.

6.2.2 *Grand Canonical ensemble*

The great advantage of this ensemble is that one can decompose each network generation in a set of L independent processes. This allows for paralelization if necessary on the case of large networks. Also, it further allows a quick an simple way to generate non-sparse networks for which the number of events $\hat{T} \gg L$. For each case (ME, W, B) it suffices to generate networks according to each obtained probability distribution (Poisson, Negative Binomial, Binomial) for each state (or pair of nodes) independently, which can be done applying any standard generation method for the case where purely linear constraints are considered.

A comment is in order however for the case where binary constraints are added. In such a case, the generation need to be performed in a two step-method (yet still independently for each state) as described in Algorithm 3.

Algorithm 3: Generating algorithm for Zero-Inflated processes to be used to generate network realizations for the case where binary constraints are considered. Note that $\langle t_{ij} \rangle = z_{ij}$ for Poisson, $\langle t_{ij} \rangle = M z_{ij}/(1 - z_{ij})$ for Negative Binomial and $\langle t_{ij} \rangle = M z_{ij}/(1 + z_{ij})$ for Binomial respectively.

```
Input: State occupation probability $\langle \Theta(t_{ij}) \rangle$ and Lagrange
       multipliers $z_{ij}$. Total trials $R$.
Output: State $ij$ occupation number $t_{ij} \in \mathbb{N}$.
begin Binary occupation
    Generate random number $r$;
    if $r < \langle \Theta(t_{ij}) \rangle$ then
        begin State occupation
            Set $rr = 0$. $t_{ij} = 0.$;
            while $rr < R$ do
                Generate $t_{ij}$ with mean $\langle t_{ij} \rangle (z_{ij})$ using the
                normal version of the statistics in each case;
                if $t_{ij} > 0$ then
                    Break
                end
            end
        end
    else
        $t_{ij} = 0$
    end
end
return $t_{ij}$
```

The two priorly mentioned ensembles just have a drawback concerning its generation: The quality of the random generator and method used. One must ensure its quality in order to obtain truly unbiased network instances for each ensemble.

6.2.3 *Micro Canonical ensemble*

Generating network realizations on this ensemble can be involved as enforcing the constraints for each network realization is a complicated process and need to be carried out by *rewiring* methods. Such methods need to be designed ad-hoc for each type of constraints and need to be expressed in terms of Markov processes that select events from the network and swaps them, conserving the imposed constraints in each step. Many rewiring algorithms exist, yet not all of them perform an unbiased sampling of the phase space: Not all network states are equiprobable and great care need to be taken when using these kind of algorithms. Furthermore, if one wants to build networks from scratch with prescribed properties (without starting from a network instance to be rewired), no general unbiased generating methods exist.

For the case of binary networks, an exact generation can be achieved in some cases, see [56, 20]. The case of non-binary networks, however, becomes more complicated because of the degeneration of the Ω space. For the case of ME networks this is not a problem, since for sufficiently large sampling ensemble equivalence is assured and hence one can generate networks using either of the two other proposed ensembles. However, this is indeed a nuisance for the B and W cases: Depending on the problem at hand, the non equivalence between ensembles can lead to substantial differences depending on the generating method chosen for the models.

In Section B.3 a more detailed discussion on the only MC ensemble explored numerically in this work is performed.

Apart from the problems in designing the generating algorithm for these ensembles, these algorithms can be slow for non-sparse networks. Being based on rewiring methods, $\mathcal{O}(T)$ accepted moves are required for a single realization, whence this methods can be lengthy for dense networks. Also, in many cases it will be convenient to assume that the data gathered contains some kinds of errors, which can be *mimicked* by using a probabilist framework rather a strictly constrained one such as the MC ensemble. Throughout this thesis, the only MC ensemble we have generated is the non-binary Multi-Edge Configuration model (MECM).

6.3 COMPUTING ANALYTICAL PREDICTIONS FOR NETWORK OBSERVABLES

An alternative to generating network instances and averaging is to compute analytically expected quantities for network observables. This can be done once the constraint equations have been solved and the Lagrange multipliers $\vec{\alpha}$ have been obtained. Sometimes exact calculations can be carried out and in the rest of cases linear approximations can be used.

Many network metrics widely used in the literature can be written as a quotient of functions of the occupation numbers $\mathcal{M} = x(t_{ij})/y(t_{ij})$. $\{t_{ij}\}$ are random variables and computing $\langle\mathcal{M}\rangle$ might not be straightforward. We thus need to rely on approximations, expanding the expressions in Taylor series around their mean values and then taking the ensemble average of the first terms of the sum.

All analytical approximations are to be carried in the GC ensemble due to the independent nature of the occupation numbers random variables involved.

$$\langle\mathcal{M}\rangle \simeq \frac{\langle x\rangle}{\langle y\rangle}\left(1+\frac{\langle y^2\rangle}{\langle y\rangle^2}-\frac{\langle xy\rangle}{\langle x\rangle\langle y\rangle}\right) = \frac{\langle x\rangle}{\langle y\rangle}\left(1+\frac{\sigma_y^2}{\langle y\rangle^2}-\frac{\sigma_{xy}^2}{\langle x\rangle\,\langle y\rangle}\right)$$

$$\sigma_{\mathcal{M}}^2 \simeq \frac{\langle x\rangle^2}{\langle y\rangle^2}\left(\frac{\langle x^2\rangle}{\langle x\rangle^2}+\frac{\langle y^2\rangle}{\langle y\rangle^2}-2\frac{\langle xy\rangle}{\langle x\rangle\langle y\rangle}\right) = \frac{\langle x\rangle^2}{\langle y\rangle^2}\left(\frac{\sigma_x^2}{\langle x\rangle^2}+\frac{\sigma_y^2}{\langle y\rangle^2}-2\frac{\sigma_{xy}^2}{\langle x\rangle\,\langle y\rangle}\right) \tag{6.4}$$

These expressions can be used to compute expected values and fluctuations of any network metric expressed as a ratio of functions of the occupation numbers $x(t_{ij}), y(t_{ij})$ provided that the moments $(\langle x \rangle, \langle x^2 \rangle, \langle y \rangle, \langle y^2 \rangle)$ in the right-hand side can be evaluated. This is usually the case when $x(t_{ij}), y(t_{ij})$ are algebraic expressions of $\{t_{ij}\}$ (which are uncorrelated random variables in the GC ensemble). For most metrics $\mathcal{M}$ widely used for non-binary networks, the calculations of the moments of $x(t_{ij})$ and $y(t_{ij})$ are lengthy, but follow from a general methodology without further difficulty[4]. In Section 6.5 we will provide explicit examples of this methodology applied to the MECM case.

6.4 CHOOSING THE RIGHT NULL MODEL FOR THE RIGHT PROBLEM: THE NON-BINARY CONFIGURATION MODEL CASE

To highlight the importance of considering an appropriate null model for the assessment of real data features, we consider the case of networks with a fixed strength sequence applied to a real dataset. Empirically observed networks usually display highly skewed node strength distributions, which have important effects in their observables. Hence, to correctly assess whether some observed feature in a dataset can be solely explained by the strength distribution, it is crucial to choose an appropriate null model to compare the data to. This situation is especially important for instance with regard to community analysis through modularity maximization for weighted networks, because the modularity function to be optimized [35] needs as input a prediction from a null model with fixed strengths ($Q \propto \sum_{i,j} \left(\hat{t}_{ij} - \langle t_{ij} \rangle\right) \delta_{u_i,u_j}$ where $\{u_i\}$ are the community node labels associated to the optimal network partition). The resulting saddle point equations for this model are

$$\hat{s}_i^{out} = \left\langle s^{out} \right\rangle \qquad \hat{s}_i^{in} = \left\langle s^{in} \right\rangle \qquad i = 1, N \tag{6.5}$$

which particularized to each case read,

$$\begin{array}{ll} \text{ME:} & \begin{cases} \hat{s}_i^{out} = M x_i \sum_j y_j \\ \hat{s}_j^{in} = M y_j \sum_i x_i \end{cases} \\ \text{W:} & \begin{cases} \hat{s}_i^{out} = M x_i \sum_j \frac{y_j}{1 - x_i y_j} \\ \hat{s}_j^{in} = M y_j \sum_i \frac{x_i}{1 - x_i y_j} \end{cases} \\ \text{B:} & \begin{cases} \hat{s}_i^{out} = M x_i \sum_j \frac{y_j}{1 + x_i y_j} \\ \hat{s}_j^{in} = M y_j \sum_i \frac{x_i}{1 + x_i y_j} \end{cases} \end{array} . \tag{6.6}$$

The ME case has an analytical solution while the others must be solved computationally using the methods presented in earlier sections. For details in the degree of precision attained, refer to Section B.2.

[4] And can be easily implemented using any standard mathematical symbolic software.

As real world dataset we use the OD matrix generated by taxi trips in Manhattan for the year 2011 [149, 4]. This dataset will be extensively used in Part iii and details about it can be found in Appendix C. The OD has been constructed as the aggregation of $M = 365$ daily layer snapshots, each node represents an intersection and each weight the number of trips recorded between them. Even if the network considered is directed throughout this section we will only show results in the outgoing direction, as the results in the incoming direction are qualitatively equal.

To analyze the difference between models, we compute ensemble expectations for different edge and node related properties suitably rescaled (fixing the original strength distribution of each dataset) and then compare the obtained results with the real observed data features.

The details and precision of the saddle point equation solving can be found in Section B.2.

Since the taxi dataset is quite dense ($\hat{E}/L = 0.43$), it has enough sampling for the wide differences between models to emerge. All cases have the same number of events $\hat{T}$ on average, but they are not distributed among connections between nodes in the same way for the different models. Being zero the most probable value for the geometric distribution, for the W case with a single layer the connection probability initially grows distinctively faster than in all the other cases leading to larger number of binary connections between low strength nodes (Figure 6.1-A,B). Yet the higher relative fluctuations of the geometric statistics also generate extremely large maximum weights in the tail of the existing occupation number distribution (Figure 6.2), which are concentrated in connections between high strength nodes (Figure 6.1). Since the total number of events incoming and outgoing each node is fixed, this means that the W case has comparatively the lowest degrees for the most weighted nodes despite counting the larger number of binary connections $E = \sum_{ij} \Theta(t_{ij}) = \sum_i k_i^{out} = \sum_j k_i^{in}$ as can be seen in Figure 6.3-A.

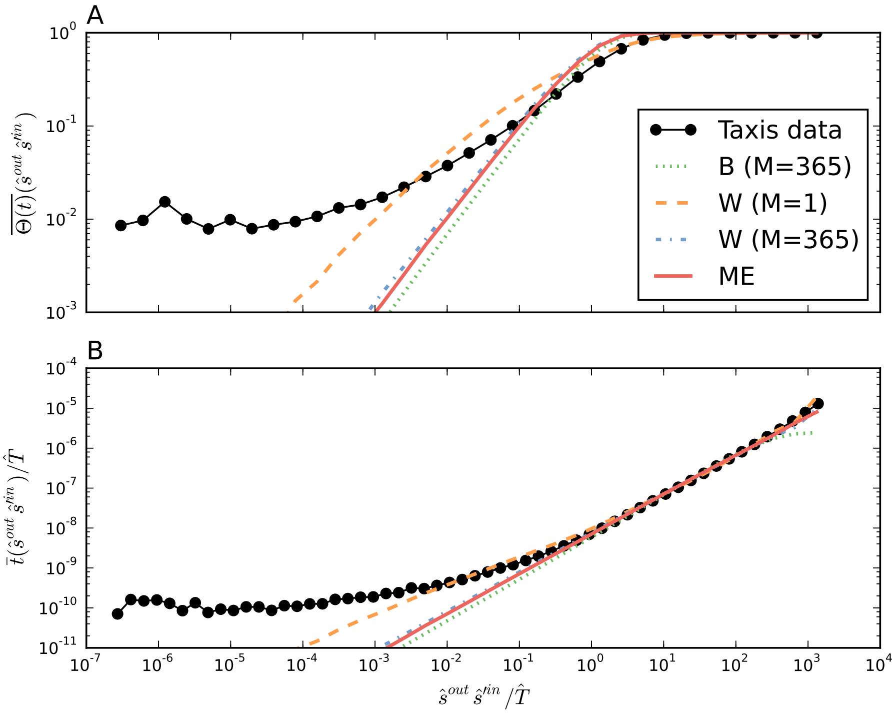

Figure 6.1: **Node pair statistics.** Binary connection probability (A) and rescaled average occupation number (B) as function of product of origin and destination node strength. Results averaged over $r = 5 \cdot 10^2$ and $r = 10^4$ realizations for the different models respectively with applied log-binning. The sudden increase for the binary pair-node connection probability can be clearly seen for the W case.

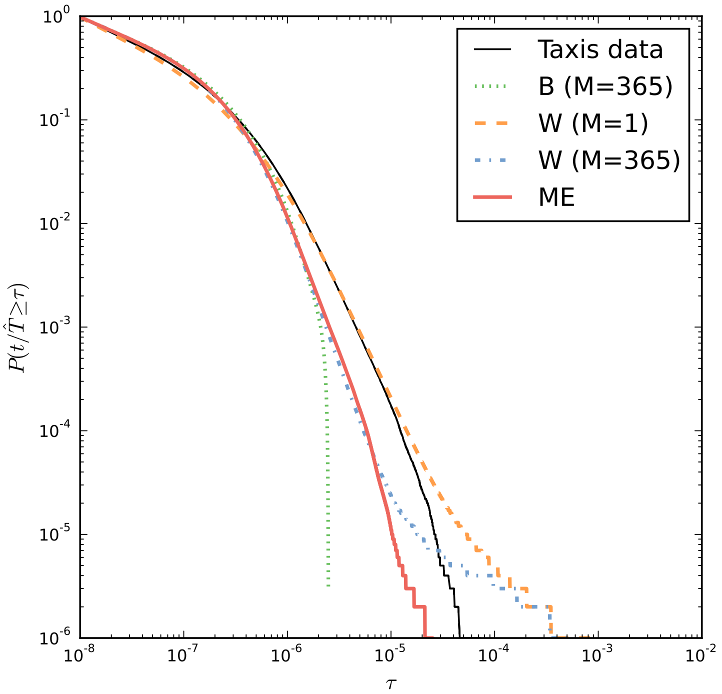

Figure 6.2: **Existing occupation number statistics.** Existing occupation number complementary cumulative distribution for the taxi dataset. Same conditions as Figure 6.1 apply. The presence of extremely large weights can be seen in the tails of the distributions for both the W monolayer and multilayer case.

These anomalies for low and high strength nodes respectively for the W case produce wild asymmetries in the allocation of weights per node, which can be studied measuring their disparity $Y_2 = \sum_j t_{ij}^2 / \left(\sum_j t_{ij}\right)^2$ (Figure 6.3-B), which quantifies how homogeneously distributed are the weights emerging from each node: It displays a U shaped form with both low and high strength nodes tending to very strongly concentrate their weights on few connections. This non-monotonic behaviour is in strong contrast with the one observed for the taxi data and usually in other datasets [124]. Concerning second order node correlations, the outgoing weighted average neighbor strength $s_{nn}^w = \sum_j t_{ij} s_j^{in} / s_i^{out}$ (Figure 6.4) again displays a large range of variation for the W case (with either one or more than one layer) in contrast with the slight assortative profile of the real data, the uncorrelated profile of the ME case and the slight disassortative trend of the B case[5]. This last case is caused by the combination of two factors: The limitation on the maximum weight of the edges cannot compensate (with large weights connecting the nodes with the larger strength) the tendency of large nodes to be connected

[5] In the context of network science, assortativity is associated with an increasing trend of s_{nn}^w with increasing node strength s while dissasortativity refers to a decreasing trend.

to a macroscopic fraction of the network, which is dominated by low strength nodes.

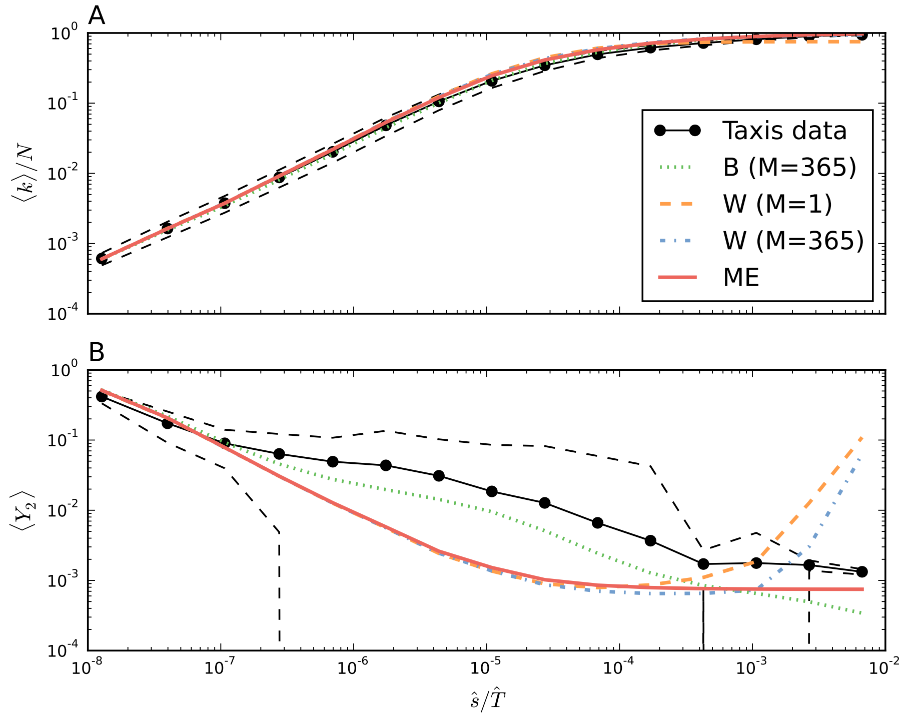

Figure 6.3: **First order node statistics.** Rescaled degree (A) and disparity (B). Same conditions as Figure 6.1 apply. Dashed lines represent log-binned standard deviation ranges for the real data. The U-shaped disparity profile is clearly seen for the W cases in sharp contrast with the monotonous behaviour of both the real data and the ME model.

Obviously none of the null models used reproduce the real data, however, the goal in model construction is rather to assess the structural impact that a given constraint (in this case a strength distribution) has on the network observables. In this sense, we have shown that different models provide very different insights about such impacts. In particular, since the taxi dataset a Multi-Edge network (people riding taxis are clearly distinguishable), the fact that the B and ME cases respectively lie closer to the real data comes at no surprise.

Once the model to use is clear, and given that the results for this particular ME case are completely analytical, we will focus on it to perform an extended study, since it will be the main model to be used in Part iii.

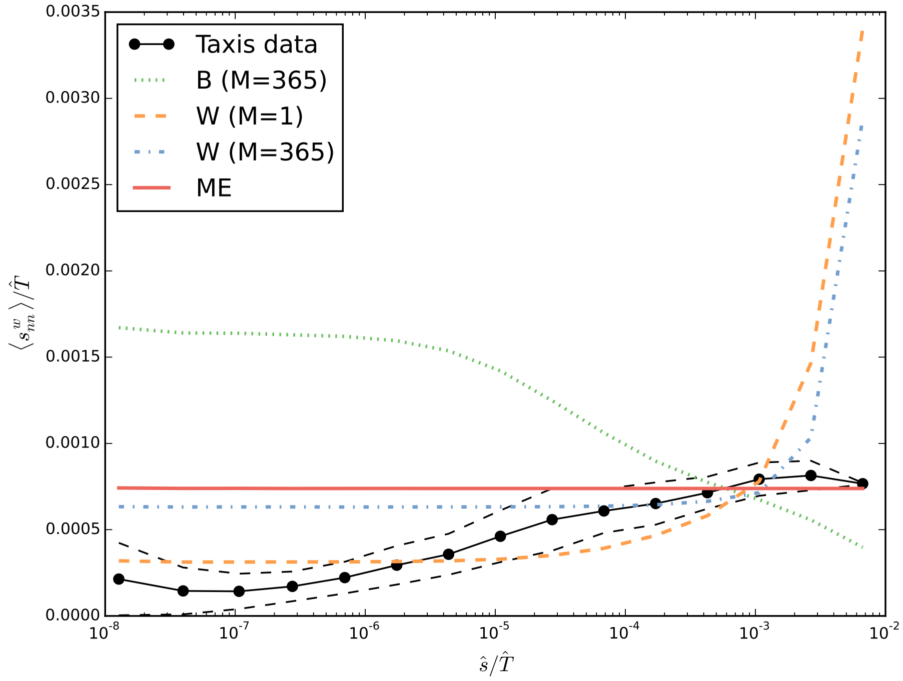

Figure 6.4: **Second order node statistics.** Rescaled weighted average strength. Same conditions as Figure 6.1 apply. Dashed lines represent log-binned standard deviation ranges for the real data. A sharp increase is clearly seen for high strength nodes in the W cases.

6.5 a case study: the me configuration model

As we have seen in earlier chapters, Multi-Edge networks with linear constraints are the only ones whose ensembles possess well defined characteristics. For this kind of networks one can compute exact statistics for the three ensembles considered (Grand Canonical, Canonical and Micro-canonical) which lead to equivalent results. One can also compute well defined event specific entropies in the Γ space.

Among all possible choices of linear constraints, the case where the strength sequence is fixed allows for analytical solving of the constraint equations. Additionally, arbitrary strength distributions can be considered, including highly skewed ones, which are commonly present in real data. This fact turns it in the ideal candidate to act as null model for skewed datasets, since it balances nicely heterogeneity and simplicity in generation and analytical treatment. Furthermore, it is one of the few cases where simple rewiring/generating algorithms can be used to generate unbiased network instances in the MC ensemble (see Section B.3 for extended discussion). Hence, it is the perfect model to use as case study to exemplify all the theoretical developments early developed.

In this final concluding section, we[6] compute analytical predictions for the full edge and node statistics as well as first order correlations providing not only average expected values for the observables but also precise bounds for its fluctuations and compare the obtained results with simulations using a stub-rewiring algorithm (see Section B.3 for details on the algorithm, also used in [159], which is based on the well known binary configuration model [30, 42, 193]), yielding excellent agreement. By particularizing the general results to the case of power law distributed strengths, commonly found in real data [53, 100**?** , 63, 25], we demonstrate how the null-model expectations of some widely used non binary network metrics, which are generally considered a sign of relevant correlations (see [140] and references therein), can instead in some cases be seen as just a consequence of the particular form of the imposed strength sequence, and hence may not represent any unexpected property of the network under study. For the sake of simplicity, I will consider equal incoming and outgoing strength sequences, so the incoming and outgoing Lagrange multipliers will be equal, (out) and (in) superscripts will be dropped and only results in the outgoing direction will be shown.

For this model, the solution to the saddle point equations reads

$$\langle t_{ij} \rangle = \frac{\hat{s}_i \hat{s}_j}{\hat{T}}, \tag{6.7}$$

the left-hand side is the ensemble average of a random variable, while the right-hand side is a result expressed in term of the constraints. In this case the $\{\hat{s}_i\}$ are the only fixed quantities and hence they must be taken as the basic variables from which to derive the rest of the network properties: All nodes sharing the same strength value $\hat{s}_i = \hat{s}$ are statistically equivalent, and possess self-averaging properties (likewise all edges connecting nodes with the same pair of strength values). In what follows, we apply the procedure proposed in Section 6.3 to obtain some particular network metrics. More details explaining the general strategy used in their calculation are given in Section A.3, stating there only the key results for all metrics considered and leaving to the reader the adaptation to other network magnitudes.

6.5.1 *Distribution of existing occupation numbers*

We start by computing the distribution of occupation numbers

$$P(t) = \frac{1}{E} \sum_{ij} \delta(t, t_{ij}) \Theta(t_{ij}) \tag{6.8}$$

which has been reported to have broad forms on empirical data for airport flow [53], cargo ship transport [100], public transport in cities [**?**], commuting [63] or face to face interactions [25] among others.

[6] This part of the work was done in close collaboration with Dr. Francesc Font.

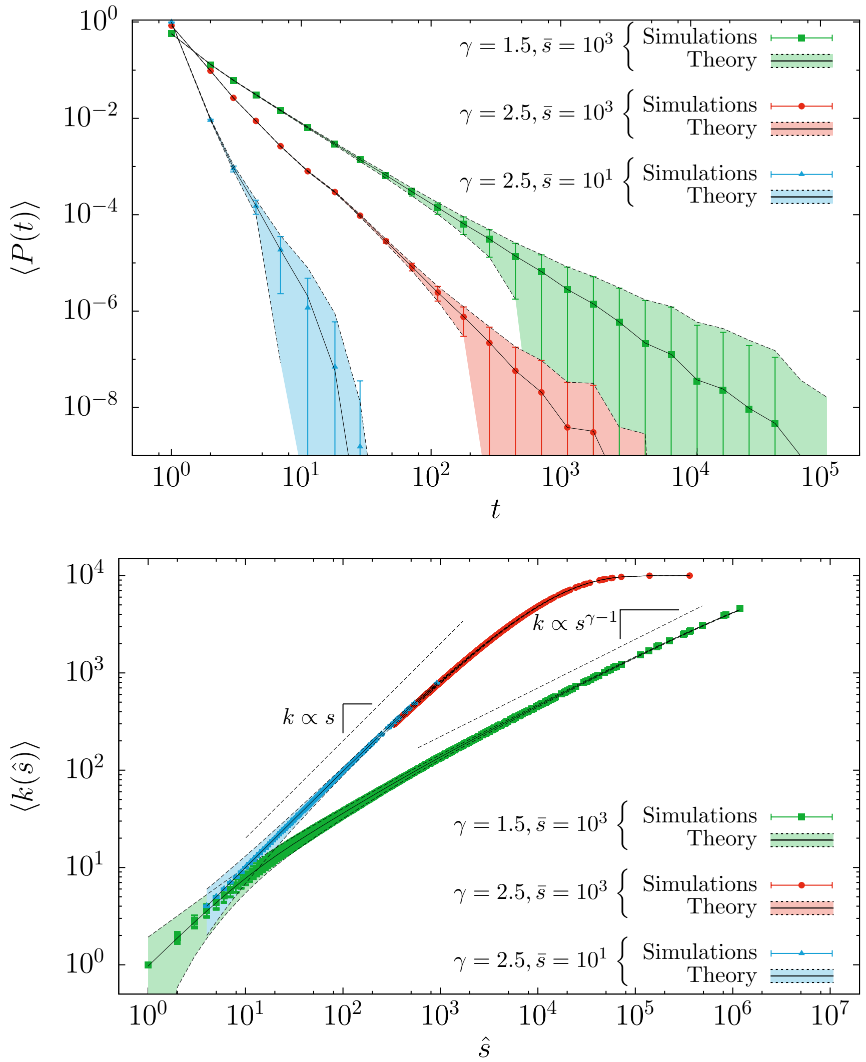

Figure 6.5: **The effect of the strength sequence on network observables.** (Upper): Ensemble average of distribution of occupation numbers over existing edges (log-binned) and analytical predictions given by expression (6.9) and its standard deviation (see Section A.3, Eq. (A.27)) for power law distributed strength sequences with $N = 10^4$ and different exponents for 1000 repetitions. The dependence on sampling $\bar{s} = \hat{T}/N$ is apparent. (Lower): Degree-strength relationship for the same networks as earlier (average) and theoretical predictions from equation (6.10). Standard deviation are represented as error bars and lines of constant slope are provided as a guide to the eye.

Applying Eq. (6.4) to the case of P(t) yields,

$$\langle P(t)\rangle = \left\langle \frac{\sum_{ij}\delta(t,t_{ij})\Theta(t_{ij})}{\sum_{ij}\Theta(t_{ij})}\right\rangle \simeq \frac{\sum_{ij} e^{-\langle t_{ij}\rangle}\langle t_{ij}\rangle^{t}}{t!\langle E\rangle} + \mathcal{O}(\langle E\rangle^{-2}). \tag{6.9}$$

Figure 6.5 shows the distribution of occupation numbers for existing edges and its associated standard deviation (see Eq. (A.27)) for three networks generated using power law distributed strength sequences ($\gamma = 1.5, 2.5$) and different graph-average strength $\bar{s} = \hat{T}/N$. We can see that the form of the resulting distribution is broad due to the imposed form of the strength sequence, and hence is not a sign *per se* of any interesting property of the multi-edge network being studied. Moreover, its shape is not stable and strongly depends on the total number of observed events $\hat{T}$.

6.5.2 *Degrees and strengths*

Having tackled the occupation number statistics of the network, in what follows we consider its node-related properties. We have that the strengths $s_i = \sum_j t_{ij}$ will also be Poisson distributed random variables, being sums of independent occupation numbers. Moreover, since the binary projection of occupation numbers $\Theta(t_{ij})$ are Bernoulli distributed variables with parameter $q(t_{ij} > 0) = 1 - e^{-\langle t_{ij}\rangle}$ one can also compute the associated degrees k_i of the nodes, which will be sums of independent Bernoulli random variables. We have that,

$$\begin{aligned}
\langle k(\hat{s}_i)\rangle &= \left\langle \sum_j \Theta(t_{ij})\right\rangle = \sum_j q_{ij}(t_{ij} > 0) = \\
&= \sum_j \left(1 - e^{-\langle t_{ij}\rangle}\right) = N - \sum_j e^{-\frac{\hat{s}_i\hat{s}_j}{\hat{T}}} \\
\sigma^2_{k(\hat{s}_i)} &= \sum_j \sigma^2_{\Theta(t_{ij})} = \sum_j e^{-\langle t_{ij}\rangle}\left(1 - e^{-\langle t_{ij}\rangle}\right) = \\
&= N - \langle k(s_i)\rangle - \sum_j e^{-2\frac{\hat{s}_i\hat{s}_j}{\hat{T}}}.
\end{aligned} \tag{6.10}$$

which constitutes an extremely accurate prediction (see Fig. 6.5 lower panel)[7] . The asymptotic cases for the ensemble averages are easy to

The distribution of a sum of Bernoulli random variables is called Poisson Bernoulli and has well-studied properties [157], albeit their moments are difficult to compute. One can, however, give bounds to the error committed whenever assuming a Poisson approximation also for the degrees.

[7] In this case we can even obtain a closed analytical expression for the case of power-law distributed strengths:

$$\begin{aligned}
\langle k(\hat{s}_i\rangle &\simeq N\left(1 - \int_{\hat{s}_{min}}^{\hat{s}_{max}} e^{-\frac{\hat{s}_i}{\hat{T}}s} p(s)ds\right) \\
&= N\left\{1 - (\gamma-1)\left(\hat{s}_i\frac{\hat{s}_{min}}{\hat{T}}\right)^{\gamma-1}\left[\Gamma\left(1-\gamma, \hat{s}_i\frac{\hat{s}_{min}}{\hat{T}}\right) - \Gamma\left(1-\gamma, \hat{s}_i\frac{\hat{s}_{max}}{\hat{T}}\right)\right]\right\}.
\end{aligned} \tag{6.11}$$

asses: For small strengths we have that $\hat{s} \ll \hat{T}/\hat{s}' \, \forall \hat{s}'$ which leads expression (6.10) to $\langle k(\hat{s}) \rangle \sim \hat{s}$ (converging to a Poisson distribution for degrees due to the properties of the Poisson Bernoulli distribution), while for large strengths one has $\hat{s}\hat{s}' \gg \hat{T} \, \forall \hat{s}'$ which leads to fully connected nodes $\langle k(\hat{s}) \rangle \sim N$ with vanishing variance.

Results comparing simulations and equation (6.10) are shown in Fig. 6.5 (lower panel), where an interesting transition is observed for $\gamma < 2$: The degrees are exactly equal to the strengths for small values of $\hat{s}$ (as expected by conservation of the edges) and evolve to a scaling of the type $k(\hat{s}) \sim \hat{s}^{\gamma-1}$ that finally leads to a saturation due to the bounded nature of the observables ($k(\hat{s}) \leqslant N$).

As I already pointed out to the reader in Section 4.2.2, we observe here clearly that *a scaling relation of the kind* $k(s) \sim s^{\beta}$ *is not always a reliable trace of relevant correlations.* More concretely, we have seen that in our framework, and for the case of power-law distributed strength sequences in particular, it is solely a consequence of the imposed constraints. In other cases, it might or might not be an indicator of correlations not imposed by the strength sequence, but one cannot assume either case *a priori*: Since this metric heavilly depends on the strength sequence, it always requires comparison with a null model. Alternatively, this also shows (for the case $\gamma > 2$) that a relation $k(s) \sim s$ does not necessarily imply $\bar{t} = \text{Ctnt}$ [24].

6.5.2.1 *Disparity, Average neighbor properties and general metrics*

In recent times, efforts have been devoted to extend well-known magnitudes on binary graphs to weighted graphs: Having appropriate null-models for multi-edge graphs permits to assess the applicability of such *weighted* extensions [12]. To this end, one can use the results of the GC ensemble to compute with high accuracy any network metric expressed in terms of t_{ij}: As an example we consider widely used magnitudes such as the disparity $Y_2(s_i) = \sum_i t_{ij}^2/s_i^2$ [160] and weighted neighbor average strength $s_{nn}^{w}(s_i) = \sum_j t_{ij} s_j / s_i$ [24]. Using again (6.4), one reaches after some algebra the following expressions

$$\langle Y_2(\hat{s_i}) \rangle \simeq \frac{1 + \frac{\hat{T}_2}{\hat{T}^2}\hat{s}_i}{1+\hat{s}_i} \left(1 + \frac{\left(\hat{T}^2 - \hat{T}_2\right)\left(2\hat{s}_i + 3\right)}{\left(\hat{s}_i + 1\right)^2 \left(\hat{T}_2 \hat{s}_i + \hat{T}^2\right)} \right) \tag{6.12}$$

$$\langle s_{nn}^{w}(\hat{s}_i) \rangle \simeq \left(1 + \frac{\hat{T}_2}{\hat{T}} \right) \left(1 - \frac{\hat{T}_2 + \hat{T}\hat{s}_i - \hat{s}_i^2}{\hat{T}(\hat{T} + \hat{T}_2)} \right). \tag{6.13}$$

where $\hat{T}_n \equiv \sum \hat{s}_i^n$. The average values and their fluctuations are in excellent agreement with the simulations, as can be seen from Fig. 6.6 panels A, B, E, F.

The results show several interesting features: On the one hand, the expectation for the disparity is not $Y_2(\hat{s}) \sim k_i^{-1}$ as assumed under a

Where $\Gamma(s, x)$ is the incomplete gamma function.

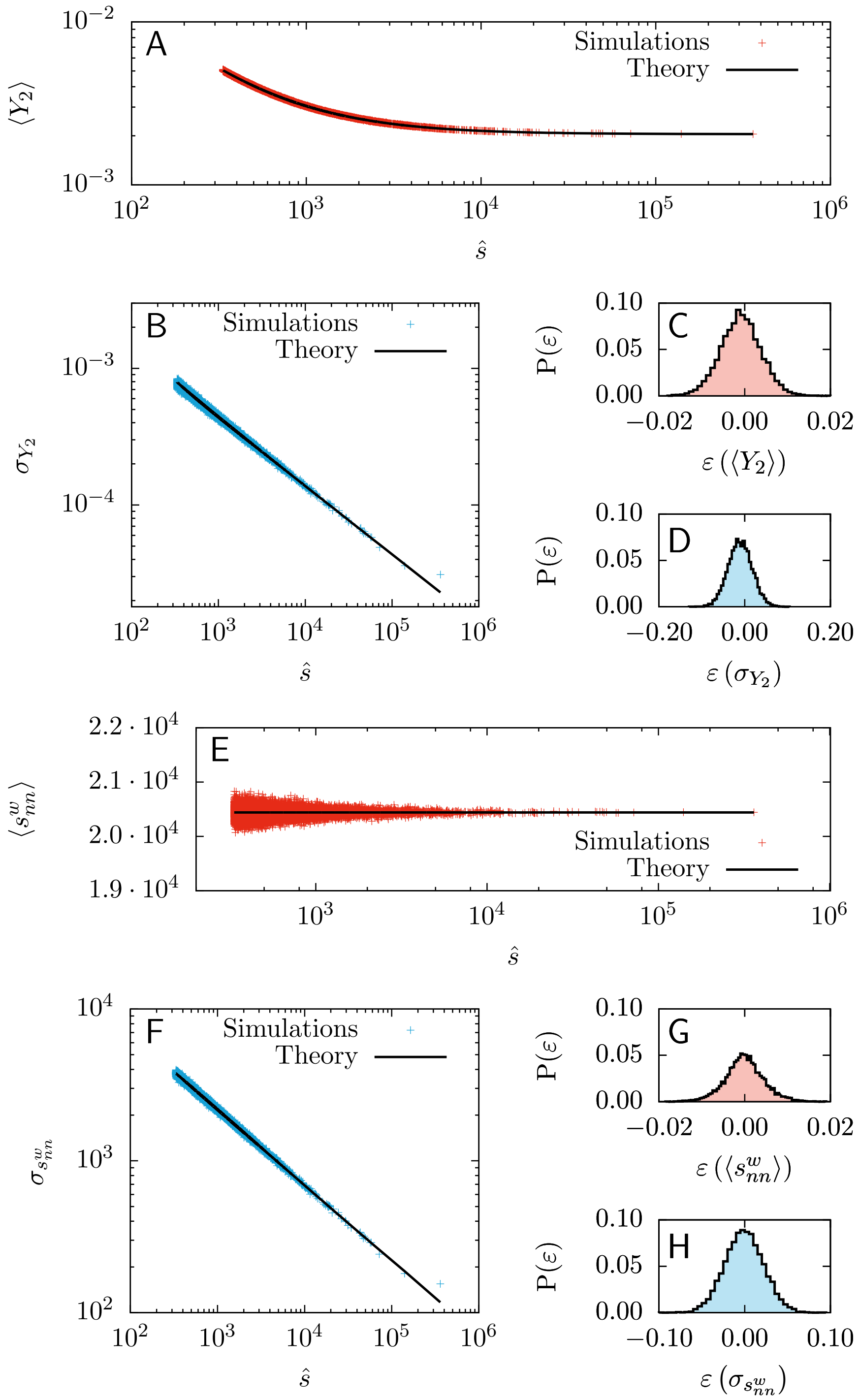

Figure 6.6: **The accuracy of the GC predictions.** Ensemble average (A, E) and standard deviation (B, F) for individual node disparity Y_2 and weighted neighbor average strength s^w_{nn} for power law distributed sequences with $\bar{s} = 1000$ and $\gamma = 2.5$. C, D, G, H: Histogram of relative error between theory and simulations averaged over 1000 repetitions for the values in Figs. A, E and B, F. A single outlier corresponding to the lowest value of σ for both the disparity and the average neighbor strength is not shown in the histogram on Figs. C, G.

total random allocation of edge weights [158], but rather decays as $Y_2 \sim \hat{s}_i^{-1}$ and rapidly converges to a plateau, independent of the chosen strength distribution. The weighted average neighbor strength displays an almost flat behavior which is a correct indicator of absence of correlations at the node level. On the other hand, the fluctuations of both magnitudes decay in a power law form as the strength of the node increases: This fact can be easily understood as increased connectedness imply higher availability of sampling.

6.5.2.2 *Comparison between simulated and predicted values for macroscopic observables*

To quantify the precision of our predictions, we computed the histograms of the relative error generated per node when calculating a given property z, $\varepsilon(z) = (\langle z\rangle_{si} - \langle z\rangle_{th})/\langle z\rangle_{si}$, where the subindices si stand for the Micro-Canonical (MC) simulations and th for the (GC) theoretical predictions in equations (6.12) and (6.13). The histograms in Fig. 6.6 panels C, D, G, H show the accuracy of the obtained results, providing numerical evidence for the equivalence between the MC simulations and the GC predictions, which is expected in the limit where an infinite sampling of events $\hat{T} \to \infty$ is available. Even when this requisite is not met, the use of the theory presented here constitutes an excellent approximation as shown in Fig. 6.7, where the relative error averaged over all nodes between ensemble expected GC magnitudes and simulations is shown for the different metrics considered for a increasing ranges of values of sampling.

I thus provide here experimental validation of the predicted equivalence between ensembles, developed in earlier chapters. Also, the presented case has served as example to show how to apply all the aspects derived in this chapter: We have generated networks using a MC approach and compared the predictions for macroscopic observables with the GC ensemble analytical approximations using linear approximations that yield excellent results. Furthermore, we have shown how differences between simulated and predicted values tend to dissapear as the sampling of the networks is increased.

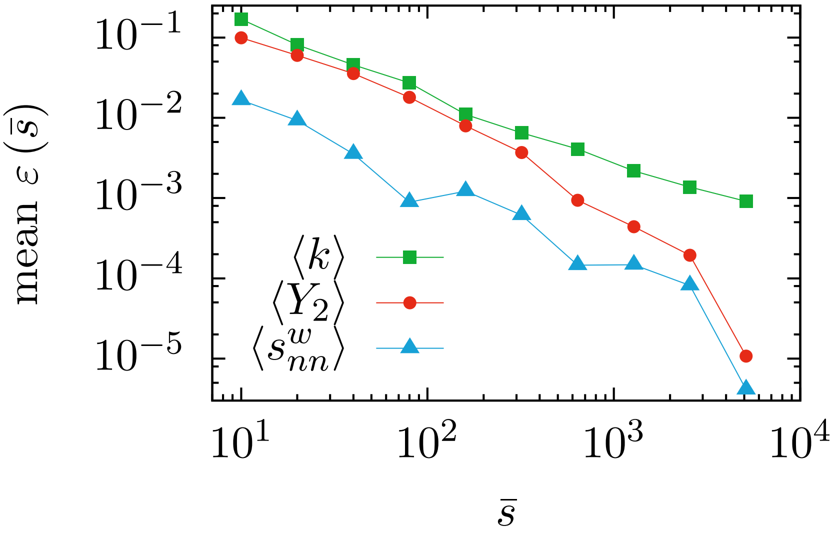

Figure 6.7: **Convergence between ensembles with increased sampling.** Relative error between ensemble average predictions and simulations, averaged over all nodes for degree, disparity and average neighbor strength for different values of sampling $\bar{s} = \hat{T}/N$ for 1000 repetitions each point, $\gamma = 2.5$ and $N = 2000$.

6.6 WRAPPING UP: NETWORK GENERATION MAIN TRICKS AND TIPS

In this last chapter of the first part of the thesis, we have mainly reviewed technical details and possible applications of the previous theoretical discussions performed in earlier chapters. In a nutshell, the present chapter can be understood as a manual for practitioners and mainly provides advices and instructions to effectively generate network instances for the wide variety of earlier discussed examples.

More specifically and reviewing what is done here,

A. **Saddle point solving:** We have provided all necessary explicit algorithms to solve the saddle point equations, with special discussion devoted to the W case.

B. **Network generation:** We have discussed the generation of networks in all ensembles considered, and given explicit recipes for both the Canonical and Grand Canonical ensemble. For the MC ensemble we have discussed the main problems of network generation and only provided a specific example for the non binary ME configuration model.

C. **Analytical prediction of macroscopic examples:** In order to avoid simulation, we have provided also a recipe to analytically approximate expected values and fluctuations of common network metric and observables. Furthermore we have exemplified such a recipe for the case of the non binary ME configuration model.

D. **Experimental comparison of cases:** By making use of the all of the above, we have proceeded to discuss the differences that can be observed whenever generating ensembles with the same constraints for different cases ME, W and B.

E. **Experimental validation of ensemble equivalence for ME linear case:** For the non binary ME configuration model, we have shown ensemble equivalence by calculating analytically in an extreme precise way predictions for average values and fluctuations for network metrics in the GC ensemble. Then we have compared them with MC simulations of the same model and validated that the (small errors) between predictions and simulations tend to disappear in the high sampling limit.

The take home message of this part of the thesis is that in order to perform a meaningful analysis on a given network, a practitioner needs to be able to select an appropriate null model, which not only depends on the endogenous constraints one considers but also on the very nature of the process one is modelling [49]. This work provides researchers with a range of maximum entropy (and maximum likelihood) models to choose from, covering a wide spectra of possibilities for the case of non binary networks. Each of this models is not *wrong or even right* in a general case despite yielding very different predictions for the same sets of constraints, but just more or less appropriate depending on the problem at hand one wants to study [186].

As a rule of thumb, when in doubt, one should use the Multi-Edge description, since it is the case which has the most desirable properties: Vanishing fluctuations, event specific entropies, well defined high sampling limit, ensemble equivalence and easiness to solve the constraint equations. Furthermore, we have seen that such a case is the limiting one for all once $M \to \infty$. Finally, usually real data collected tends to be time-stamped, which introduces an effective distinguishability between the events forming a network, hence converting the Multi-Edge description in highly satisfactory.

To conclude this part of the thesis, a careful word of advice need to be once more repeated: The blessing of big data may also be its dearest danger! High dimensionality data sets require sophisticated null-models to detect the effects of the system constraints on the given observables and hence comparison with a null model to assess the relevance of observed features in real data is always needed. The present application of ensemble theory to networks aims to draw attention to this problem and to close this gap for the case of non binary networks.

Part III

DATA ANALYSIS AND MODELLING OF URBAN MOBILITY

Any mathematical formulation, *physical* theory or scientific hypothesis is incomplete if it remains unchallenged. Data on phenomena extracted from the *real* world, despite its many possible limitations, is the only way by which we can test our predictions and conjectures. In this part, we proceed to apply the mathematical theory developed earlier to study empirical datasets on human urban mobility. Additionally we propose explicit, ready-to-use applications of the ensemble approach earlier developed to solve open problems in the field. Also a critical review on existing models on mobility generation is performed.

7 URBAN MOBILITY DATA: OVERVIEW

The purpose of computing is insight, not numbers.

— Richard Hamming [92]

We live in the so-called *Big Data era*. Most of our daily activities on either the virtual or real world leave a trail of breadcrumbs in the form of data that is captured and stored by external agents (more often than not private corporations). The increased pervasiveness of information and communication technologies is enabling the tracking of human mobility at an unprecedented scale. This in turn is fostering research on mobility from indirect sources, based on the availability of these data sources (and the willingness of their owners to share them). Massive call detail records from mobile phone activities [88, 38] and the use of global positioning systems (GPS) in large vehicle fleets [28] for instance, are generating extraordinary quantities of positional and movement data available for researchers who aim to understand human activity in space. Other data sources, such as observations of banknote circulation [46, 181], online location-based social networks [152, 151], radio frequency identification traces [51, 148], or even virtual movements of avatars in online games [180] have also been used as proxies. These studies have provided valuable insights into several statistical aspects of human mobility, uncovering distinct features of human travel behavior such as scaling laws [46, 168] or predictability of trajectories [169] among others.

Besides empirical studies, the surge of available data on human mobility has also opened the door to validate, revisit and innovate on abundant previous work on theoretical models of mobility at several scales. Such models can have diverse applications in a wide variety of disciplines ranging from epidemiology to urbanism [94, 29, 52, 23], with special importance in city planning and policy action [27].

Despite the potential hinted by this successful studies, the theoretical development of tools and techniques for handling massive data sets of human mobility and for assessing their possible biases is still a road full of obstacles. Besides ethical and moral questions concerning data-acquisition, privacy and data-ownership, also methodological problems (old and new) are still open.

The philosophical issues derived from the data collection at huge scales will be shortly discussed in Part iv.

One could argue that the main challenges are the size of the data analyzed, the multiple scales involved, the highly skewed statistical nature of human activities [129] and the lack of strict control on the data collection processes and protocols. Enunciating those in the form of open questions, this boils down to:

A. **Sampling:** How much sampling is needed to get an accurate picture of a mobility process? How can we compare datasets with different sampling?

B. **Normalization:** How can we compare mobility traces generated by the same mechanism in distinct environments (different geographical layouts with different characteristic scales)?

C. **Characterization:** Does mobility data exhibit *common* patterns? Can we characterize them into intensive indicators (independent of size and or sampling)?

D. **Modelling:** Can we explain the origin of the detected common patterns? Or, at least, can we exploit them for predictive purposes?

E. **Validation:** Human generated data usually displays highly skewed distributions and highly non-gaussian statistics. How can we appropriately validate proposed models of mobility? Which quality indicators must we use?

This second part of the thesis, is devoted myself to the analysis of mobility demand at urban scales, understanding mobility as the process by which people choose the destination of their trips in cities. To do so, I shall consider that the two main elements shaping this process are the decisions taken by the citizens (on average) and the context in which they take them, that is, the city. These decisions are affected by the constraints or limitations of their personal circumstances (budget or *expected gain* from a trip) but also by the context. Not all places in a city are equally *important* or appealing nor equally visited. Once more, we must emphasize that this thesis is not concerned with the routing of the trips, once their origin and destination has been chosen. It is centered on the previous stage, that of establishing the origin and destination of each trip.

I am thus interested in people moving in cities. It does seem necessary then to face the questions enunciated above and tackle the challenges they contain. In the subsequent three chapters I will try to do so while exploiting the theoretical tools developed in the first part of this document. In this chapter in particular, I review the main methodological problems I believe a practitioner faces when dealing with urban mobility data and I present the solutions adopted in the present case. I shall put special emphasis on the solutions that network science offer to this particular data mining problem.

In Chapter 8, I describe the main common empirical features of urban mobility considering temporal, topological[1] and spatial patterns from a network science point of view. In Chapter 9 I perform a critical

[1] In this part of the thesis, whenever referring to *topology* we mean the features of a certain dataset from a network yet spatially agnostic perspective.

review on existing models of mobility and their validation/applicability. Finally, Chapter 10 is devoted to the application of the insights and methodologies developed in the present thesis to solve specific urban human mobility related problems.

7.1 PEOPLE IS WEIRD: EMPIRICAL DATASETS FOR THE ANALYSIS OF URBAN MOBILITY

We want to study real mobility problems, hence, in order to perform a comprehensive study of urban displacements, we need a reliable source of empirical data. Furthermore, since we want to deal with data normalization challenges, we want this data to accurately track the activity generated by the same process in distinct urban environments.

Throughout this thesis, the main source of data used will be that generated by Taxi displacements. I have at my disposal time-stamped datasets that contain GPS trajectories of taxis in four different cities, which are located in 3 different continents (with inhabitants correspondingly having diverse cultures). Those are New York (NY) and San Francisco (SF) located in North America, Singapore (SI) located in Asia and Vienna (VI) located in Europe.

Sadly, two out of the four datasets are not public yet the majority of the analysis will be centred in the dataset containing the most sampling[2] which happens to be public, that of the borough of Manhattan in New York. In all cases, the studied trajectories are only those where the taxis were occupied by customers and thus these datasets represent their mobility choices and not those of the drivers.

Appendix C contains all detailed information about the datasets used in this thesis.

Albeit one can obtain traces of human mobility from a variety of sources, the use of Taxi fleets acting as proves has many advantages.

To begin with, all the data at our disposal are evidently generated by displacements using the same mode of transport. Also, one can consider each trip as a statistically independent event from the others, since in a city with enough population, the probability of a person taking enough taxis as to be able to bias an entire dataset seems extremely remote (in contrast to the skewed profiles of activity detected for online social networks such as geo-located Twitter activity [114] or Foursquare [131]). This allows a safe aggregation of all user generated traces to obtain a dataset from where to extract statistically sound conclusions. These two factors are important to consider since a variety of studies on human mobility using other data sources [46, 168] have been shown to generate somewhat *spurious* macroscopic statistical features that can be explained by the aggregation of a diversity of modes of transport [194] and specific user characteristics [88, 133].

[2] During this part of the thesis, the adjectives *big* or *small* will be used as synonyms of *dense* or *sparse*, relating to the total sampling of each dataset.

Furthermore, for this mode of transport a direct relation between distance travelled and costs for the user can be established (the more distance/time one travels, the more one pays) so the possibility to introduce budget related mobility constraints is clearly open. This might not be so obvious for the case of public transport records [**?**] because of the existence of Fare zones[3]. Also, Taxi trips are well defined trajectories between two points in space and time, and this data is not opportunistically recorded (such as Foursquare check-ins [131] or call detail records (CDR) [88]). Yet, it is passively obtained without asking users to fill in tedious surveys (with the possibility of errors) as would be the case of usual census studies [117]).

One of the possible shortcomings of our datasets is that although they are extensive, their coverage of the entire population under study might be limited. Data from taxis, or from other vehicle fleets [28] are typically obtained from a single company, which usually represents only a small fraction of the actual number of vehicles circulating in a city. In our case, however, the dataset obtained from the city of NY does overcomes this limitation as all the yearly generated trips are recorded, and I will precisely use this situation to effectively tackle the problem of limited sampling in Section 10.1.

Even if Taxi data seems to be satisfactory for the objectives at hand, one must not forget that this corpus of data is also prone to errors derived from imprecisions of the GPS. In our case, I will be only interested in the starting and ending points of trips (as to simplify the already complicated analysis). To eliminate unreliable data, we have matched those points with the nearest intersection obtained from the street network (openly available from [132], see Figure 7.1) for the different cities. All trips with either (or both) the starting and ending point further away from immediate walking distance (200 m) of an existing intersection have been eliminated from the dataset. More details are provided in Appendix C.

Finally, a note must be made about commuting: We are interested in mobility in cities caused by spontaneous actions of everyday life, and hence we shall not focus ourselves here in the phenomenon of commuting. Commuting is understood as the movement of people from their homes to fixed work places and vice-versa, which displays strong temporal and spatial correlations which are very likely not captured by Taxi datasets. Other passive recording methodologies such as Call Details Records (CDR) of data from online social media (OSM) most surely do not capture this fact either, although [114] shows that CDR and Twitter data lead to quantitative similar results when focused on urban environments. They are nevertheless more susceptible to be affected by it, since taxis seem an unlike source of transportation for everyday use.

[3] It is true that in some cases taxis also have defined Fare zones, but the vast majority of trips happen in general within the main urban zone.

7.2 cities are complicated: the correlation-causality problem

Now that we have obtained an accurate proxy of human mobility at urban scales, we must consider the layout in which these trajectories are generated: The city.

Recently, a *science of cities* [27] has been proposed to study urban systems from a complexity science perspective. Cities are *alive* entities where many processes happen simultaneously. They are composed by many interacting sub-systems which are highly heterogeneous and co-evolve at many temporal scales. Cities are thus inherently anisotropic, densely populated and highly distinctive in terms of geographical layout, form and size across the world. In fact, there does not seem to be a consensus on how to measure their spatial extent [119]. Furthermore, cities concentrate various groups of people with diverse cultures, motivations, objectives and life standards.

The interaction between all these ingredients shapes the traces one is able to observe while tracking mobility. The data we record is clearly a mix of individual freedom (represented by each person mobility choices), their constrained reality (budget, needs) and their adaptation to this reality.

To accurately understand a mobility process, one must try to isolate its characteristics as much as possible from the underlying framework where it is developed, which in this case is the already diverse urban context. We must hence take special care in analyzing the intrinsic dynamics of mobility within the city at the temporal and spatial level, in order to see whether a meaningful analysis can be carried out and at which level of detail.

Furthermore, if our objective is to study the same phenomenon in different contexts (mobility using taxis) and we are implicitly assuming that such a phenomenon shares common traits in the different cities, one must be able to effectively compare the different datasets accounting for differences in sampling (number of trips recorded), geographical layout (associated typical scales) and possibly temporal dynamics. In the subsequent parts of this chapter we analyze the different common features the data shares when regarded from an urban point of view.

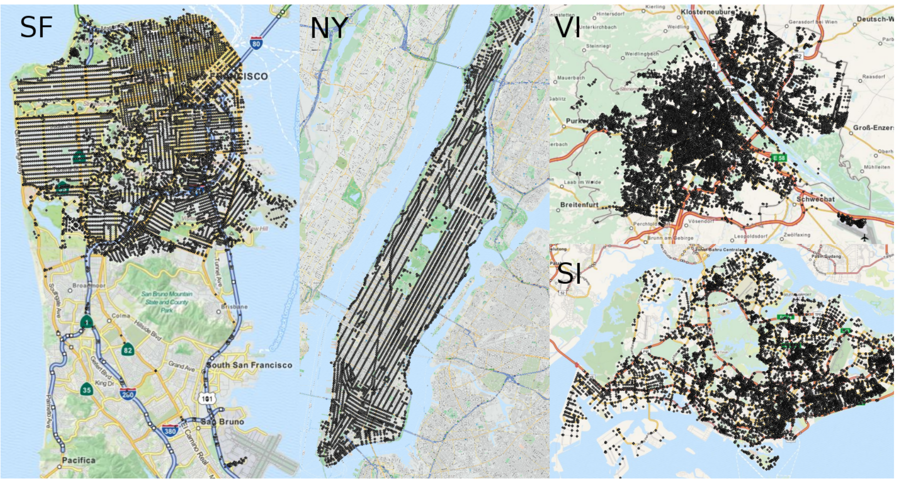

Figure 7.1: **Different layouts for the same mobility process.** Maps of the four different cities for which we have data on Taxi displacements with the considered intersections for trip matching overprinted using black dots. The SF, SI and VI cases include the airport as it concentrates a relevant fraction of the total traffic. Background maps obtained from [132] and image generated using [142].

7.2.1 *Cities are different: Same phenomena in different layouts*

For the reasons given above, directly comparing urban mobility datasets can be problematic, as the influence of the specific geographical layout of each city can heavily influence the observed mobility patterns on it. In our case, the four different cities for which we have data have very different geographical structures as can be seen in Figure 7.1: While the borough of Manhattan or the city-state of Singapore are islands, Vienna is situated in the mainland and San Francisco in a peninsula.

Not only their shapes are different, but also their sizes. Accordingly, for each city, the definition of *large distance* may vary depending on its extension and density: The relative position of distinctive spots such as Airports (which usually accumulate a macroscopic fraction of the observed Taxi traffic), which may be situated at different distances from the center, influence heavily space dependent quantities. If we compare the four different cities at hand using a distance based metric such as the distribution of trip lengths (Figure 7.2)[4], we clearly observe an initial sudden increase (people tend not to ride taxis for very short distances) followed by a steady exponential decrease (in accordance with studies in other cities [116]). Such a decrease, however, has different slope for each city under study due to varying size. It is furthermore interrupted at a value of 20 km (the order of magni-

[4] In all cases where lengths are considered in this thesis, those are approximated considering the projected Euclidean distance between starting and ending points for each trip.

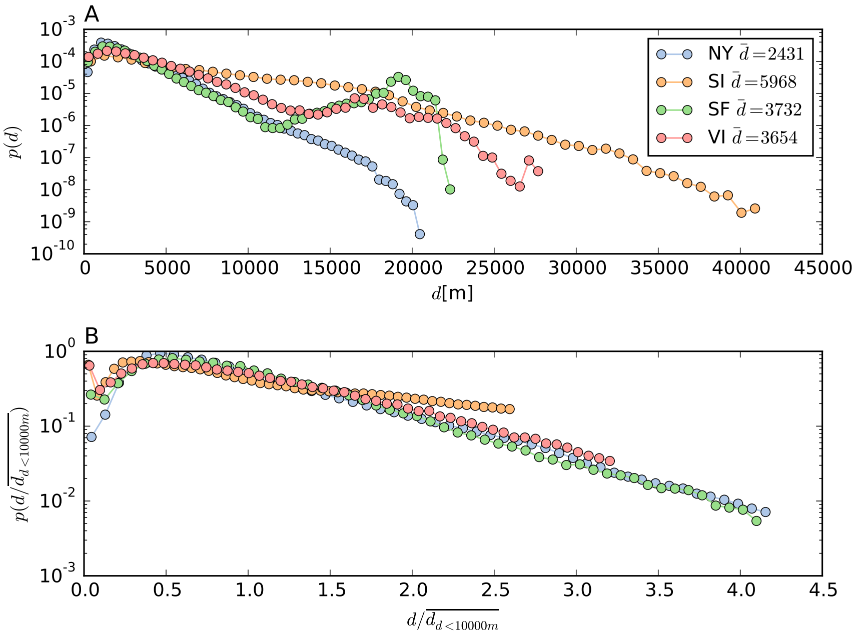

Figure 7.2: **Geographical layout effects on occupied trip length distribution.** The presence of airports which accumulate a large portion of the traffic in the SI,VI and SF datasets at different distances from the city center induce notable changes in the tail of the distribution of trip lengths (A), which can otherwise be approximated by an exponential function as seen in (B), where the distribution of distances shorter than $d = 10$ km is shown with rescaled values over the conditioned mean.

tude of the distance between airports and city centres) with a sudden increase for VI and SF, since the airport in these cases is isolated from the city while in Singapore the airport is embedded in the island and the effect is smoothed.

To avoid these problems related to geography, one must use null models that effectively balance the different city layouts and leave us with a somewhat context independent picture of the phenomenon of mobility. To this end, one can generate a mobility network considering each location of the city as a node. Such networks are usually called Origin-Destination matrices (OD) and each of their integer valued entries count the number of trips between two given locations [70, 60, 189]. This modelling framework combined with considerations made earlier about the basic ingredients at work when facing the problem of mobility generation makes this problem fully compatible with the ensemble treatment derived in the first part of this thesis.

Although the network approach may be seen as a convenient way to wash away geographical effects, there is still one issue which must be discussed. Networks are discrete objects but points in space from where we build these networks, in contrast, *live* by nature in a continuous metric. Wherefore the way to aggregate these points is not unique and can have a significant effect in our analysis. Considering the methodology I followed to filter the data, a logical choice to do this aggregation is to consider the intersections of each city (Figure 7.1) as the nodes from which the network is formed, assigning trips to events connecting the closest nodes to their starting and ending points. However, this is only one of the possible ways to perform this aggregation and so deserves a comment[5].

Finally, even if the network approach seems convenient to tackle the geographical differences across datasets, there are still a few issues related to city structure and activity that must be observed. They are mainly two: *a) Which is the right (network) null model to choose?* and *b) How can we deal with the temporal dimension of taxi trips and seasonality effects using a static description such as an aggregated network?*

7.2.2 *Cities are heterogeneous: Anisotropy of population density and activity*

In principle, choosing the right constraint we would be able to build the appropriate null model to compare to the data applying the formalism earlier derived. The main problem reduces now to finding realistic constraints and understanding the type of network under study. As previously mentioned and shown in Section 6.4, different Taxi trips represent fully distinguishable events, and hence we shall deal with Multi-Edge networks (ME). This has several positive implications. These models have well defined properties in the high sampling limit and hence, by using its definition, one is provided with an effective normalization strategy for datasets of different size.

Recalling the definition of high sampling limit, (3.18), the important variables at play will be those related to the relative fraction of trips allocated between intersections, that is:

$$\hat{p}_{ij} = \frac{\hat{t}_{ij}}{\sum_{ij} \hat{t}_{ij}} = \frac{\hat{t}_{ij}}{\hat{T}}, \tag{7.1}$$

hence a natural scaling variable for occupation related metrics will be the total number of observed trips $\hat{T}$.

The complete analysis of the different datasets from a Multi Edge Network perspective is performed in Chapter 8.

In all the datasets studied, the distribution of strength per node (total number of people entering and leaving each location), either in the incoming or outgoing direction, is highly anisotropic across space and also displays highly skewed distributions (see Figure 7.3): There are orders of magnitude of difference between the trip generation/attraction of nodes in the city. Furthermore, the spatial allocation of these nodes is not arbitrary but depends on the specifics of the city.

[5] The non unique way to aggregate spatial points is known as *modifiable area problem* (MAUP) [48, 192] in geography studies and remains an standing open problem.

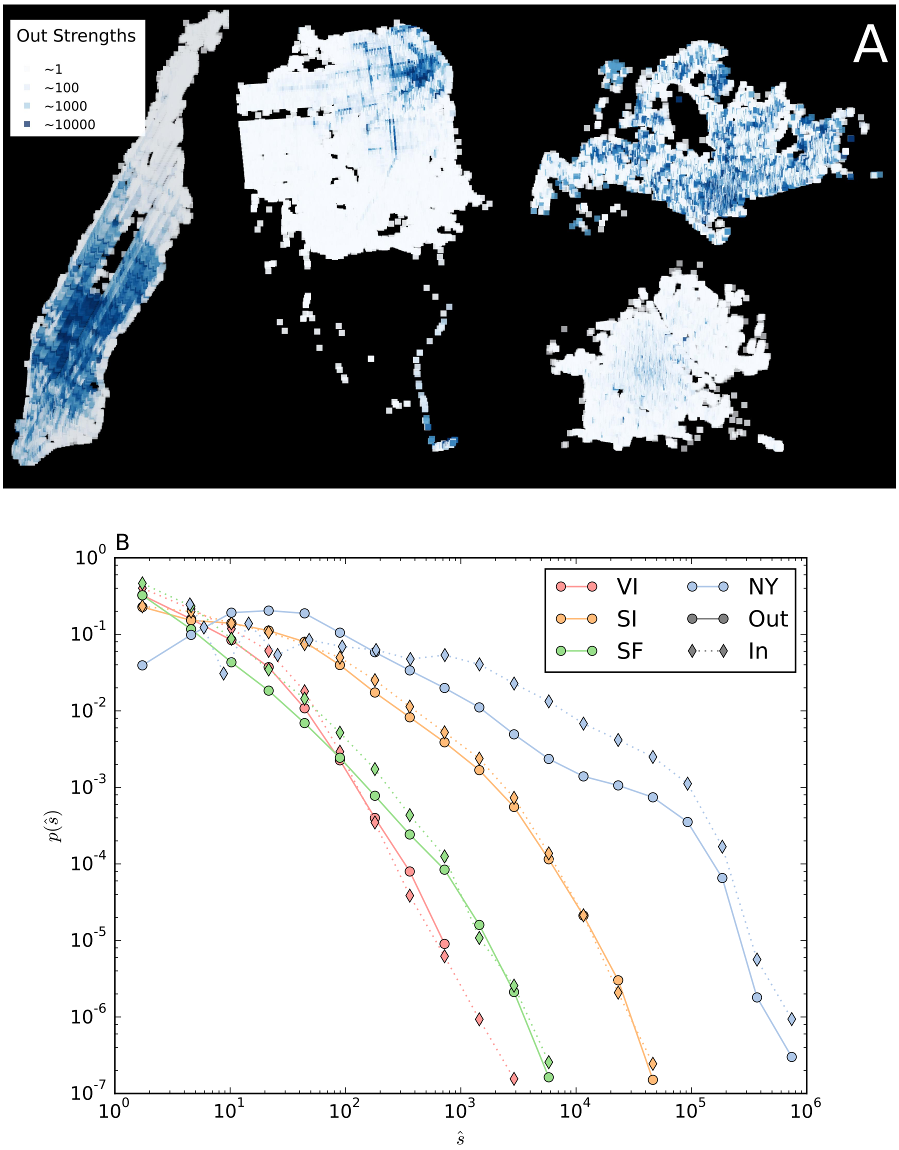

Figure 7.3: **Node distribution of strengths across different cities.** Spatial inhomogeneities of outgoing strength distribution (A) and its highly non-gaussian distribution (B). Intersections are represented as squares in the map whose color corresponds to their outgoing strength using a logarithmic based scale. Figure B uses log-binning for the representation of the curves.

Considering the detected anisotropies, the most basic null model balancing easiness of use and approximation to reality would then be the already studied Multi Edge non binary configuration model (MECM, see Section 6.5). By using such a model, we assume that the mobility needs and accessibility are given by external factors such as population and density of work places or infrastructures, hence each node has a certain *attractivity* which macroscopically, is related to the total number of taxi trips entering and leaving each location (incoming and outgoing strength pairs per node $\vec{s} = (s^{out}, s^{in})$). The dynamics of change of all these factors encoded in the attractivity of a node change slowly in time (in a timescale of years), at a pace much slower than the typical time a urban trajectory takes. Hence assuming them as "initially given" is a good starting point for the analysis.

7.2.3 *Cities are alive: Circadian rhythms and seasonality effects*

While the built structure of cities evolves slowly in time, many dynamic and behavioural processes that take place within a city unfold relatively fast, and in principle could be strongly variable across time. However, human activity in cities exhibits highly regular patterns when observed over well defined periods of time, such as circadian or weekly rhythms. Cities are not an exception, and as *alive entities*, display cyclic behaviours at different scales (daily, day of the week, weekly, monthly and so on). If we are to use a static description of mobility, to what extent are we missing those dynamics?

The good news in this case are that Taxi trip generation seems to be very stable across time, and although it shows some seasonal fluctuations, these can be considered as second order effects. Figure 7.4 shows the temporal trip statistics of the four available datasets. The number of recorded trips as our observation time advances (Figure 7.4-A) can be very well fitted by a linear relation, indicating that the trips generated per day are concentrated around their mean value. In fact, their distribution closely resembles that of a gaussian distribution, see the quantile-quantile plot in Figure 7.4-B.

The effects of weekly activities can be seen in Figure 7.4-C, where the number of trips averaged per weekday is shown: Taxi usage peaks on Fridays and Saturdays (leisure days where people tend to go out) and drop on Sundays with a stable behaviour during work days. The overall variation, however, does not lie further away than two standard deviations from the mean in any case.

Finally, city *circadian rhythm* can also be observed in the hourly distribution of trip starting times. As expected, a valley is observed during night times when activities in the city significantly decrease. Furthermore, the general statistics of trip generation across hours in the city validate our original hypothesis of event independence among trips. Figure 7.5 shows that the inter-event time distributions between

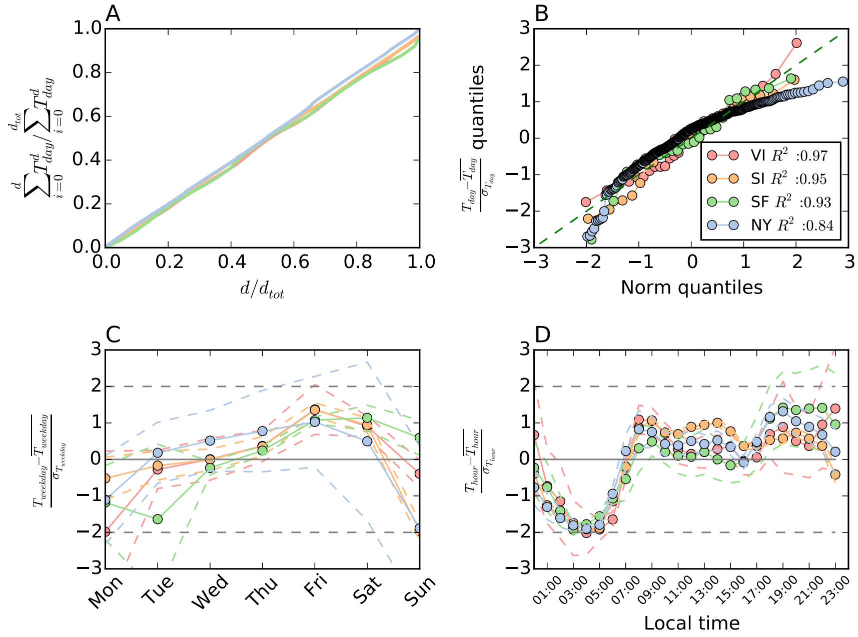

Figure 7.4: **Temporal statistics of trip generation across cities.** Fraction of generated trips in the datasets as a function of elapsed days (A) and quantile-quantile plot of distribution of generated trips per day compared to a normal distribution (B) with linear coefficient of determination R^2. The trip generation is very stable across time at day level. Variabilities are detected when computing averages across weekdays (C) and day-hours (D), yet the statistics are not extreme with the higher variability not exceeding two standard deviations in each case (the values are standardized).

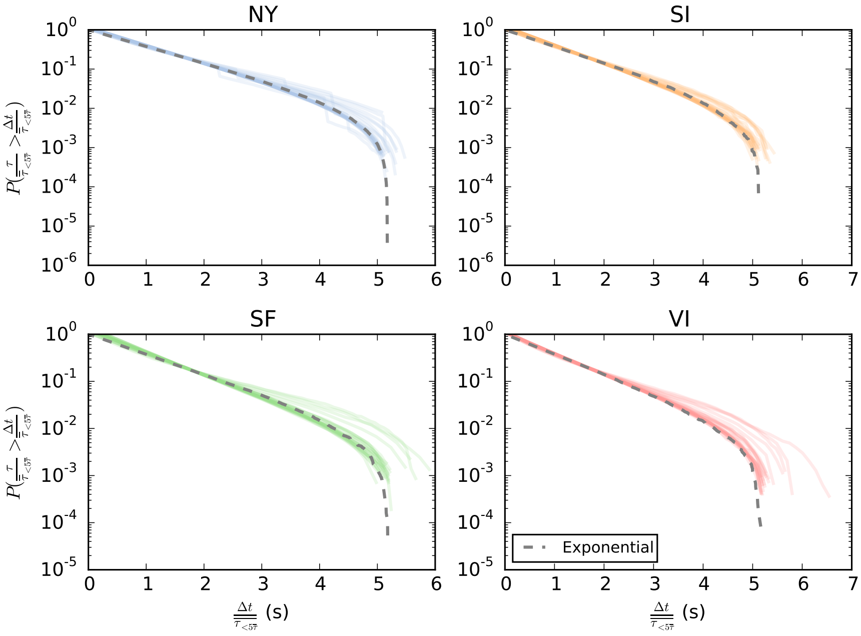

Figure 7.5: **Evidence of hourly Poisson trip generation across cities.** Complementary cumulative distribution of hourly inter-events starting trip times aggregated over the entire days of the dataset and rescaled over their mean values. Outliers lying further away than five times the mean of the distribution (< 2.5% of the data in the worst case) have been omitted and subsampling considering only trips generated by 500 randomly chosen taxis has been applied to the NY and SI cases to account for dataset finite resolution (seconds). Transparency has been applied to each of the 24 lines in each plot representing a given hour of the day as to appreciate the concentration of points. A line generated with exponential synthetic data corresponding to inter-event statistics of Poisson generated events is also shown.

successive trips agree very well with an exponential distribution, signature of Poisson trip generation.

To study yearly seasonal effects, one can take the NY dataset which spans a complete year. As can be clearly seen in Figure 7.6, holidays and extraordinary events (in that year the hurricane Irene hit New York and this can be seen with a substantial decrease of Taxi trips) affect the statistics, which are otherwise very stable (but clearly not random!).

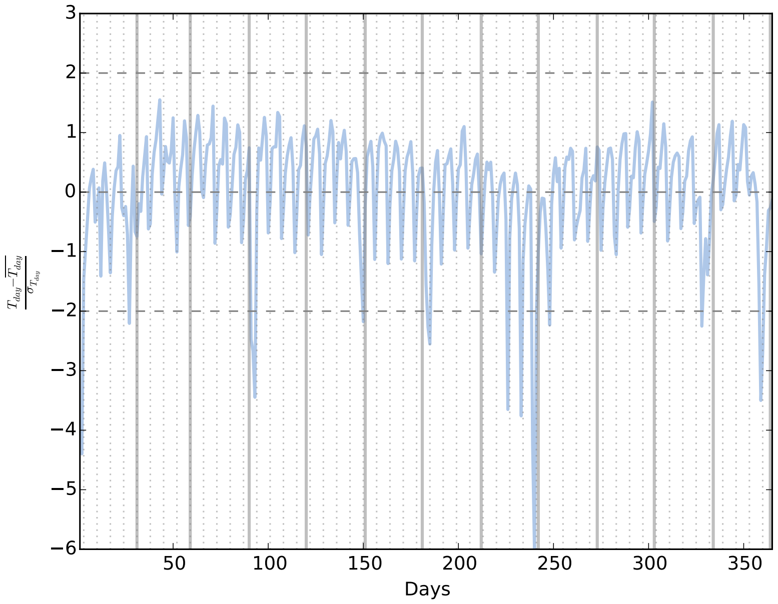

Figure 7.6: **Seasonal yearly dynamics of Taxi trip generation in the New York dataset.** Number of recorded trips per day over the entire year. Dotted lines mark the start of each week (Mondays) and filled lines the start of each month. The effect of the hurricane Irene and also holidays is apparent together with the weekly circadian rhythms.

All the above observations suggest that temporal patterns are sufficiently regular with regards to mobility and hence can be overlooked at a first level of approximation. If one wishes to go further, the observations indicate that the next step to consider would be an hourly separated analysis (also static in each case) with increasing level of detail accounting for weekday-weekend separation and finally seasonal effects.

In this case, I will restrict my analysis to a static picture, yet all of the proposed tools and considerations can be refined using the above reasoning to take into account the different variations caused by cyclic effects. To discard the effects that unexpected events on particular days (such as hurricanes) may have on the dataset, the statistics of daily trip generation will be approximated as gaussian and all days with values exceeding two standard deviations from the mean will be eliminated from the datasets.

7.3 WRAPPING UP: CONSIDERATIONS ABOUT URBAN MOBILITY ANALYSIS FROM A NETWORK PERSPECTIVE

In the present chapter, that serves as introduction to this second part of the thesis, important preliminary considerations about the methodological challenges to be faced when studying urban mobility patterns have been presented. Apart from considerations about general issues to keep in mind in mobility analyses, the main ingredients and assumptions we will consider are discussed. In a nutshell, the discussion can be grouped in different aspects:

A. **Empirical Data:** The justification for the use of Taxi data and its strength and weaknesses compared to other mobility sources have been discussed. Also the filtering procedure applied to datasets is described.

B. **Spatial dimension:** A network approach has been proposed to mitigate the effects that different city layouts have on distance based observables.

C. **Temporal dimension:** An analysis of temporal patterns has been performed to justify the static framework used for the analysis that will be carried out due to recurrent periodicity of trip generation.

D. **Null model choice:** Based on observed spatial anisotropy and skewness of node related features, the choice of the MECM studied in earlier chapters as null model has been justified.

E. **Common pattern detection:** In all the studies carried out in the different datasets, common patterns have been observed which justify the purpose of this part of the thesis of analysing general features in mobility data across different cities.

To conclude this introductory chapter, I emphasize the practical approach taken in this document: My aim is to provide interested readers with appropriate, ready-to-use tools and methodologies for the analysis of mobility. While the specific applications presented here apply to the case of Taxi mobility, they are discussed with the focus placed on generality, with the aim to be easily adaptable to other situations.

8 URBAN MOBILITY NETWORKS EMPIRICAL FEATURES

It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is ? if it disagrees with experiment it is wrong. That is all there is to it.

— Richard p. Feynman [75]

In the preceding chapter, I have tried to justify the modelling point of view that we will take with regards to the analysis of urban mobility. While advocating for the network approach, I have mentioned that since Taxi trips performed by different customers are clearly distinguishable events, if one wishes to study mobility networks one must do so taking as reference the earlier introduced Multi-Edge structures (ME). These structures are generated from the data by aggregating over a certain timespan τ (depending on each dataset) all trips recorded between a given set of locations taken as nodes (road intersections). Obviously, the longer our observation time is, the more total trips will be recorded in the dataset and eventually, if we were to wait a very large (infinite) amount of time, we should in principle recover a fully connected network, i. e. an adjacency matrix where all entries are non-zero. This is so because in principle there are no forbidden connections between nodes (all trips can be performed) yet we obviously expect some of them to be much more frequent than others.

Having said all of the above, a relevant question to ask now is whether given the conditions fulfilled by Origin Destination matrices (ODs), those can be conveniently modelled as ME structures or not[1].

8.1 BUT CAN ORIGIN-DESTINATION MATRICES REALLY BE MODELLED USING MULTI-EDGE NETWORKS?

Recalling our previous theoretical work on Multi-Edge networks, it is worth noticing that mainly two characteristics will be useful on our analysis. On the one hand, we know that overlay multi-layered networks (non-binary networks constructed by aggregation of individual layers) are totally indistinguishable from Multi-Edge networks with a single layer once we observe their statistics. On the other hand, we

[1] Mobility networks are usually called Origin-destinations matrices (ODs) in transport terminology [60, 189]. We shall use both words indistinctly.

know that the L possible states (pairs of nodes) where trips can be allocated should display Poisson related statistics and that for a given sampling $\hat{T}$, the expected number of observed trips is $\hat{T}p_{ij}^{\infty}$ where p_{ij}^{∞} is a property of the node-pair independent of the sampling[2]

The combination of the two prior characteristics gives us a simple way to check whether we can model Origin-Destination matrices as Multi-Edge networks. Given that the generation of trips over days fluctuates around a stable value as seen in Figure 7.4-A, we can relate the sampling $\hat{T}(\tau)$ with the time dimension by using a simple linear relation $\hat{T}(\tau) \propto \tau$. If we assume that the constraints under which urban mobility is generated do not change over time, we can take our datasets and slice them in a series of aggregated snapshots over a certain time period τ', and then study the statistics of $\hat{p}_{ij} \equiv \hat{t}_{ij}(\tau')/\hat{T}(\tau')$ averaged over the different temporal slices. We will expect two things:

A. For two arbitrary snapshots with different sampling τ, τ', for all the intersections their values $\hat{p}_{ij}(\tau)$ and $\hat{p}_{ij}(\tau')$ should lie very close. This will give us an effective way to normalize our measures.

B. We should be able to calculate the two first central moments of the variables p_{ij} by knowing that they are quotients of Poisson distributed random variables. They should correspond to the empirically measured ones over time slices of any duration. For Poisson distributed variables we have, using (6.4),

$$\begin{aligned}
\langle p_{ij} \rangle_{\tau'} &= \left\langle \frac{t_{ij}}{\sum_{ij} t_{ij}} \right\rangle_{\tau'} \simeq \frac{\langle t_{ij} \rangle_{\tau'}}{\langle T \rangle}_{\tau'} \simeq \frac{\hat{t}_{ij}(\tau_{max})}{\hat{T}(\tau_{max})} = \hat{p}_{ij}(\tau_{max}) \\
\frac{\sigma^2_{p_{ij}}|_{\tau'}}{\langle p_{ij} \rangle |_{\tau'}} &\simeq \frac{1}{\langle T \rangle_{\tau'}} \frac{1 - \langle p_{ij} \rangle_{\tau}'}{\langle p_{ij} \rangle_{\tau}'}
\end{aligned} \tag{8.1}$$

To validate these hypothesis I have split the available data into n_τ equal daily time intervals and computed the relative dispersion of the values accumulated over the entire data set $\hat{p}(\tau_{max})$ around the measured values $\langle p \rangle_{\tau'}$,

$$\varepsilon = \frac{\hat{p}(\tau_{max}) - \langle p(\tau') \rangle_{\tau}'}{\langle p(\tau') \rangle_{\tau}'}, \tag{8.2}$$

where τ_{max} is the time at the end of the full observation period and the averages are performed over all the time slices of length τ' (Ta-

[2] A clarification on notation: For a random variable x, $\hat{x}$ corresponds to its measured value in the empirical dataset, $\langle x \rangle_{\tau'}$ to its (empirical) average over timespans of duration τ' and $\langle x \rangle$ to its ensemble average. $\hat{x}(\tau)$ corresponds to the aggregated value of random variable x over a timespan τ.

DATASET	TIME WINDOWS n_τ	$\langle p_{ij} \rangle$: $\bar{\varepsilon}_{inter}$ (± STD)	OUTLIERS
NY	359	0.008 ± 0.09	0.03
SI	23	0.007 ± 0.09	0.09
SF	28	0.038 ± 0.18	0.07
VI	31	0.017 ± 0.13	0.04

Table 8.1: **Variability of node-pair statistics over time.** Average relative number of trips between intersections $\langle p_{ij} \rangle_\tau$ computed using daily temporal snapshots of the dataset and averaged over the full network compared to final values at the end of the observation period (equation (8.2)). Time units with a total number of trips at least two standard deviations apart from the adjusted yearly mean have not been considered in the average to account for seasonal variations. For the pairs of intersections ij, only pairs with at least one non-zero appearance on the time slicing have been considered for the graph-average. The fraction of data with absolute relative error larger than two standard deviations is also reported as *Outliers*.

ble 8.1). The graph-average of ε is very close to zero and highly concentrated around this value for all the time windows considered[3].

Figure 8.1 shows the correlation between the relative error and the relative importance of occupied links. The fact that $\sum_{ij} \hat{p}_{ij}(\tau) = \sum_{ij} \langle p_{ij} \rangle_\tau = 1$, coupled with second order seasonality effects induces an uneven distribution of errors: An overestimation of some values in the collection $\{\langle p_{ij} \rangle_\tau\}$ will forcefully induce an underestimation in some other values of the collection. Despite this issue, we can clearly see that the vast majority of the mass of relative errors is concentrated around zero (see points in background for Figure 8.1). For the NY case, seasonality effects are washed out (and hence fluctuations are much smaller) since a complete year - $n_\tau = 365$ - is available for averaging, as opposed to roughly a month of data for the other datasets.

We thus observe a clear validation of a sort of *conservation law* implicit in our formulation of high sampling limit for Multi-Edge networks (see (3.18)). Now that we know that $\{p_{ij}\}$ statistics are (very) stable across time, we can similarly compute their higher order statistics by observing their relative fluctuations.

From Figure 8.2 we observe how the relative fluctuations of $\{p_{ij}\}$ for a wide range of values do conform to our expect theoretical values (8.1) (see background points). It can be however observed how for large values of $\{\langle p_{ij} \rangle_\tau\}$ a plateau is reached, most possibly caused by

[3] Days with *abnormal* number of trips further away than 2 standard deviations from the mean have not been considered for the slicing (< 5% of data in the worst case).

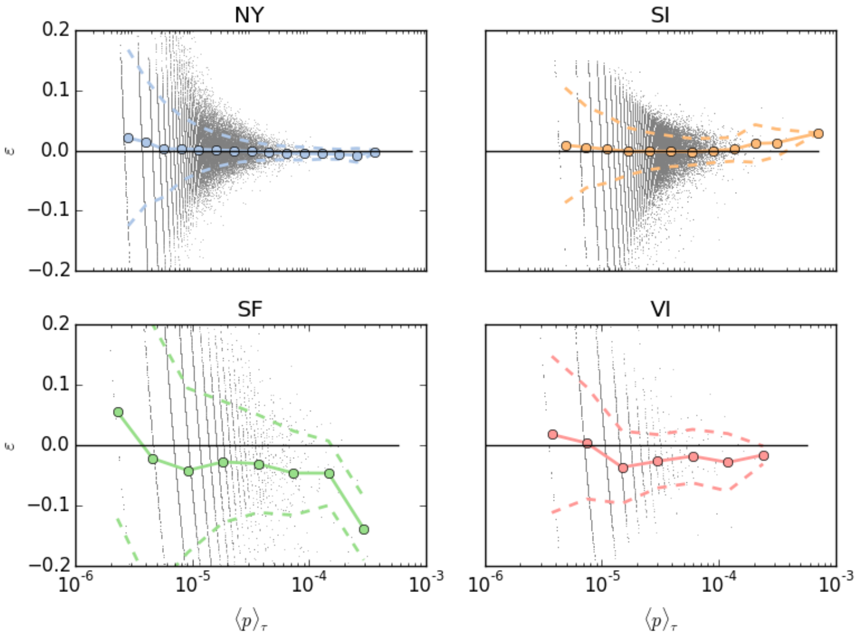

Figure 8.1: **The effect of sampling on intersection pair temporal stability.** Correlation between measured values of intersection pair $\langle p_{ij} \rangle_\tau$ averaged over daily slices and relative dispersion around the mean (8.2) for the aggregated data over the entire observation period. Error bars represent standard deviations on the log-binned data. Raw data is shown in the background. For visual clarity, NY panel only shows a random subsample of 1/100 of the original points.

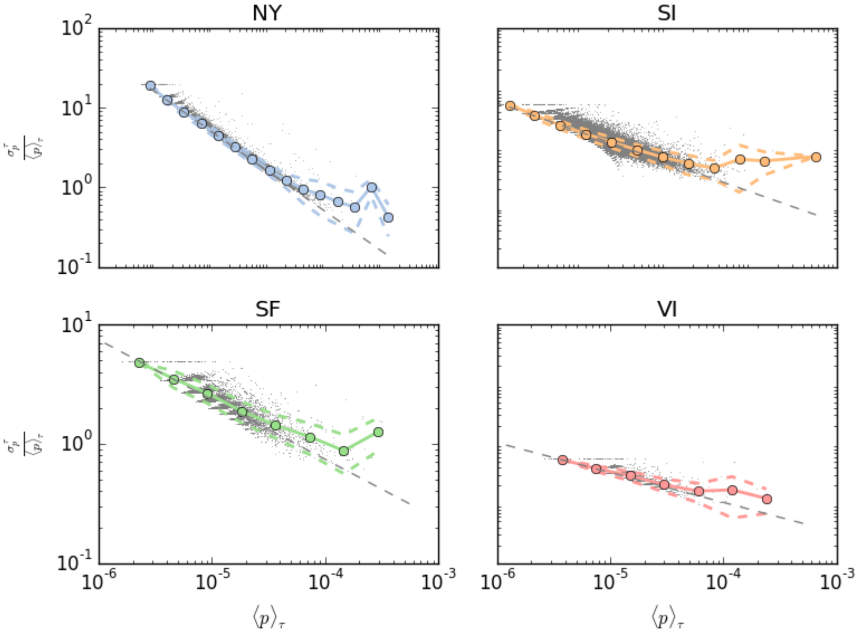

Figure 8.2: **Relative fluctuations of** $\{p_{ij}\}$ **over time.** Relative fluctuations of $\{p_{ij}\}$ as a function of the average values $\{\langle p_{ij} \rangle\}$ for the daily sliced values of all the datasets. Dashed lines represent theoretical predictions from (8.1). Error bars represent standard deviations on the log-binned data. Raw data is shown in the background. For visual clarity, NY panel only shows a random subsample of 1/100 of the original points.

time fluctuations which generate gaussian-like statistics on the most important node-pairs due to the Central Limit Theorem[4].

From the observations above, we can hence conclude that our datasets represented in an OD can indeed be well modelled using a maximum entropy Multi-Edge description.

The picture for an alternative scenario such as considering a Weighted case is numerically explored and found to be distinctively different in Section D.5.

8.2 topological patterns

Now that we have uncovered the relevant rescaling variable on our datasets, we proceed to analyze their main topological properties, bearing always in mind the need to compare our results with those obtained by the chosen MECM model.

Since our statistics are time-stamped, we can study the temporal evolution of their global topological variables in Figure 8.3. The relative number of nodes covered follows a fast saturation curve, with $> 60\%$ of the total nodes covered when only 10% of the observation time is considered in the worst case. This is not the case for the ac-

[4] For highly used nodes, the seasonal and weekly fluctuations become relevant and our analysis does not take this into account, since it considers all days as samples from the same process.

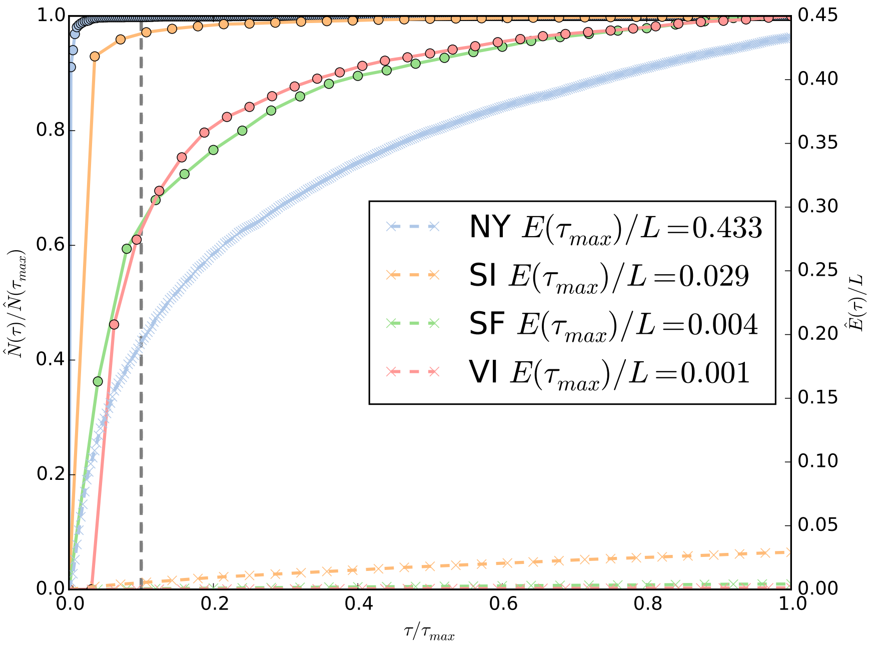

Figure 8.3: **Temporal evolution of number of distinct observed nodes $\hat{N}(\tau)$ and binary edges $\hat{E}(\tau)$.** Fraction of distinct observed nodes relative to the total number of nodes present in the accumulated data (circles) and distinct observed binary edges relative to total number of node-intersections (x-markers) as relative observation time τ/τ_{max} advances. A majority of the nodes is covered with very few days of accumulated data while the growth of observed binary edges is very slow and tends to saturate. A vertical line marking 10% of relative observation time has been added for visual clarity.

cumulated number of binary edges $\hat{E}$ which grows at a much slower pace. In fact, for the densest of our datasets (NY), at the end of the observation time we only have covered 45% of the total node-pairs $L = N^2$ which indicates a significantly skewed distribution of trips among binary edges (we have seen in contrast that the trip generation rate is almost constant in the preceding section, Figure 7.4).

In Section A.4, a detailed list of network metrics used in this thesis can be found, with explicit formulas and descriptions of their use and meaning.

Focusing now on the static picture of all trips aggregated by the end of the observation period τ_{max}, Figure 8.4 displays the (rescaled) node related properties of the generated ME networks with the empirical data. We have already seen that the strength distribution is highly skewed (Figure 7.3), and now we see that, at the node level, the empirical features do not diverge much from MECM predictions (dotted lines): The assortativity profile (s^w_{nn}) is almost flat and the empirical degrees are only slightly smaller than the predictions of the MECM. This difference causes greater discordance between empirical data and model predictions in the disparity profile. For small sampling, the overall differences are hard to notice (SF and VI) while for NY and SI they become obvious.

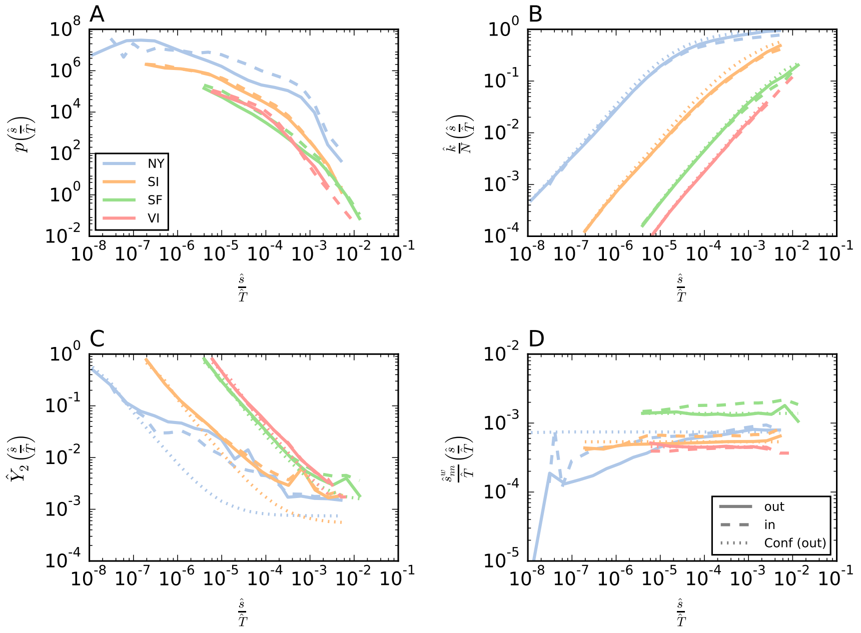

Figure 8.4: **Empirical taxi Multi-edge network node properties.** Rescaled strength distribution (A) and node related properties as function of strength [degrees (B), disparities (C) and average weighted neighbor strength (D)]. Results show averages over log-binned bins in the x axis. The properties of the raw data display close properties to those of a MECM averaged over $r = 10^3$ instances (dotted lines).

The resemblance between data and MECM impedes to clearly quantify differences between both using only log-log plots and thus a closer look is called for. To further explore at a finer scale the differences between empirical data and null model, Figure 8.5 displays the relative differences between node features, $\varepsilon_x = (\hat{x} - \langle x \rangle_{\mathrm{MECM}}) / \langle x \rangle_{\mathrm{MECM}}$ (x being strengths, degrees, disparities and weighted neighbor strengths).

Obviously the strengths are coincident[5] while the difference in degrees grows up to *only* the -30% in the worst case for medium strength nodes. Small strength nodes coincide with the configuration model due to the correlation between strength and degrees ($s = 1 \implies k = 1$) and large strength nodes are almost fully connected and thus also coincide with the model ($k \sim N$). Medium sized nodes have less connections than expected, hence concentrate more of their incoming-outgoing trips in these, but this cannot explain the relative differences observed in disparity[6] (much larger than those seen in the node de-

[5] The small differences for low-strength nodes are caused by biased and finite sampling (statistics are only computed for realizations where nodes have non-zero strength values) but this effect quickly vanishes in all datasets.

[6] Disparity values are difficult to interpret and compare to a null model where degrees are not fixed, because the disparity $Y_{2,i} \in [1/k_i, 1]$ is a bounded measure. We use them in the thesis as they are considered a standard quantity in weighted (usually

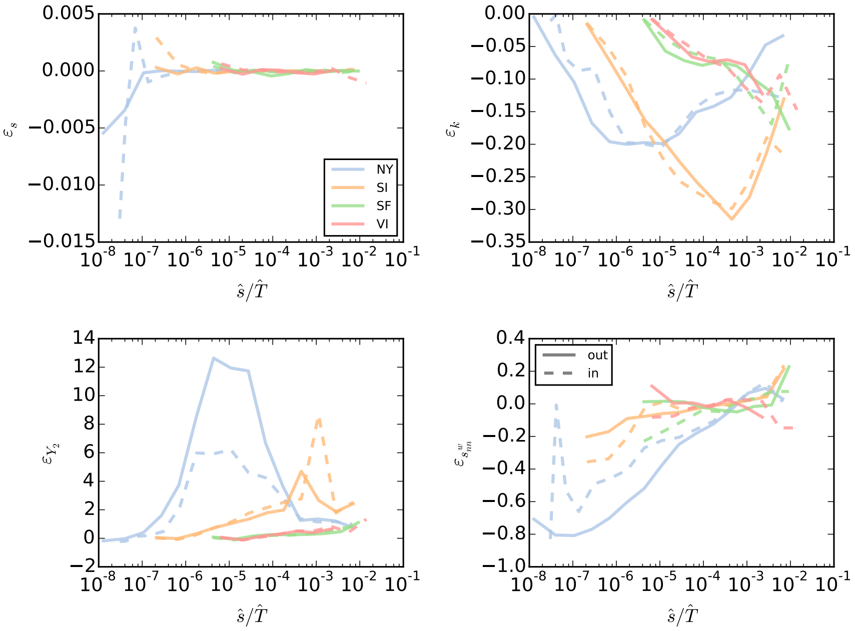

Figure 8.5: **Relative difference $\varepsilon = (\hat{x} - \langle x \rangle_{\mathbf{MECM}}) / \langle x \rangle_{\mathbf{MECM}}$ between empirical node properties of Figure 8.4 and configuration model predictions.** The relative errors for strengths (A) are non-existent while node degrees (B) are slightly smaller for the empirical dataset, as well as incoming node neighbor correlations (D). More relevant differences are detected for disparities (C).

grees): Such an hypothesis can be tested and is rejected by applying a model fixing both strengths and degrees, as done in detail in Appendix D. Finally, the assortativity profile is clearly ascending, indicating that nodes of small strength tend to connect with larger-than-expected weights among themselves.

Apparently, the directionality of the network does not seem to be important as trends in outgoing and incoming direction are qualitatively equal, however, taking a look at node asymmetry, we observe some differences. Relative differences between node incoming and outgoing strengths are encoded in the asymmetry coefficient:

$$\Delta_i^{\hat{s}} = \frac{\hat{s}_i^{\mathrm{out}} - \hat{s}_i^{\mathrm{in}}}{\hat{s}_i^{\mathrm{out}} + \hat{s}_i^{\mathrm{in}}} \qquad -1 < \Delta_i^{s} < 1. \tag{8.3}$$

Their graph-average across nodes, as well as the Spearman correlation coefficient among incoming and outgoing strengths is reported in Table 8.2. *Significant*[7] ranking correlation exists between the variables, but it is not perfect. Concerning the strength asymmetry, even

Spearman's [171] rank correlation coefficient is a measure of the association in rank *or importance among two sets of variables.*

sparse) networks literature. It is used to compare empirical weight allocations among nodes to a node-based null model that assumes uniform random weight allocation per link.

[7] In this case, the word significance is used in a wide sense, meaning that in all cases the coefficients display values larger than 0.8.

DATASET	SPEARMAN	$\overline{\Delta^s}$ ($\pm$STD)
NY	0.94	-0.25 ± 0.40
SI	0.89	-0.09 ± 0.41
SF	0.87	-0.28 ± 0.48
VI	0.84	-0.14 ± 0.48

Table 8.2: **Assessing directionality importance in the empirical Taxi networks.** Average graph node asymmetry and Spearman rank correlation coefficient comparing outgoing and incoming strength sequence pairs $(s^{out}, s^{in})_i$. p values correspond to alternative hypothesis of uncorrelated variables are in all cases $p \simeq 0$.

if a conservation rule applies to strengths, i. e. $T = \sum_j s_j^{out} = \sum_i s_i^{in}$, the graph averaged values are negative, indicating that the incoming-strength distribution is more heterogeneous (see Figure 7.3) than the outgoing one (hubs[8] accumulate a larger share of the total incoming traffic). Taxis tend to be booked for trajectories from centres of attraction towards scattered locations, hence nodes on average accumulate larger incoming than outgoing trips.

Figure 8.6 displays edge related properties where wider differences between configuration model and empirical data are seen: The occupation number distribution displays larger values in the tail which follows a slow decaying trend (for the cases of VI and SF with smaller sampling, *similar* to a power-law, although concavity appears for the NY and SI denser datasets).

We can study the rescaled average weight of links according to their incoming and outgoing node strengths in Figure 8.7.

$$\bar{p}(\hat{s}, \hat{s}') \propto \frac{\sum_{ij} \delta_{\hat{s}_i^{out}, s} \delta_{\hat{s}_j^{in}, s'} t_{ij}}{\sum_{ij} \delta_{\hat{s}_i^{out}, s} \delta_{\hat{s}_j^{in}, s'}}. \tag{8.4}$$

In such a plot, a configuration model prediction corresponds to the over-printed white lines ($\hat{s}^{out}\hat{s}^{in} = \text{Ctnt}$). We see that the apparent scaling of conditioned average occupation number as product of their source and destination node strengths is roughly linear.

At a finer level of detail, Figure 8.8, we show the quotient between expected values of the model and empirical data[9]. There, we can clearly observe that links among small nodes have distinctively larger occupation than expected (also among hubs) and also among same-strength nodes (values along the diagonal $\hat{s}^{out} \sim \hat{s}^{in}$).

So in a nutshell, we have an intricate structure where nodes can be roughly classified into three groups:

[8] Throughout this part of the thesis, *hub* is used to refer to nodes with large values of incoming or outgoing strength.

[9] Since the comparison is done with one run of the model, we set undetermined values $0/0 = 1$ and $0/K = \infty > 4$ for $K \neq 0$ throughout the thesis for this kind of plots.

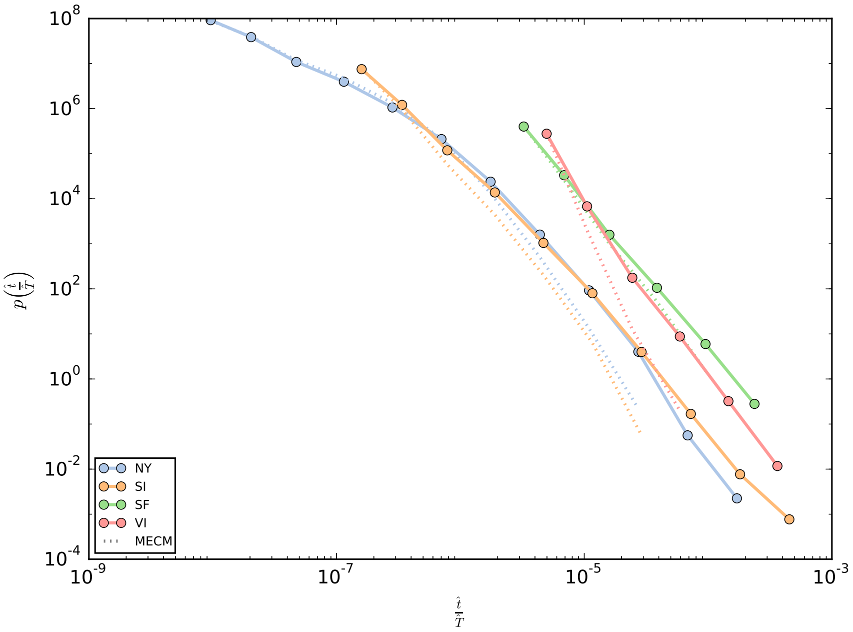

Figure 8.6: **Empirical taxi Multi-edge network occupation number distribution.** Configuration model dotted lines are computed over a single realization of the model.

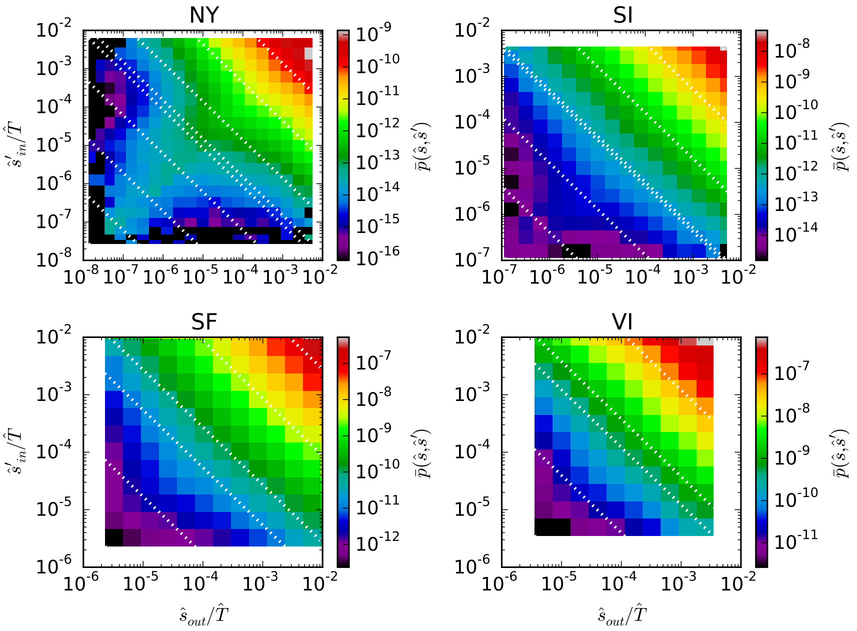

Figure 8.7: **Scaled average occupation number of links as function of their origin and destination node strengths.** Bidimensional plot showing (8.4) computed using log-binning. White lines correspond to MECM predictions of constant occupation number value.

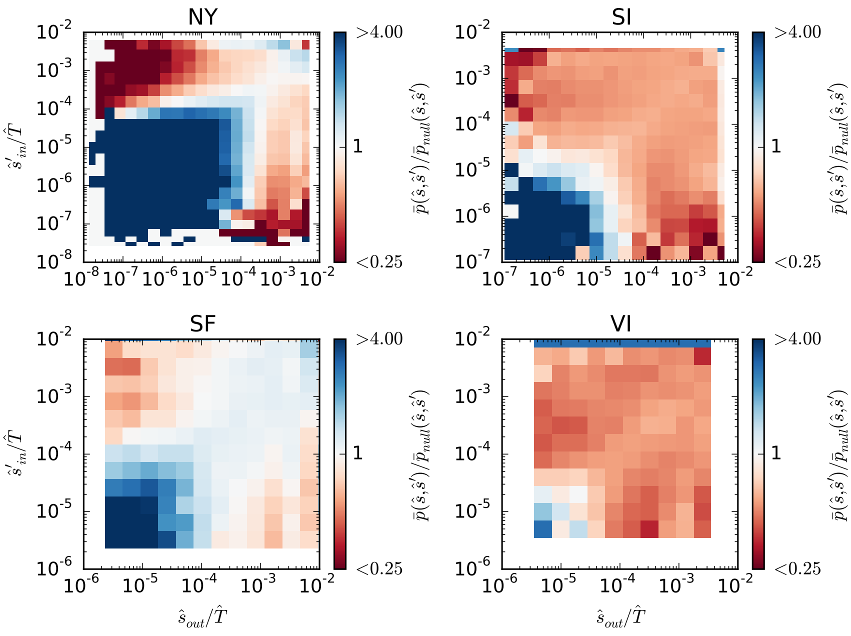

Figure 8.8: **Comparison between empirical data and MECM at edge level.** Relative scaled occupation number as function of starting and ending node strength comparing empirical data and MECM model over a single run. Both cases are normalized over the bins.

A. *Low sized nodes*[10]*:* Their naturally coincident degree and strength induces small differences with configuration model predictions in degree or disparity. Their connection pattern is noisy, since they connect to hubs but there are also many statistically (unexpected) trips between pairs of scattered locations that induce instabilities in their average neighbor strength values.

B. *Medium sized nodes*: These constitute a relevant fraction of nodes in the network and display interesting features. On the one hand, they are unbalanced favouring being destination of trips. On the other hand, they accumulate less binary edges than expected. Their binary edges thus *carry* over-expected traffic, directed mainly to nodes belonging to the same group (as seen in the assortativity profile and throughout the diagonal in Figure 8.8).

C. *Large sized nodes*: These nodes are connected to a large fraction of the system and hence their properties are averaged out in general. Their main difference with regards to a configuration model is the slightly larger values of occupation on the links connecting them, which cause larger values in the distribution

[10] During this part of the thesis, node-size, node-importance and node-strength will be used to refer to the same concept.

of existing occupation numbers. In general, they tend to produce more trips than they receive.

The previous list being made, we must bear in mind that the differences observed are not *extreme* and that hence the main zero-approximation features of this network are well explained by their skewed strength distributions, which can be attributed to exogenous factors such as population, job or housing density. However, additional factors add many more trips between some connections than there should be under random conditions. Hence, the distribution of node strengths across the city has a strong influence on the trip allocation, and needs to be taken into account when modelled, but needs extra ingredients to fully account for the observed pattern of connections.

In Section 10.2 a methodology is proposed to filter out *the influence of MECM features.*

8.3 SPATIAL PATTERNS

A final important analysis is related to space: We have observed the topological and temporal features of the networks under study, but we cannot forget that we deal with a network rooted in a metric space. The *first law of geography*, enunciated by W. Tobler [182], states roughly,

> Everything is related to everything else, but near things are more related than distant things.

We see that at very dense environments such as a city and with regards to mobility this is also true. Distance does have an effect on trip frequency as seen by the characteristic exponential decay detected in many transportation systems (also taxis). In Figure 8.9 we can observe how being the null model agnostic with respect to node distances, its probability to allocate longer trips is larger and its form smoother (ratio between average trip distance $\bar{d}^{\text{data}}/\bar{d}^{\text{MECM}}$ is 0.7 for NY, 0.6 for SI, 0.9 for SF and 0.7 for VI), hence, this feature of the empirical data can be marked as distinctive. In Figure 8.9, the internodal distance distribution for each geographical layout is also shown to emphasize the fact that even if the MECM model allocates trips at random, the geographical distribution of node strengths plays a very important role in shaping the obtained trip distance distribution[11].

From a network point of view, one can cluster the nodes into spatially cohesive regions. To do so, we have used a standard network

[11] Our findings partially confirm those in [115], where a combination of an exponential decrease of radial density of strength in nodes around local centers and a gravity-like model for traffic prediction with power law deterrence function (see Chapter 9) is assured to be the cause of the observed exponential trend in the distribution of trip lengths, which is obtained from Taxi displacements and other sources. Sadly enough, in that study the authors do not apply a distance agnostic value so it is hard to assess whether only the strength allocation could already produce the exponential decrease observed in trip distances.

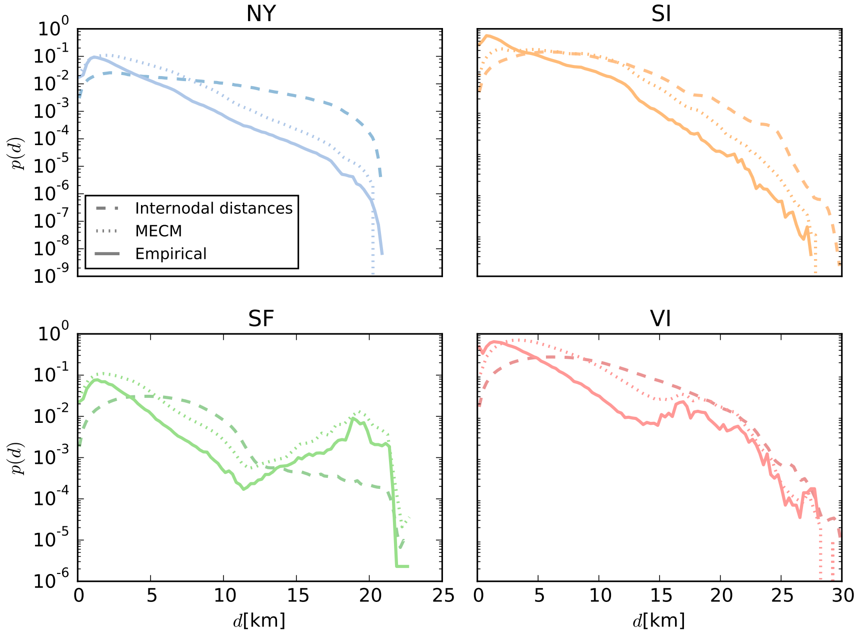

Figure 8.9: **The effect of cost on displacements.** Histogram showing the trip length distribution for the empirical datasets compared to MECM with linear binning (internodal distance distribution is also shown).

analysis tool that is useful for community detection in directed, non-binary graphs. A community is understood as a partition of labels $\{u_i \, \forall i = 1, N\}$ over a set of nodes that accumulate more intra-group connections than those that would be expected under a null model. A common way to find such a partition consists on optimizing an objective function known as the graph-modularity Q [127]:

$$Q = \frac{1}{\hat{T}} \sum_{ij} \left(\hat{t}_{ij} - \langle t_{ij} \rangle_{\text{null}} \right) \delta_{u_i u_j}. \tag{8.5}$$

Refer to [79] for a comprehensive review on community detection methods.

Several methods are available for this, in this case, I have used an implementation of the Louvain algorithm [39] openly available [184]. In order to assess the significance[12] of the obtained modularity values, we have simulated $r = 10$ samples of networks with the MECM and computed a histogram of their modularity and number of detected clusters $\mathcal{N}_c$ from which we have extracted z-scores for these magnitudes[13]. From the results displayed in Table 8.3, we see that even if in

[12] Absolute modularity values are uninformative about the community structure of the network, they must be compared to averaged realizations of the null model.

[13] The statistics of these random variables are approximately Gaussian. As a reminder, a z score for an observed variable $\hat{x}$ is computed as,

$$z = \frac{\hat{x} - \langle x \rangle}{\sigma_x}. \tag{8.6}$$

DATASET	$Q(z_Q)$	$\mathcal{N}_c\ (z_{\mathcal{N}_c})$
NY	0.007 (968)	9 (0.6)
SI	0.19 (502)	9 (−8.8)
SF	0.14 (115)	56 (−22)
VI	0.28 (104)	46 (−45)

Table 8.3: **Community structure empirical network scores.** Empirical modularity Q and total number of clusters $\mathcal{N}_c$ obtained from application of the Louvain modularity optimization method on the empirical datasets and corresponding z scores compared to configuration model averaged over $r = 10$ runs.

some cases the obtained values of modularity might be seen as *small*, they are always relevant[14]. Figure 8.10 shows a map of the cohesive regions detected by the algorithm.

Taking a closer look, we see that even if the effect of distance is patent, communities do not form totally cohesive regions mainly due to the effect of hubs. Since they are connected to a macroscopic fraction of the system, their influence combined with the topological resemblance between empirical networks and configuration model blurs the spatial form of the detected communities.

Such spatial coherence, however, supports the theory of city policentricity [**?**]. Even though city centres are undeniably the most busy areas in terms of mobility, economic activity, traffic and others, the influence of other areas has a strong influence in a variety of processes, among which, mobility. The city organizes itself in hierarchical levels that are intrinsically related with their spatial layout. In our case, the paradigmatic case is Singapore, since on the one hand it is an island and on the other their city areas are naturally delimited by the presence of a significant geographical obstacle (a mountain).

A deep discussion on the theory of policentricity and the study of the city from a geographical point of view lies outside of the scope of this thesis. A recommended read is [119] and references therein.

The spatial analysis shows how the mobility in the city is shaped by mainly two competing *forces*: The extremely heterogeneous topological (and spatial) distribution of activities (encoded in the node strengths) and the cost associated with trips, which favours the emergence of communities spanning spatially compact domains which is blurred by the influence of hubs (their spatial agnostic influence can also be *felt* in the anomalies to the exponential regime detected in trip lengths in Figure 8.9).

[14] It is important to note that community detection algorithms are spatially-agnostic and hence no information about node relative-coordinates is provided as input, wherefore a spatially coherent node partition can be considered as moderately surprising fact.

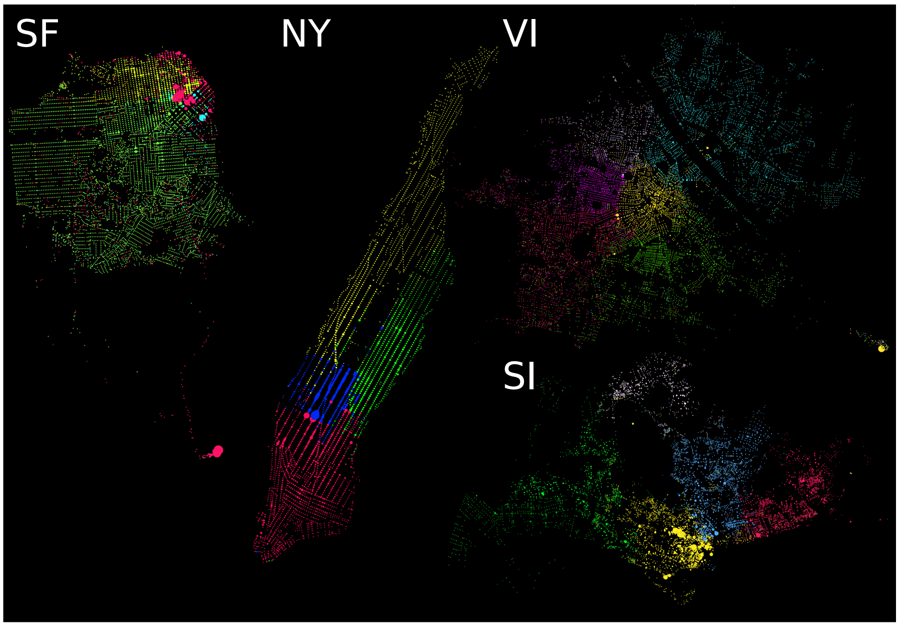

Figure 8.10: **Emerging spatially cohesive communities from modularity optimization.** Map displaying the emerging cohesive communities obtained from the modularity optimization (only stable clusters spanning more than 1% of the system are shown). Color encodes group partition while size is proportional to incoming node strength.

8.4 WRAPPING UP: MAIN FEATURES OF URBAN MOBILITY NETWORKS

In this chapter we have performed a comprehensive analysis on the Taxi empirical datasets represented as non-binary networks. Many common features to the four considered datasets have been detected, both at the spatial and topological level:

A. **Stable time statistics for node-pair sites:** We have confirmed that a Poisson description, rescaled by the total number of trips in a given time period, is compatible with the temporal statistics of trips among pairs of nodes present in the data. This is an evident signature that a maximum entropy modelling approach considering Multi-Edge structures with linear constraints as studied in Part ii constitutes an excellent framework for the study of these datasets.

B. **Topological patterns close to configuration model:** At the topological level, we have confirmed our hypothesis that the distribution of node importance, which is empirically detected by the strength of each node is the main driving force shaping the mobility in the city. A zero order approximation to model these networks is thus the MECM, which yields expectations for the main network observables within the same order of magnitude as those empirically observed. However, some differences have been pinpointed:

   - **Edge-level differences:** The main topological differences detected at the node-pair level are the tendency of hubs to accumulate larger number of trips in the links connecting them and also the statistically relevant presence of trips among scattered locations. Also, a significant tendency of nodes to connect to other nodes with similar strength is observed.
   - **Node-level differences:** The differences at edge level are seen here with a slight assortative profile (nodes with small strengths tend to connect with more trips than expected to nodes with equal or larger strengths). This in turn causes an overall decrease in the expected number of binary connections per node (degrees) which is specially important for nodes with medium size.

C. **Spatial effects of cost and hubs:** We have extracted communities from the network that conform spatially coherent regions, in sharp contrast with the null model. This is probably caused by the exponentially decaying shape of the trip length distribution, which is altered by the presence of trips directed towards large hubs such as airports lying at far distances from the city.

D. **Node unbalance:** The effect of the considered transportation mode (taxis) is patent in the unbalance in out and in strength detected at node level. Hubs tend to be origin of trips rather than destinations, while mid sized nodes tend to be destinations rather than origins. We hypothesize that the correlation at the spatial level between the locations of the latter places with other transportation means such as large bus, metro or train stations might be behind this fact, and thus needs to be taken into account on an overall analysis of city mobility from a multiplex perspective [81].

The qualitative and quantitative coincidence of features among the datasets show that indeed we can detect common features to diverse Taxi urban mobility datasets despite their different geographical layouts. With the insights obtained in this chapter, we will now proceed to examine ways into which such structures can be reproduced using models, as to be able to find mechanisms that explain the origin of their characteristics beyond simple city structure.

Finally, since we have seen that the influence of strengths overwhelms the secondary order effects that also have an effect on mobility, in Section 10.2 we propose a methodology to filter out their contribution. This will also allow for effective network visualization applications. It is worth noting also that the techniques used here and some of the conclusions drawn from them are coincident with a large study [118] based on call detail records (CDR) over a large number of spanish cities. In this case, trips are clustered into four groups (which may be identified with the elements in each of the quadrants in Figure 8.8), one of which are the trips between small strength locations. Such trips, termed "random", are shown to have larger expression than under a random model (using a variant of the MECM model), and to have a relative weight $\hat{t}_r/\hat{T}$ increasing with population (sampling in our case).

9 URBAN MOBILITY MODELLING

One important idea is that science is a means whereby learning is achieved, not by mere theoretical speculation on the one hand, nor by the undirected accumulation of practical facts on the other, but rather by a motivated iteration between theory and practice.

— George E. P. Box [43]

Up to the present moment we have evaluated and surmounted the main technical issues to be faced for the analysis of OD matrices in urban environments. Using these insights, we have uncovered common temporal, topological and spatial patterns in our datasets. We have argued that these common features are best made apparent when studying the datasets under the common frame of Multi-Edge networks. Our ultimate objective is to profit from the obtained insights and develop practical applications related to urban mobility.

One of these applications is related to forecasting traffic generation in cities. Several *predictive* models exists for traffic production[1] forecasting at general scales, and it is time to test whether these models can perform also well in the city environment.

The objective of this chapter is to review the existing models proposed in the literature on human mobility and assess the challenges that the adaptation of these models to the urban environment pose. In order to begin, however, we have already seen that the majority of statistics encountered are highly skewed and non-gaussian, so prior to any discussion we need to develop indicators in order to quantify to which extent any proposed model aiming to *predict* trip generation (an OD for a city) performs adequately when compared to an empirical sample.

In this chapter, I start by developing a collection of indicators to quantify the quality of any proposed model of trip generation. To do so, we shall use a combination of information theoretical related metrics, topological metrics and matrix similarity indicators. Then, using the tools developed, I perform a critical review of the existing models to finally assess which performs better in the current taxi case under study. We will study three broad classes of models: Heuristic, Opportunistic and Maximum entropy based models.

[1] Traffic generation deals with the prediction of how many people travel from two given places, while traffic routing or allocation relates to which trajectories this trips take. The latter case lies outside of the scope of the present thesis.

9.1 MODEL VALIDATION

The first thing to do when facing the problem of model validation is to define what we mean by it. In the present context, for any given model, our aim is to detect its shortcomings and its successes with regards to reproduction and estimation of empirical data. Since, as we have seen, our records -even if studied from an aggregated, static, point of view- display a highly intricate topological and spatial structure, assessing model quality must be a holistic process.

The inherent heterogeneous nature of the studied data forces us to take a multi-angled perspective with regards to model validation. It is highly unlikely that a single general metric will be able to capture the effectiveness of any model to reproduce real data, and so that is why we will have to use a diversity of indicators. A natural question emerges: What will be the best model, that which *gets right* the majority of the data (small nodes-trips) or the most influential (large nodes-trips, a minority)?

The answer to this question is that neither will be: The aim of model validation is to detect the strengths and weaknesses of each considered trip generation mechanism, which should be the ultimate objective with regards to explanatory modelling. Hence, we shall use a cocktail of indicators taking different perspectives to explore the likeness between data and predictions.

We will focus on information theory related indicators (such as entropies), topological ones studying the network structure of the predicted ODs and finally similarity metrics related to matrix comparison. For all these indicators, we will also need to consider their sensibility to sampling, which as we have seen is an important dimension to be taken into account.

9.1.1 *Topological metrics*

We have made extensive use of network metrics in the previous Chapter 8. Those are fully described in the Appendix (Section A.4) and place the emphasis on the pattern of connections among nodes. Whenever possible, the assessment of differences between model and data should be done by analyzing relative differences between indicators at node level, since most of these metrics span multiple orders of magnitude. However, log-log plots should be used with care as their limited resolution provides limited information for models close to real data, as seen in Section 8.2.

At the node level, we shall focus the analysis on the predicted binary degree per node and assortativity profile[2] as we have been doing

[2] We have seen that disparity values Y_2 are highly dependent on the binary structure of the network, and hence not very informative for non-sparse networks, but we report them here for completeness.

in previous chapters. At the edge level, we will put the emphasis on the average occupation of binary links as a function of the strength of their source and destination nodes.

Finally, since we have taken a Multi-Edge network description, all the indicators are scaled (i. e. lie in the range $[0, 1]$) as to limit the influence of sampling on them.

9.1.2 *Information metrics*

Any complete model should be able to provide not only average values for expected trips between locations $\vec{T}$, but also the probabilistic form from which those averages are obtained, $P(\vec{T})$. Ideally, such a probability should be derived from first principles (*a generating mechanism*), allowing to test with empirical data whether the proposed generating mechanism is correct or not.

If a probability of occurrence for a given instance of a network is given, then we can define an ensemble of networks for this model (even if it is not a maximum entropy one), an we can extract some information theoretical related measured from it.

From $P(\vec{T})$, one can define likelihood and entropy measures. One can measure the surprise per event $-\mathcal{L}(\hat{T}|\text{model})/\hat{T} = -\hat{T}^{-1} \ln P(\hat{\vec{T}}|\text{model})$ of the empirical OD, $\hat{\vec{T}}$ according to the model defined by $P(\vec{T}|\text{model})$. Furthermore, given that we know that the statistics of of the observed networks should behave as Multi-Edge network structures, we can compute the maximum likelihood estimation of the difference per event between the Γ entropy of the model and the empirical data based on a single run,

$$\frac{\Delta \hat{S}^{\Gamma}}{\hat{T}} = \sum_{ij} \left(\frac{\hat{t}_{ij}^{\text{data}}}{\hat{T}} \ln \frac{\hat{t}_{ij}^{\text{data}}}{\hat{T}} - \frac{\hat{t}_{ij}^{\text{model}}}{\hat{T}} \ln \frac{\hat{t}_{ij}^{\text{model}}}{\hat{T}} \right). \tag{9.1}$$

Ideally, in the previous expression we should use $\langle t_{ij} \rangle$ instead of $\hat{t}_{ij}^{\text{model}}$ for any given model, however we will use this approximation, given that some of the models we will examine do not yield analytical expectations for $\{\langle t_{ij} \rangle\}$. This indicator can be interpreted as the average amount of information (bits)[3] per event that is needed to differentiate any instance from the proposed model from the empirical data.

In some cases, for simplicity, we will also present the relative loglikelihood of empirical data to belong to the proposed model relative to the loglikelihood of the empirical data considering Poisson statistics for the pairs of nodes ij with average equal to the empirical one

Other examples of information theoretic metrics have been used practically in the network literature, see for instance [35].

[3] The base of the logarithm defines the scale of the units of information used. For instance, using a base 2 logarithm one would be using bits. We will use the natural logarithm for simplicity in the calculations.

$\hat{t}_{ij}$, in order to simplify model comparison. The baseline for this corresponds to,

$$\mathcal{L}_0 = \ln P(\{\hat{t}_{ij}\} \mid \{\langle t_{ij} \rangle\}) = \sum_{ij} \ln \left(e^{-\langle t_{ij} \rangle} \frac{\langle t_{ij} \rangle^{\hat{t}_{ij}}}{\hat{t}_{ij}!} \right) \tag{9.2}$$

with $\langle t_{ij} \rangle = \hat{t}_{ij} \, \forall \, ij$. Incompatible Loglikelihood values will not reported in tables (such as cases where $\langle t_{ij} \rangle = 0 < \hat{t}_{ij}$ or $|\hat{t}_{ij} - \langle t_{ij} \rangle| \gg 0$).

Being the statistics of $\{\hat{t}_{ij}\}$ highly heterogeneous (see Figure 8.6-A), the only shortcoming of loglikelihood measures is that they place larger emphasis on the tails of the distribution of occupation numbers, hence models predicting larger values for the maximum weights are favoured over others, even if the latter ones my capture better the overall network structure[4].

9.1.3 *Similarity metrics*

To compare matrix structures, we will use two different indicators. A global indicator capturing the overall similarity between expected and empirical networks and the correlation between recorded and predicted individual occupation numbers.

A recently used global indicator is the Common Part of Commuters (CPC), introduced in [113] based on a similar index developed in biological studies (the Sorensen-Dice indicator [67, 170]).

$$\mathrm{CPC}_{\mathrm{sample}} = \frac{2 \sum_{ij \mid \hat{t}^{\mathrm{data}}_{ij} > 0} \min(\hat{t}^{\mathrm{data}}_{ij}, \hat{t}^{\mathrm{model}}_{ij})}{\sum_{ij} \hat{t}^{\mathrm{data}}_{ij} + \sum_{ij \mid \hat{t}^{\mathrm{data}}_{ij} > 0} \hat{t}^{\mathrm{model}}_{ij}}. \tag{9.3}$$

The different versions of this indicator have values in the range $[0, 1]$, where $\mathrm{CPC} = 1$ indicates total coincidence between data and model and $\mathrm{CPC} = 0$ total disagreement. However, for sparse data sets with a skewed distribution of $\{\hat{t}_{ij}\}$ values, equation (9.3) may return values excessively lower than 1, even for models very close to reality. To mitigate the effects that finite sampling have on the datasets, in some cases (those where analytical expectations of expected values can be obtained) we will use a slightly modified version of the indicator:

$$\mathrm{CPC} = \frac{2 \sum_{ij \mid t^{\mathrm{data}}_{ij} > 0} \min(\hat{t}^{\mathrm{data}}_{ij}, \langle t^{\mathrm{model}}_{ij} \rangle)}{\sum_{ij} \hat{t}^{\mathrm{data}}_{ij} + \sum_{ij \mid t^{\mathrm{data}}_{ij} > 0} \langle t^{\mathrm{model}}_{ij} \rangle}. \tag{9.4}$$

Additionally, plots comparing the pairs $(\hat{t}^{\mathrm{data}}_{ij}, \hat{t}^{\mathrm{model}}_{ij})$ have been popularized after its introduction in [164]. Since the visual plots on a

[4] Albeit not shown, this causes for instance the configuration model to display greater likelihood for the Weighted case than the Multi-Edge one in the comparison done in Section 6.4, even if the latter captures much better the topological features of the network.

f	$\langle \mathrm{CPC}_{\mathrm{SAMPLE}} \rangle$	CPC	R^2	$\mathcal{L}(\hat{\hat{T}})/\hat{T}$	$\Delta \hat{S}^{\Gamma}/\hat{T}$
0.0001	0.63	0.996	-108.45	1.003	-0.57
0.001	0.65	1.000	-7.24	0.946	-0.53
0.01	0.71	1.000	0.32	0.726	-0.37
0.1	0.84	1.000	0.94	0.332	-0.13
1	0.94	1.000	1.00	0.090	-0.03

Table 9.1: **Evolution of global model performance indicators with sampling for the NY dataset.** f stands for the fraction of uniform and random subsampling on original dataset. Average values performed over $r = 1000$ instances. Standard deviations in all cases smaller than $1/1000$. R^2 and $\Delta \hat{S}^{\Gamma}/\hat{T}$ computed using $\hat{t}_{ij}^{\mathrm{data}} = \hat{t}_{ij}^{f}$ and $\hat{t}_{ij}^{\mathrm{model}}$ obtained from a single run of the Poisson model (see (9.2)).

log-log plot may be misleading (we will always plot empirical data against a single run of any given model), we also report the coefficient of determination R^2 in all tables , based on the comparison between real data and average values of the model on the existing edges, assuming an identity relation[5] $\hat{t}_{ij}^{\mathrm{data}} = \hat{t}_{ij}^{\mathrm{model}}$.

$$R^2 = 1 - \frac{\sum_{ij|\hat{t}_{ij}^{\mathrm{data}}>0} \left(\hat{t}_{ij}^{\mathrm{model}} - \hat{t}_{ij}^{\mathrm{data}} \right)^2}{\sum_{ij|\hat{t}_{ij}^{\mathrm{data}}>0} \left(\hat{t}_{ij}^{\mathrm{data}} - \overline{\hat{t}_{ij}^{\mathrm{data}}} \right)^2}. \tag{9.5}$$

9.1.4 *The effect of sampling*

As has been already pointed out, for distributions displaying slow-decaying tails, total available sampling on empirical data is expected to have an important influence on the observed metrics. To test this hypothesis, we have proceeded to subsample the NY dataset to different levels fractions f of the original data to see the evolution of the proposed indicators as less and less data is considered. To this end, we proceed to generate instances of a Multi-Edge network with Poisson node-pair statistics $\langle t_{ij} \rangle$ equal to the empirical measured ones $\hat{t}_{ij}^{f}$ for each level of subsampling f and compute the proposed indicators. Results are displayed in table Table 9.1.

We observe how the CPC indicator computed over a single instance displays important variations with sampling. The effects of sampling are also patent on the coefficient of determination R^2 and for the event specific proposed indicators. Due to this dependence, all results comparing model to reality will be shown related to the ones obtained

[5] We use this indicator as a general measure of quality of the fit between model and data, yet, it has many limitations due to the heterogeneity of the underlying data, as can be seen by its negative values which indicate bad agreement between both.

by comparing empirical data with one run of a Poisson model with $\left\{\langle t_{ij}\rangle = \hat{t}_{ij}^{\text{data}}\right\}$, which will enable us to assess to which extent any indicator is *far* or *close* to the empirical data. Note, however, that in all cases for finite sampling the observed realizations are less informative than the ensemble entropy ($\Delta\hat{S}^{\Gamma}/\hat{T} < 0$) because the observation will be sparser in number of binary links on average[6].

9.2 CRITICAL REVIEW OF EXISTING MODELS

Modelling is a verb that is used to identify two similar yet different processes in a variety of fields. In some cases it refers to better understand an observed phenomenon by postulating a (microscopic) mechanism by which such a macroscopic observable may be reproduced. In others, it refers to a framework to be applied in practical situations, i. e. to generate estimates of a given (unknown) quantity of interest with the highest possible precision, without explicit explanatory primary objective. From a historical scientific point of view, normally the latter view tends to integrate elements of the first view in an iterative process that ends up conforming a complete theory.

Trouble is, specially with the advent of the Big Data era and its focus on empirism, that this symbiotic cycle seems to be slightly disrupted. More often than not models are considered to *predict the future* or to *explain* phenomena, when, if confronted to an objective and critical analysis they partially fail at either of them.

Whenever quantifying how good a model is (compared to another) several aspects need to be taken into account:

A. **Equal grounds comparison:** Models need to be compared on equal grounds. Step zero thus is to consider the same number of trips per model, and, if possible, the same number of constraints.

B. **Input importance, not parameters:** For any given model, the importance does not lie on the number of parameters it contains, but rather on the relation input/output: Fitting parameters is a technical issue, but the important matter here refers as how much we feed the model externally and how good the outcomes of it are.

C. **Feasibility of parameter fit:** Of course, models need to have an application. Hence, if dependent on a set of parameters, these need to be able to be calculated/estimated from real data with adequate precision and in a non-arbitrary fashion.

[6] Note that $\hat{E} - \langle E\rangle = \sum_{ij}(\Theta(\hat{t}_{ij}) - 1 + e^{-\langle t_{ij}\rangle}) = \sum_{ij|\hat{t}_{ij}>0} e^{-\hat{t}_{ij}} > 0$, even more so if the distribution of $\{\hat{t}_{ij}\}$ is heterogeneous and favours small values of occupation numbers, as is the case with real data.

d. **Normalization and skewness:** We have seen that human mobility processes display highly heterogeneous structures, and fat tailed statistics are ubiquitous to data describing it. Hence, specially care should be taken with regards to data normalization across different datasets (any model should provide insights about this) and influence of finite size effects. Also, some models might be better than others to reproduce a given aspect of the data, but fail in other aspects: A balanced view concerning as many angles of view as possible is beneficial for model assessment.

e. **Large sampling limit:** Finally, any model should yield predictions valid in a certain range of sampling, and also for the level of data-density at which fluctuations do not matter anymore (a sort of high sampling limit).

Based on the above criteria, we review the two majoritarian approaches used in mobility modelling literature. To this end, we will discuss the general approach behind each one of them and exemplify them presenting and analyzing one (of the many) model versions belonging to each type.

9.2.1 *Empirical models*

Historically, the first approaches to model human mobility were performed in economics and sociology based on empirical observations. Those pointed out that, as we have been stressing through this thesis, perceived *cost* and *importance* are two key elements that affect the number of trips between two given locations. Mainly two ways of associating cost to distance and to *number of lost opportunities* have been proposed.

One constitutes what is known as the family of gravity models of transportation while the other has given rise to rank based models. In either case, the problem with these models is that in general they are not expressed in mathematical terms suitable for analytical treatment. Hence, no probabilities can be associated to each model $P(\vec{T})$ and no entropy, surprise nor insights about high sampling limits can be obtained. This is so because more often than not, those models are based on analogies but not in complete theories developed from first principles and hence those can be interpreted as interesting starting points (they do seem to represent real data acceptably) but should not be considered as satisfactory models.

9.2.1.1 *The family of gravity models of transportation*

A first step towards modelling mobility was based on an analogy with a physical law which gave rise to the term *gravity law of transportation* proposed by Zipf in [198]. This model is based on assuming that each location has a mass M_i (which can be related to population, job availability, interest...) and that the number of trips between two locations can be approximated by:

$$\langle t_{ij} \rangle = K M_i M_j f(d_{ij}). \tag{9.6}$$

Where K is a normalizing constant ensuring that $\sum_{ij} \langle t_{ij} \rangle = \hat{T}$ and $f(d_{ij})$ is an arbitrary (decreasing) function of the cost or distance, known as deterrence function. Its most common version uses a power law $f(d_{ij}) = d_{ij}^{-\gamma}$ but also exponential versions are present in the literature. In the urban context, we will only consider exponential versions of the deterrence function for any model, because there are self-loops in our empirical data (and $f(d_{ij}) = d_{ij}^{-\alpha}$ diverges at the origin) and because the typically dense environment of the city does not allow for representative distances to span multiple orders of magnitude, hence the effect of a power law relation is not expected to be important.

For the attractivity of each location, versions ranging from power law dependency on population $M_i = P_i^{\beta}$, node strength $M_i = (\hat{s}_i^{out,in})^{\beta}$ and others have been proposed. All these parameters are fitted by multilinear (logarithmic) regression of the observed values $\{\hat{t}_{ij}\}$.

For a very complete review, discussion and analysis on the gravity laws of transportation see [70].

This approach is extremely popular in a variety of fields and has become mainstream due to its simplicity [99, 107, 100, 87]. While it is not based on any model, hypothesis or mathematical formulation, just a loose analogy, its advocates defend that its "overwhelming empirical evidence of success" validates the proposition. However, it is a model very prone to overfitting through a circular process. It is designed by construction to optimize the regression $\{\langle t_{ij} \rangle\}$ vs $\{\hat{t}_{ij}\}$, yet it is validated using again the same regression. In the end, its based on optimizing a fit of a single (increasing) function based on 3 parameters. Such a problem is common to other approaches depending on parameters that need to be fitted (for instance [101]).

Precisely because we already know that distance and importance influences mobility, the effectiveness of the fitting method comes at no surprise. Due to this fitting procedure, different exponents, dependencies and forms for the deterrence function are observed for different geographical layouts. Furthermore, this method is highly dependent on sampling: The logarithmic multilinear regression only takes into account existing edges and only those that have values larger than unity (and as we have seen, those with value 1 are majority due to their skewed distribution and for sparse cases the majority of inter-node links are not occupied and hence are not considered in the fit). It is an approach that takes as input 2N inputs relating node strength and importance and 3 additional parameters. For the reasons provided above, we will not consider this approach in the subsequent parts, since in our opinion, it is more a fitting procedure than a complete model able to provide insights. Also, using a relation

In Section D.5 a multilinear regression analysis in the spirit of the gravity models of transportation is performed for completeness, and its shortcoming made patent.

for attractivity of the type $M_i = \hat{s}_i^{\beta}$ (in the incoming or outgoing direction) leads to inconsistencies when considering different sampling and $\beta \neq 1$, because either the incoming rescaled strength $\hat{p}_{s_i^{in}} = \hat{s}_i^{in}/\hat{T}$ or the outgoing one are not conserved, and hence $\hat{p}_{ij}$ will change, in clear contradiction with our empirical observations, see Section 8.1.

9.2.1.2 *Other empirical models*

More recently, other types of models have also been proposed from the field of computer science. A prominent example is what I call *Sequential gravity model* [114]. This model allocates trips in a configuration-like fashion stochastically according to probabilities that are updated at each step,

$$p_{i\to j}(t) = \frac{s_j^{in}(t)f(d_{ij})}{\sum_j s_j^{in}(t)f(d_{ij})} \tag{9.7}$$

being $s_j^{in}(t)$ the incoming strength of location j at the time of allocation of the trip t. For the deterrence function an exponential form $f(d_{ij}) = e^{-\gamma d_{ij}}$ has been proposed with an exponent not fitted by empirical data but obtained from a *universal* function depending on the area under study. This is an interesting feature, since on the one hand avoids some of the problems caused by the modifiable area problem (MAUP, see Section 7.2) while on the other hand hints at the existence of some common dependence on distance across (very) different geographic areas (the model is validated using an extensive dataset on regional human mobility).

The problem of this model is that while it has a mechanistic explanation, it requires total knowledge by the travellers of the current state of the system: each user emerging from a given location takes a decision based on the actual state of allocation of trips. Moreover, while the algorithmic description is an interesting step, developing mathematical formulations is difficult due to the history dependent nature of the stochastic process involved in the trip allocation. Finally, problems may arise in its algorithmic implementation when the number of trips is extremely large (large sampling limit) since the naive complexity of the algorithm is $\mathcal{O}(\hat{T}N)$ because the probabilities need to be recomputed at each timestep, and thus averaging over instances becomes a laborious task. This model uses 2N inputs (node strengths) plus an additional parameter γ.

A final empirical model worth mentioning, which is a good representative of the opportunistic approach in cities (see below) is that proposed in [131]. Its authors propose that the probability to move from two given locations i, j reads,

$$p_{i\to j} \propto (\text{rank}_i(j))^{-\alpha} \qquad \text{rank}_i(j) = |\{w : d(i,w) < d(i,j)\}|. \tag{9.8}$$

It is a model where the probability of transition between places depends on the density of points of interest or destinations (POIs) between them measured as a rank, not on distance. In their interesting study using foursquare check-in data, the authors show that distance alone is not a good indicator of mobility in dense environments such as the urban habitat because observed trip distributions differ when observed in different cities (as we have seen in previous chapters). Their measurements show that the value of the rank exponent $\alpha = 0.84$ is highly stable across all the cities studied, but no particular justification is given for this. This finding again suggests a universality in human mobility behaviour across cities (which differ in size and form). The novelty of this model is that its formulation is continuous across space (no network aggregation is needed avoiding the MAUP). Again this model uses 2N inputs and one parameter and the version provided by its authors has complexity $\mathcal{O}(\hat{T})$ albeit it can be speeded using a multinomial node based approach to $\mathcal{O}(L)$ (in a similar way as the Radiation model, see below) to avoid problems related to high sampling limits. However, precisely the fact that the model formulation is continuous leads to inconsistencies in the large sampling limit (in such a limit, one expects that the density of points between locations diverges, making some kind of aggregation necessary and hence calling for the introduction of weighting in the computation of ranks).

9.2.2 *Analytical models*

The insights provided by the aforementioned empirical models inspired some authors to propose analytically tractable models taking into account both approaches. In a broad manner, those can be classified in entropic models and opportunistic ones, albeit in some cases applying suitable transforms, both models can be related.

9.2.2.1 *Entropic models*

What we call *entropic models* are those related to conservation laws that can be circumscribed in an ensemble non-binary description of a mobility process in the framework provided in the first part of this thesis. As we have seen, an appropriate approach for human mobility is the one provided by Multi-Edge structures with linear constraints. For the particular case where strengths and average trip distance is fixed, we reach an expression reminiscent of the previously introduced *gravity laws*,

$$\langle t_{ij} \rangle = x_i y_j e^{-\gamma d_{ij}}. \tag{9.9}$$

This was first developed by A.G. Wilson [190]. One may be tempted to relate this model with the previous gravity models but their differences are noteworthy: In this case, a non arbitrary way of obtaining $\{x_i\}, \{y_j\}$ and γ, which is not based on a fit on empirical data, can be prescribed. Also, a particular form for the probabilities associated to each entry t_{ij} is provided and discussion on sampling is also possible. It uses hence $2N+1$ parameters that are obtained from $2N+1$ inputs by likelihood maximization and its complexity is $\mathcal{O}(L)$.

For a complete critique of the Wilson entropic model see [150].

Some authors argue that its main weakness is that it relies on a general maximization rule and not a mechanistic description at the agent level. Why should people be interested in maximizing entropy? While this is certainly true, this is not a model aimed at explaining human psychology (highly diverse, complicated and complex) but rather a modelling methodology that balances predictive power with analytical tractability, allowing to derive conclusions from clearly identifiable constraints (that of fixed average cost per trip). Additionally, it allows to increasingly introduce information (constraints) to the model (as we will see in Section 10.1). It only assumes fixed allocation of importance in the places of the city, fixed average trip distance and otherwise total randomness expressed mathematically as maximization of entropies.

It is obvious that any redefinition of the perceived cost between locations d_{ij} can lead to a huge variety of deterrence functions $f(d_{ij})$, Wilson himself showed how to recover Stouffer's opportunistic model (see below) using this framework. We have furthermore shown that the statistics of t_{ij} in our empirical observation do behave according to this model, hence to follow the path opened by entropic model, we need to focus on uncovering the particular relation between strengths and importance in a city (job, housing or POI density for instance) and a good predictor for perceived cost[7]. Finally, it is patent that even if for each city the indicator γ may differ, if the model is successful at reproducing real data, it means that the starting hypothesis are likely to be true (fixed cost per trip), hence no contradiction may be found in disparity of Lagrange multipliers values across different geographical layouts [26]. Also, even if it is not a model defined in continuous space (it requires binning), since the computation of parameters is performed from first principles, its sensibility to the MAUP is small (the solutions of the Lagrange multipliers might change with binning, but the starting hypothesis remains unchanged).

In relation to the criticism expressed against the model from economic fields and spatial geography [150], for me, the problem of the model lies not on the methodology itself, but on the conclusions Wilson and others draw from it. One must resist the temptation to disclose a universal relation between $(x_i, \hat{s}_i^{out})$ and $(y_j, \hat{s}^{in})$. This would be misleading since each value x_i (y_j) depends not only on the strength of the node $i(j)$, but also on the geographical layout of the network (encoded in the cost matrix $\vec{D}$), the average trip cost $\bar{\hat{d}} = \hat{D}/\hat{T}$ and the total incoming and outgoing strength sequence. This is so because all

[7] Different metrics can be used such as topological network distances [144], Manhattan, time to destination [197] and others. However, the cost does not need to be necessarily a formal metric in the strict mathematical sense.

the values are obtained from an overall maximization, and thus care should be taken when trying to interpret parameters such as γ using analogies like the relation between temperature and Boltzmann's β in classical statistical physics. γ is loosely related to the average cost of a trip, and indeed for two ODs $\hat{\vec{T}}_1$ and $\hat{\vec{T}}_2$ rooted in the same geographical layout with the same strength sequence but with different $\bar{\hat{d}}_1 > \bar{\hat{d}}_2$, we expect $\gamma_1 < \gamma_2$. But this might not necessarily be true if the strength sequences for the two networks (or their geographical layout) are different. Conversely, for a given OD, a change of cost of trips, would incur in a change of the number of trips among locations, which would in turn change the strength sequence considered and conversely would alter all of the considered Lagrange multipliers. This means, once again, that interpretations of the Lagrange multipliers are hard, and hence theories "not far from equilibrium" analyzing the effects that change of travel costs may have are complicated to do analytically, as there is no guarantee that the Lagrange multipliers change smoothly (and hence a constant approximation is correct) with the constraints, because they are in no ways related to an equilibrium situation, as would be the case for physical systems.

9.2.2.2 *Opportunistic models*

Opportunistic models are related to the idea proposed by Stouffer that humans move between locations depending on the density of opportunities between them, not on distance alone. Stouffer enunciated mathematically this idea in his famous paper presenting the Intervening Opportunities model [177]. In our notation[8], his model states:

$$\begin{aligned} \frac{\sum_{j|d_{ij}\in[d,d+\Delta d]} \langle t_{ij} \rangle}{\Delta d} &\propto \frac{1}{\hat{S}_{ij}(0,d)} \frac{\hat{S}_{ij}(d,d+\Delta d)}{\Delta d} \\ \hat{S}_{ij}(r_1,r_2) &= \sum_{j|d_{ij}\in[r_1,r_2]} \hat{s}_j^{in}. \end{aligned} \tag{9.10}$$

The cost d_{ij} may be measured in terms of time, perceived cost or euclidean distance in space and is continuous in its original formulation. $\hat{s}_j^{in}$ is related to the number of opportunities at location j. Stouffer states that any desired metric may be used to account for distances and any desired function may be used to relate opportunities and distance, or opportunities and city traits, and hence his model is fully flexible (yet he uses euclidean distances in a discretized grid and proportionality between strengths and opportunities). The model uses 2N inputs and one or various parameters relating the density of opportunities and observed traffic for each location. Subsequent authors

[8] The original model assumes somehow a continuous description where $\sum_{j|d_{ij}\in[d,d+\Delta d]} t_{ij}$ is identified as Δy and x is a cumulative measure of opportunities, which initially Stouffer relates to housing density but ends up stating it to be proportional to what we know as aggregated incoming node strength.

have re-interpreted the model: Apart from the previously work done by Wilson, including it in the collection of entropic models (and hence displaying all positive previously mentioned aspects) other authors have succeeded in relating it to a stochastic model [154].

More recently, in [164] a novel opportunistic approach was proposed, which in addition yielded interesting analytical results giving birth to what has been called *the Radiation model*. The radiation model is based on assuming that each user that wishes to move (there are $\hat{s}_i^{out}$ such users in each location) is interested in reaching the place j nearest to his present location i that has the maximum value z_j (superior to his current value z_i) of a random variable z that encodes the opportunities present in each place. Hence, for each place k, $\hat{s}_k^{in}$ random trials of the random variable z are drawn and its maximum value in each location is chosen. Then, the user selects the closest location to his home such that this maximum value is greater than the one obtained at their current location. Mathematically, it can be proven that the trips between locations can be finally expressed as[9],

$$\begin{aligned} \langle t_{ij} \rangle &= \hat{s}_i^{out} p_j^{(i)} \\ p_j^{(i)} &\propto \frac{\hat{s}_j^{in}}{(\hat{s}_i^{in} + \hat{s}_j^{in} + S_{ij}(0^+, d))(\hat{s}_i^{in} + S_{ij}(0^+, d))}. \end{aligned} \tag{9.11}$$

The latter expression corresponds to a multinomial process centered in each node with $\hat{s}_i^{out}$ trials over normalized probabilities $\sum_j p_j^{(i)} = 1$ (this implementation is much faster than the stochastic based approach described and allows to rapidly obtain network instances). This model has attracted a lot of attention due to its lack of fitting parameters and exceptional predicting power defended by its authors. However, some caveats can be identified. First of all, it needs as input 2N values related to population plus an additional parameter to establish the proportionality relation between population and strength of a location (or otherwise only 2N incoming and outgoing strength input pairs). Additionally, the model in its present form entails some mathematical and algorithmic problems when considering the high sampling limit. First of all, concerning its stochastic version, the model relies on maximum value statistics, but if the number of trips is effectively infinite, the model predicts no movement at all (the maximum value of z will coincide with $\max z$ in every location), also its complexity is larger than $\mathcal{O}(N\hat{T})$, complicating the generation of samples and subsequent averaging. Furthermore, considering its node-based multinomial form (complexity[10] $\mathcal{O}(L)$) one can see that while the outgoing number of trips is conserved for each location ($\hat{s}^{out} = \langle s^{out} \rangle$)

[9] In here, we have adapted the original expression in terms of our variables, substituting population by node strength. Note also, that this model predicts zero flows for nodes with either zero incoming or outgoing strength, and hence for sparse datasets, it must be adapted (it must also be adapted to allow self-loops).

[10] We do not consider here the computation of S_{ij} values.

MODEL	COMPLEXITY	INPUTS	$P(\vec{T})$	CALIBRATION	$\hat{T} \rightarrow \infty$
GRAVITY	-	2N+3	-	Yes	No
SEQ	$\mathcal{O}(N\hat{T})$	2N+1	$\hat{T}$ Multinomial	Yes	Yes
RAD	$\mathcal{O}(N)$	2N	N Multinomial	No	No
WILSON	$\mathcal{O}(L)$	$2N+1$	L Poisson	Yes	Yes
MECM	$\mathcal{O}(L)$	2N	L Poisson	No	Yes

Table 9.2: **Comparing network generation models.** All models conserve the total number of trips $\hat{T}$ and $\hat{T} \rightarrow \infty$ indicates whether the model is well defined in the high sampling limit so that $\left\{p_{ij}^{\infty}\right\}$ are well defined.

and the scaling on $\hat{T}$ behaves according to a ME description, the incoming strength cannot be matched to the original data ($\langle s^{in} \rangle \neq \hat{s}^{in}$), and hence $\hat{p}_{ij}$ predictions must be necessarily wrong, which is in contradiction with our observations. Finally, some problems related to normalization have been pointed out by other authors and doubts on its exceptional predicting power with regards to other models have been raised [123].

Some additions have been proposed [195, 101, 196, 144] and also a space-continuous extension [165] but the essence of the model remains unchanged. Due to recent interest, we shall use it as representative of the Opportunistic family of analytical models in our comparison.

The main features of all considered models are depicted in Table 9.2.

9.3 SO WHICH IS THE BEST PERFORMING MODEL FOR THE TAXI CASE?

To compare the different approaches taken, we proceed to generate predictions using different models belonging to each of the types explained. As empirical model we use the Sequential gravity model and as analytical models we use the radiation model to represent the opportunistic class of models and Wilson's doubly constrained, maximum entropy gravity model with exponential deterrence function to represent the family of entropic ones. Also, as baseline, we use the MECM studied in earlier chapters, which, after all, is a distance-agnostic model only taking into account the density of *sinks* and *sources* of mobility in the city and which we know from earlier chapters that performs moderately well as a first level of approximation.

In Table 9.3 we present the general model indicators earlier proposed for each model (additional graphics showing network structure and correlation $(\hat{t}_{ij}^{data}, \hat{t}_{ij}^{model})$ can be found in Section D.5 of the Appendix).

The first important aspect to notice is the earlier mentioned problem of the sampling. As datasets become sparser, the quality of our indicators decreases enormously reaching the extreme case of VI where almost all models display equally (poor) results. At first sight we see how loglikelihood and entropic measures are only appropriate tools to discriminate the quality of models with the same underlying mechanisms, and specifically for entropic measures, this is only true in the case where large sampling is available[11]. However, the level of detail provided by network related topological features (see Section D.1) is still a good tool to discriminate among models.

In Section 10.1 a possible solution to the sampling problem is proposed, studied and applied.

Furthermore, as we already know, we see that the MECM performs acceptably well despite being agnostic with respect to distances (in fact, for the SF case it ranks among the best performing models). It displays distinctively better results than the Radiation model, and similar (and in some cases superior) results as the Sequential gravity model. In my opinion, this model should always be used as benchmark in this kind of studies, and we hope it will be more adopted in the future (as for instance in [118]), as it serves as a good basis to assess the capacity of other models to take into account the importance of distance.

The Radiation model fails completely at predicting flows in the city environment (finding confirmed in [115]). The reasons behind this fact can be seen both on the topological structure of the resulting networks and the trip distance distribution. The structure of the network is extremely assortative and trips tend to over-concentrate among large sized nodes (see under-expected binary degrees, large disparity values and extreme assortative profile in Figure D.1), which are mostly concentrated around city centres (with the exception of airports), fact which can be seen in the very fast decay of the trip distance distribution. This is due to the extreme density of the city environment, i. e. high strength locations of the city accumulate so many opportunities that hardly any trips are allocated (starting or ending) outside these areas. Also, it is worth noticing the extreme heterogeneity of obtained results (see large bars of the box-plot in Figure D.2), which suggest a high level of variation among different instances of the same model.

The Sequential gravity model can be considered as a middle ground between the MECM and the Wilson model. Despite being different,

[11] Note that for VI and SF, $|\Delta\hat{S}^{\Gamma}/\hat{T}|$ for the empirical Poisson model displays worse results (larger absolute values) than either the MECM, WILSON and SEQ models. As already pointed out, the reason behind this is that for the empirical models over a single run, the probability to observe a sparser network with $E < \hat{E}$ is large leading to less informative realizations. This circumstance is only partially avoided averaging the results sufficiently over model instances, since it strongly depends on sampling (the further $\min\{\hat{t}_{ij}\} > 1$, the better).

MODEL	$\langle \mathrm{CPC}_{\mathrm{SAMPLE}} \rangle$	R^2	$\mathcal{L}/\hat{T}$	$\Delta \hat{S}^{\Gamma}/\hat{T}$
NY				
SEQ	0.60	0.13	-	-0.20
RAD	0.05	-145.84	0.001	-5.07
WILSON	0.69	0.47	0.38	0.29
MECM	0.64	0.42	0.51	0.44
Empirical	0.94	1.00	0.09	-0.03
SI				
SEQ	0.34	-0.77	-	-0.62
RAD	0.04	-69.05	2.05	-4.34
WILSON	0.37	0.19	1.92	0.40
MECM	0.30	0.08	2.26	0.62
Empirical	0.78	0.94	0.53	-0.25
SF				
SEQ	0.21	-5.41	-	-0.08
RAD	0.07	-76.42	0.14	-3.26
WILSON	0.23	-0.65	3.05	0.15
MECM	0.23	-0.13	3.07	0.18
Empirical	0.70	0.46	0.77	-0.40
VI				
SEQ	0.04	-6.08	-	0.02
RAD	0.04	-42.28	6.91	-2.08
WILSON	0.04	-2.95	5.50	0.11
MECM	0.03	-3.00	5.85	0.13
Empirical	0.65	-1.86	0.93	-0.52

Table 9.3: **Comparing global performance model indicators.** Average values performed over $r = 100$ instances (Sequential model averaged over $r = 10$ due to slow convergence). R^2 computed using $\hat{t}_{ij}^{\mathrm{model}}$ over a single run and t_{ij}^{data} obtained from empirical data using non-zero entries. Standard deviation for $\langle \mathrm{CPC} \rangle < 1e-4$ in all cases not reported.

at its basis, the form of expected average trips does not differ from the Wilson model, yet its generating mechanism is essentially like a biased Micro Canonical configuration model. The apparently surprising small values of entropy excess $\Delta\hat{S}^{\Gamma}/\hat{T}$ are explained precisely by this fact: The micro-canonical nature of the model heavily constrains its variability or randomness, and this is reflected in information related indicators. Another consequence of the configuration-like inspiration of the model is related to convergence. If no self-loops are allowed, then the algorithm has no guarantee of being able to match all node-stubs [65] (and this problem grows in importance as the skewness of the underlying strength distributions grow). Even if self-loops are allowed, this results in the majority of final trips allocated along the process resulting in self-loops, and hence distinctive difference appear between in and out binary degree pairs per node (specially for large strength nodes) as can be seen in Figure D.4 (Appendix).

Finally, Wilson's model is distinctively the best performing model. Although it is true that it uses and additional parameter with respect to the other models, its performance allows us to extract some conclusions from the observed mobility. Essentially, Wilson's model only adds an additional ingredient to the city layout encoded in the distribution of node strengths, that of average fixed cost per trip, i. e. the (small but distinctive) influence of cost in displacements. Since scattered locations tend to be at large distance from high strength nodes, the assortative network profile from empirical data which the MECM could not reproduce is partially (but not fully) recovered (see Figure 9.1 and Figure 9.2), because trips among small (but close) nodes are favoured. Also, having fixed the average cost of displacement, the trip distribution is very closely reproduced (Figure 9.4).

Even if this model is the best performing among the ones considered here, the most interesting information one can extract from it is by putting the focus on the aspects of mobility it fails to properly characterize. And the most noteworthy aspect continues to be the general presence of abnormally large flows among locations in the tail of the distribution of values of $\hat{t}_{ij}$ and the over-expression of trips among origin and destination low-strength nodes. The only exception being Vienna, in which case the distance of the airport to the center of the city combined with its isolation favours highly concentrated short trips between airport nodes (hence $\max\left\{\hat{t}_{ij}^{\text{data}}\right\} < \max\left\{\hat{t}_{ij}^{\text{model}}\right\}$).

The values $\gamma, \{x_i\}, \{y_j\}$ for the Wilson model have been obtained solving the associated saddle point equations of the model using the recipe in Section B.1 with error in all cases inferior to 10^{-7}. The package to do so is implemented in [10].

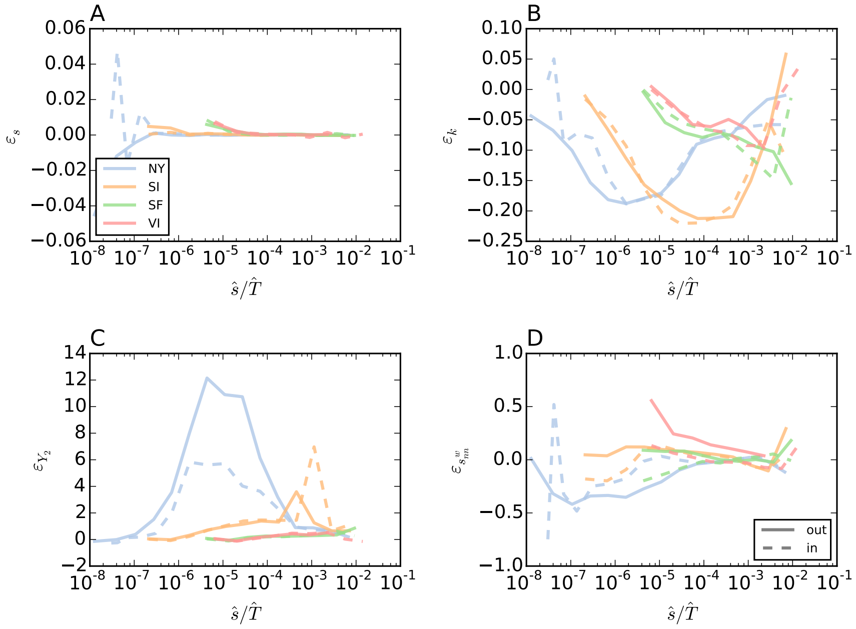

Figure 9.1: **Relative difference $\varepsilon = (\hat{x} - \langle x \rangle_{\textbf{Wilson}}) / \langle x \rangle_{\textbf{Wilson}}$ between empirical node properties and Wilson model predictions.** Relative strength difference (A), degree (B), disparity (C) and node neighbor strength correlation (D) averaged using logarithmic binning over $r = 10^2$ instances of the model.

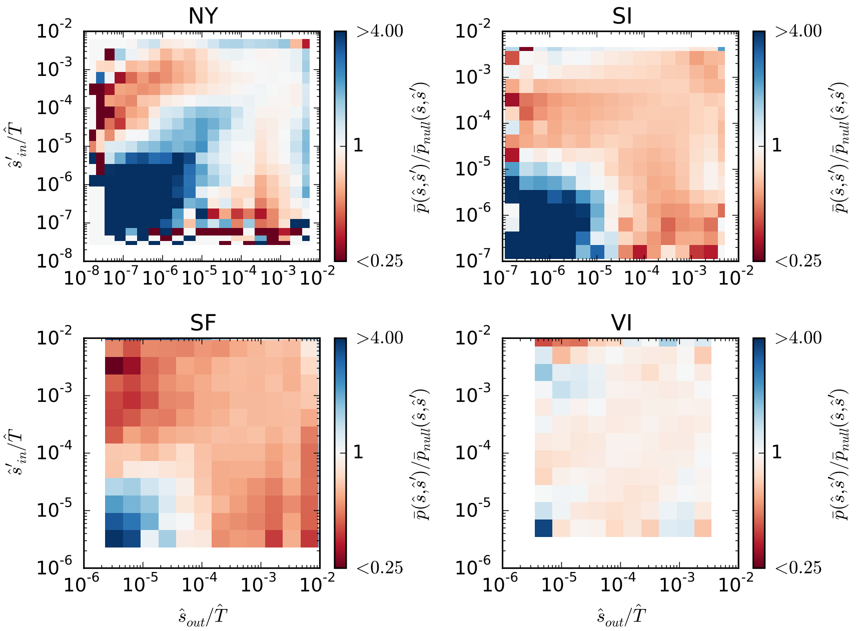

Figure 9.2: **Comparison between empirical data and Wilson model at edge level.** Relative scaled occupation number as function of starting and ending node strength comparing empirical data and Wilson model over a single run. Both cases are normalized over the bins.

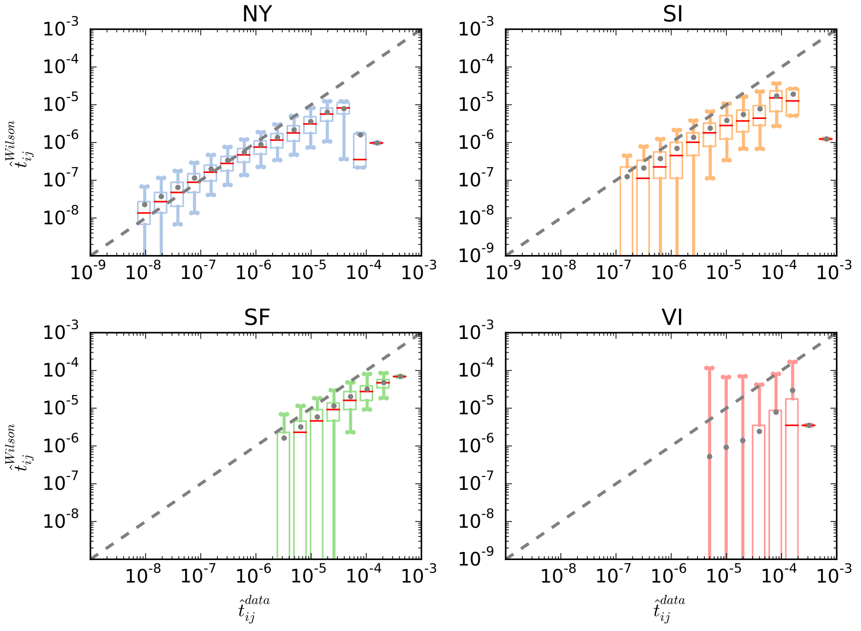

Figure 9.3: **Comparison between empirical data and Wilson model at node-pair level.** Box-plot showing correlation between relative scaled occupation number between model predictions and data over a single run. Solid lines mark the [5%, 95%] interval, median marked as red horizontal line and average value with grey dotted point.

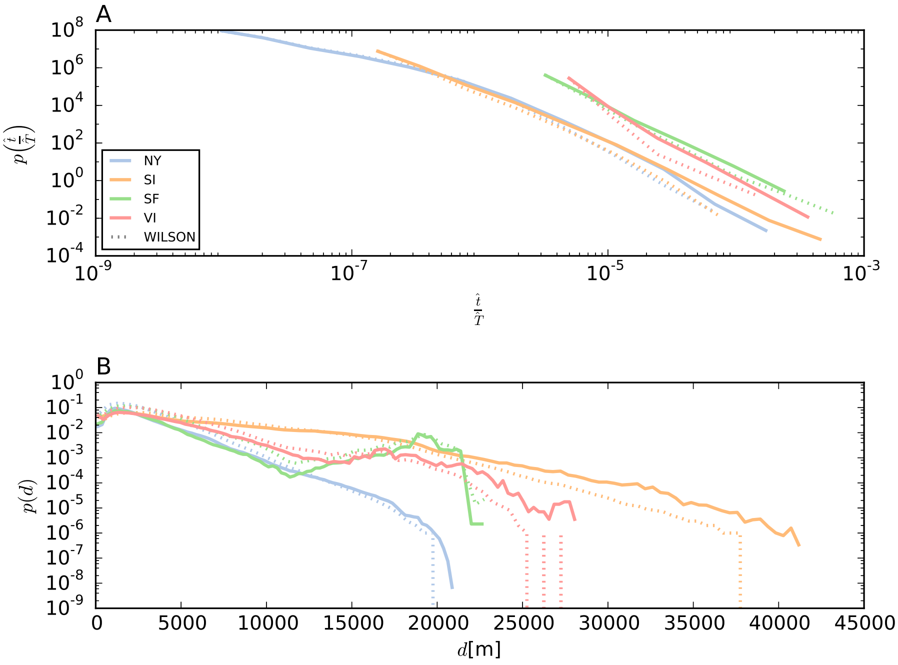

Figure 9.4: **Wilson model occupation number and trip length distribution.** Occupation number distribution (A) and trip length distribution (B). Dotted lines correspond to the model averaged over $r = 10^2$ instances while filled lines to empirical data.

In my opinion, this model is very flexible and adaptable because it adds the possibility to encode known info using our theoretical developments, such as for instance the intrinsic importance of trips among special pair of locations due to their sociological importance, which can hardly be captured by any model. Precisely this circumstance will be used in the upcoming chapter to generate realistic and precise predictions of mobility based on existing data. As conclusion, it is worth noticing that our results agree well with those depicted in [112], where a similar comparison is performed (however, for the Wilson model the γ value is not extracted from maximum likelihood arguments). The results described here also confirm the lines of though presented in [118] and discussed in earlier chapters.

9.4 WRAPPING UP: MAIN APPROACHES TO OD MODELLING AND IMPORTANT ASPECTS OF MODEL VALIDATION

In this section we have performed an exhaustive analysis of different approaches to model trip generation in the form of Origin-Destination matrices. We have presented different types of models and reviewed its main ingredients, strengths and weaknesses. In order to do so, firstly we have justified, reviewed and introduced a range of indicators focusing on different aspects of mobility that allow for a holistic assessment of models. The main conclusions drawn are summarized below.

A. **Model indicators:** We have presented three types of model indicators: Network topology metrics, information indicators and matrix similarity indicators. For each of them, their main characteristics are:

- **Topology indicators:** Network related measures, while spatially agnostic, are a good tool to assess model performance beyond single number indicators and are able to give an overview of the performance of each model while not being extremely affected by sampling.
- **Information indicators:** We have seen how entropy related measures are only good indicators to discriminate among models belonging to similar ensembles but with different constraints (hard-constrained models cannot be evaluated in equal terms to soft-constrained unless all node-pair have large sampling, which is hardly the case for real data). Similar principles apply to likelihood measures, which tend to give better result for models predicting large weights (close to real data in the tail), with less regard for precision for small values of trip occupation. However, this kind of indicators are only robust for datasets where a large fraction of node-links are occupied, $\hat{E} \sim L$ and cannot be used to compare models with different underlying generating mechanisms, hence must be used with care.
- **Matrix similarity indicators:** We have also seen that matrix similarity indicators are prone to be affected by sampling, yet, are useful in providing information encoded in single numbers (specially the CPC index) to concentrate model performance in simple indicators, and also are useful for comparison of models generated with different mechanisms and underlying statistics.

B. **Model review:** Among all the considered models, for the taxi datasets, we have justified that the most promising approach is the one provided by the entropic models earlier presented.

> Even in the distance-agnostic case of the MECM, it beats or performs equally well than other proposed models, with the additional advantage of displaying flexible and convenient analytical properties (such as adequate behaviour in the large sampling limit). We have furthermore confirmed the importance of city layout and node-strength allocation across the city, and the non-negligible influence of distance or cost-perception, which is needed to accurately model human flows. However, we have also seen that these two ingredients cannot solely account for the observed traffic. Context-dependent, intrinsic sociological and urban features of each city need to be accounted to explain the most frequently observed trips.

Concerning explicitly the urban environment, I feel that the reason behind the failure of the radiation model (but in general all opportunistic-based models) to capture the essentials of mobility are not only related to the large spatial concentration of highly busy locations, but mainly because they are all based in assuming a certain radial isotropy around nodes in the city, which is in clear contrast with observations (for a complete critique see [119]). A possible way to merge the success of entropic models (such as Wilson's one) with the reasoning behind opportunistic based ones, would possibly be to break this radiality and consider the cost associated to distance to opportunities falling in a certain (not radial) region of space between two given points (such a model would also introduce the insights of the earlier mentioned empirical-ranked model [131]).

Along these lines, and given that the interplay among node importance and distances seem to play a prominent role in shaping mobility, exploring the possibility of embedding ODs in non-conventional metric spaces such as hyperbolic planes (which have been shown to be able to capture interesting features of binary complex networks [108, 41, 161, 14, 45]) could seem a promising approach.

For the afore-mentioned reasons, we argue that in the future, efforts coming from the complexity physics side of research should be focused in interpreting the relation between cost and distance (and the way humans perceive it [105]) and also in relating incoming and outgoing traffic (strengths) of each location with city and urban attributes or indicators. Concerning the models, effort should be focused not in presenting new "successful" models with new data, but rather in expanding in a critical way the ones already present, or at least, performing systematic reviews such as the one performed here (or for instance [112]) in order to discard, update and improve them. Otherwise, given the enormous richness of data available (but bearing in mind the limitations of such data and the challenges they pose), it will be very hard to extract meaningful, universal conclusions related to mobility.

In a nutshell, what we must bear in mind is that, quoting the famous aphorism by George E. P. Box [44]:

> Essentially, all models are wrong, but some are useful.

The importance of models lies in their weaknesses rather than in their strengths: What pushes science forward is the critical exercise of assessing the limitations of any proposed theory and try to address them, rather than highlighting their exploits.

10 PUTTING MAXIMUM ENTROPY NON BINARY NETWORK MODELS TO USE: APPLICATIONS TO MOBILITY

> *You look at science (or at least talk of it) as some sort of demoralising invention of man, something apart from real life, and which must be cautiously guarded and kept separate from everyday existence. But science and everyday life cannot and should not be separated.*
>
> — Rosalind Franklin [121]

All the results presented in this thesis until the present moment are either developments on a theoretical side or a purely data exploration and analysis perspective. In this final chapter, my aim is to present two practical examples that may help in depicting how we can exploit the previously gained insights on urban human mobility to solve data related problems.

Using the maximum entropy framework for model generation and the greatly stable topological and temporal statistics uncovered in the previous analysis can be extremely helpful for facing matters ranging from sampling to visual representation of networks, for which, in the following, two examples are shown.

10.1 SOLVING THE SAMPLING PROBLEM: TAXI SUPERSAMPLING

In some scenarios, fully grasping a certain mobility-related phenomenon may require modelling the entire population of interest. For instance, it was shown that the fraction of taxi trips that can be shared in the city of New York is an increasing (albeit not simple) function of the number of daily taxi trips [149]. Hence, if a certain data set covers only a fraction of the daily taxi trips performed in a city, the taxi sharing potential cannot be fully unveiled. The above discussion motivates the need of extrapolating urban mobility data starting from a subset of the population of interest. Although a number of urban mobility studies have applied such methods [139, 172], a definition and assessment of a statistically rigorous extrapolation methodology is so far lacking.

In this section we fill these gaps by introducing a methodology to tackle the problem of obtaining an accurate picture of a mobility process when only a limited observation of it is available, both in time and volume. Based on our previous observations related to the normalization, stability (Chapter 7 and Chapter 8) and performance of maximum entropy trip generation models (Chapter 9) and exploiting the regularity of trip allocation among intersections encoded in $\{p_{ij}^{\infty}\}$, we use a maximum entropy approach combining empirical data to model the occurrence of the core of frequent trips with an exponential gravity model [190, 70, 2] modelling the least-frequent trips.

We apply the method to accurately reconstruct the NY data set using small fractions sub-sampled from only a month of recorded data. We finally assess and validate the statistical accuracy of the proposed *supersampling* methodology using the model validation tools earlier developed.

The general maximum entropy based theory for model generation earlier derived allows us to efficiently exploit both the observed temporal stability features and the heterogeneous topological properties of the network to solve the *supersampling* problem at hand. Assuming that the mobility process is driven by some general (unknown) constraints, such as population density or budget, we have earlier shown that for any desired level of sampling T_d the statistics of trips for each pair of nodes can be well described by a set of L independent Poisson processes with mean $\langle t_{ij} \rangle = \hat{T}_d p_{ij}^{\infty}$.

It seems clear that from knowing the *real* values of the collection $\left\{p_{ij}^{\infty}\right\}$, *supersampling* a mobility data set would be a trivial operation of generating L independent Poisson processes using the provided proportionality rule. Therefore, the problem now reduces to inferring the collection of values $\left\{p_{ij}^{\infty}\right\}$ from an available data set. We assume that only one snapshot of the aggregated mobility network is available to this end (thus assuming no temporal information is available on the trip data) as is usually the case in mobility studies. The maximum likelihood estimation of such values corresponds to

$$p_{ij}^{\infty}|_{ML} = \frac{\hat{t}_{ij}}{\hat{T}} = \hat{p}_{ij}. \tag{10.1}$$

There is, however, a practical issue in this formula related with the normalization condition for the random variables $\{p_{ij}\}$ and the presence of empty intersection pairs in the available observed data, as already pointed out in Chapter 8 (see specially Section 8.1). For such intersections, using the formulas above, we have that $\hat{p}_{ij} = \hat{t}_{ij} = 0$, whereas their *real* p_{ij}^{∞} value is unknown but fulfils $p_{ij}^{\infty} \in [0, \hat{p}_{min} \simeq \hat{T}^{-1}]$. Since by definition both collections $\{p_{ij}^{\infty}\}$ and $\{\hat{p}_{ij}\}$ need to be normalized, and denoting the set of active edges as $\mathcal{E} = \left\{ij|\hat{t}_{ij} > 0\right\}$, we have,

$$\begin{aligned} \sum_{ij} p_{ij}^{\infty} = 1 = \sum_{ij|ij\in\mathcal{E}} p_{ij}^{\infty} + \sum_{ij|ij\notin\mathcal{E}} p_{ij}^{\infty} \\ \sum_{ij} \hat{p}_{ij} = \sum_{ij|ij\in\mathcal{E}} \hat{p}_{ij} + 0 = 1, \end{aligned} \tag{10.2}$$

from which we see that $\sum_{ij|\in\mathcal{E}} p_{ij}^{\infty} \leqslant \sum_{ij|\in\mathcal{E}} \hat{p}_{ij} = 1$.

Hence, in general we cannot consider the empirically observed probabilities $\hat{p}_{ij}$ as a good proxy for the *real* values of p^{∞}_{ij} unless the number of empty intersection pairs is very reduced. Given that the percentage of active edges (pairs of nodes for which $\hat{t}_{ij} > 0$) is a very slowly increasing function of the sampling (see Figure 8.3), inferring directly the set of probabilities $\left\{p^{\infty}_{ij}\right\}$ empirically would take an enormous data set – note that even with over a year of data for the NY dataset only roughly 45% of edges are covered.

For the reasons given above, a simple proportionality rule using equation (10.1) is not a good *supersampling* strategy, specially for skewed and sparse data sets.

10.1.1 *Inferring intersection pair trip shares* $\{p^{\infty}_{ij}\}$

Based on the previous discussion, we now present the methodology for *supersampling* an urban mobility data set that consists in inferring the collection of $\{p^{\infty}_{ij}\}$ values from a set of aggregated empirical trips $\left\{\hat{t}_{ij}\right\}$. We should bear in mind that despite the maximum likelihood formula in equation (10.1) cannot be directly used for the empty intersection pairs in the data, it does perform well for non-empty intersections (see Figure 8.1).

The maximum entropy based framework can naturally combine any constraint driven model with the rich information encoded in the trip sample. We propose a method to predict trips based on the theory mentioned earlier: Taking the L intersection pairs (being them active edges in the data set or not), we split them into two parts, the subgroup of *trusted* trips defined as $\Omega = \left\{ij|\hat{t}_{ij} > t_{min}\right\}$ and its complementary part Ω^C. The value t_{min} is a threshold modelling a minimal statistical accuracy that depends on the amount of data available, and which may be set to 1 in practical applications. We keep the proportionality rule $\hat{p}_{ij} \propto \hat{t}_{ij}$ for the *trusted* trips, while for the remaining trips we apply a doubly constrained exponential gravity model. In other words, we generate a collection of $\left\{p^{\infty}_{ij}\right\}$ values,

I chose the doubly constrained exponential gravity model because it has been shown to be the best performing model in Chapter 9, however, we could have chosen any other maximum entropy models discussed earlier.

$$p^{\infty}_{ij} = \begin{cases} \frac{\hat{t}_{ij}}{\hat{T}} = \hat{p}_{ij} & ij \in \Omega \\ x_i y_j e^{-\gamma d_{ij}} & ij \in \Omega^C \end{cases} . \tag{10.3}$$

The values γ and $\left\{x_i, y_j\right\}$ are the $2N+1$ Lagrange multipliers satisfying the following equations

$$\begin{aligned} \hat{s}^{out}_i - \hat{T} \sum_{i|ij \in \Omega} \hat{p}_{ij} &= \hat{T} x_i \sum_{i|ij \in \Omega^C} y_j e^{-\gamma d_{ij}} \\ \hat{s}^{in}_j - \hat{T} \sum_{j|ij \in \Omega} \hat{p}_{ij} &= \hat{T} y_j \sum_{j|ij \in \Omega^C} x_i e^{-\gamma d_{ij}} \\ \hat{D} - \hat{T} \sum_{ij|ij \in \Omega} d_{ij} \hat{p}_{ij} &= \hat{T} \sum_{ij|ij \in \Omega^C} d_{ij} x_i y_j e^{-\gamma d_{ij}}, \end{aligned} \tag{10.4}$$

where $\hat{D} = \sum_{ij} \hat{t}_{ij} d_{ij}$ is the total euclidean distance of the observed trips (d_{ij} stands for the distance between intersections i and j). Note that, by construction, the values are properly normalized, i.e., $\sum_{ij} p_{ij}^{\infty} = \sum_{ij \in \Omega} \hat{p}_{ij} + \sum_{ij \in \Omega^c} x_i y_j e^{-\gamma d_{ij}} = 1$.

Equations (10.3) are solved using the balancing Algorithm 2 and its implementation is provided in [10].

The model presented earlier needs to deal with the issue of inactive nodes that do not appear in the original data due to poor sampling, i.e., nodes for which $\hat{s} = 0$ either in the incoming or outgoing direction. This issue has a minor impact in our case due to the previously observed rapid coverage of the number of active nodes (see Figure 8.3). In any case, it can be solved: given that the geographic positions of the nodes are available, we could always artificially assign a certain relative strength to the nodes not present in the data using complementary call detail records [114], census data or points of interests (POI) data, or assign them some values according to a chosen distribution depending on the data at hand. For simplicity, in our case we have chosen to keep only the nodes present in the original data.

10.1.2 *Assessing the quality of the supersampling methodology*

To test the *supersampling* methodology, we have proceeded to select a timespan of the NY data set corresponding to an observation period of $\tau = 1$ month (February 2011) from which we further randomly subsample different fractions f used as *training sets* to compute $\left\{p_{ij}^{\infty}\right\}$ applying equations (10.3) and (10.4). We then reconstruct the OD using the proportionality rule $\left\{\langle t_{ij} \rangle = \hat{T}_d p_{ij}^{\infty}\right\}$ for both the complete and reduced data set, $\hat{T}_d = \hat{T}(\tau' = 1 \text{ year})$ and $\hat{T}_d = \hat{T}(\tau = 1 \text{ month})$. Finally, we compare the model predictions with the set of empirically observed trips in these periods.

The results for the supersampling method quantified using some of the earlier developed indicators[1] (see Chapter 9) are summarized in Table 10.1 and a specific example for $f = 0.1$ (reconstruction using only 10% of the original data of the monthly data set compared to yearly data) is shown in Figure 8.1 for different network indicators

[1] R^2_{cond} is a modification of the R^2 indicator defined in (9.5), using the average occupation conditioned on link existence $\langle t_{ij} | t_{ij} > 0 \rangle = \langle t_{ij} \rangle (1 - e^{-\langle t_{ij} \rangle})^{-1}$ instead of $\langle t_{ij} \rangle$, with also a change in the denominator,

$$R^2_{\text{cond}} = 1 - \frac{\sum_{ij|\hat{t}_{ij}^{\text{data}}>0} \left(\left\langle t_{ij}^{+,\text{model}} \right\rangle - \hat{t}_{ij}^{\text{data}}\right)^2}{\sum_{ij|\hat{t}_{ij}^{\text{data}}>0} \left(\left\langle t_{ij}^{+,\text{model}} \right\rangle - \overline{\left\langle t_{ij}^{+,\text{model}} \right\rangle}\right)^2}.$$

Both indicators tend to converge for high sampling, so differences with respect to this indicator are negligible. For the case of the CPC indicator, since in this case explicit expression for $\{\langle t_{ij} \rangle\}$ are always available, we will use the version less prone to sampling effects (9.4) (instead of (9.3)). This chapter is based on the published work [4] and we keep the original formulation appearing there.

	$\tau = 1$ month				$\tau = 1$ year		
f	f_Ω	$\mathcal{L}/\mathcal{L}_0$	CPC	R^2_{cond}	f_Ω	CPC	R^2_{cond}
1.00	0.8855	1.46	0.92	1.00	0.07090	0.78	0.91
0.75	0.6417	1.64	0.83	0.98	0.05138	0.76	0.89
0.50	0.4014	1.88	0.77	0.94	0.03214	0.74	0.86
0.25	0.1711	2.26	0.68	0.83	0.01370	0.69	0.78
0.10	0.0492	2.64	0.60	0.65	0.00394	0.65	0.63
0.01	0.0012	-	0.58	0.21	0.00010	0.66	0.26
0.005	0.0003	-	0.59	0.12	0.00002	0.66	0.18
MECM	-	-	0.57	-0.87	-	0.64	-0.22
Empirical	1	1	1	1	1	1	1

Table 10.1: **Validation of the supersampling method.** See details on each indicator in Section 9.1. The number of *trusted* trips fed to the model relative to the entire number of generated trips $f_\Omega = \sum_{ij|ij\in\Omega} \hat{t}_{ij}/\hat{T}_d$ is reported in the third column. The *Supersampled* models with different fractions f are only generated using subsamples of the training set (1 month observation period). *Empirical* stands for the model generated using the empirical probabilities $\hat{p}_{ij}$ (equation (10.1)) of the full data set and MECM stands for the multi-edge configuration model applied to the full data set.

proposed. For comparison, results using both the MECM and the empirical values $\{\hat{p}_{ij}\}$ using the information encoded in the complete dataset are also shown.

We observe an accurate reconstruction of the mobility network for a wide range of values of f, which shows the validity of the proposed supersampling methodology. At the global scale, even at extreme levels of subsampling, our model is successful at reconstructing the original dataset. Also at the topological scale, despite the heterogeneity in the underlying distributions, the methodology generates very accurate predictions. The predictions for the least frequently visited nodes display higher relative errors due to the presence of inactive nodes in the training dataset (1.6% of total nodes for $f = 0.1$).

Upon close inspection, our inferred values $\left\{p^{\infty}_{ij}\right\}$ slightly over-estimate low-valued weights and underestimate large-valued weights and strengths, yet the errors are small as we can see in Figure 10.1-B. See Figure 10.1-B (green line) and Figure 10.1-E (green dots) where we can observe a gap around $t \sim 100$ and $\left\langle t^{+}_{ij} \right\rangle \sim 100$, respectively, which corresponds to the separation between the trusted empirical data (separated points in the background belonging to the group of trusted trips Ω) and the reconstructed trips (clustered cloud of points). The minor seasonal fluctuations found in our temporal analysis (Chapter 7 and Chapter 8) together with these over- and under-estimations explain the mi-

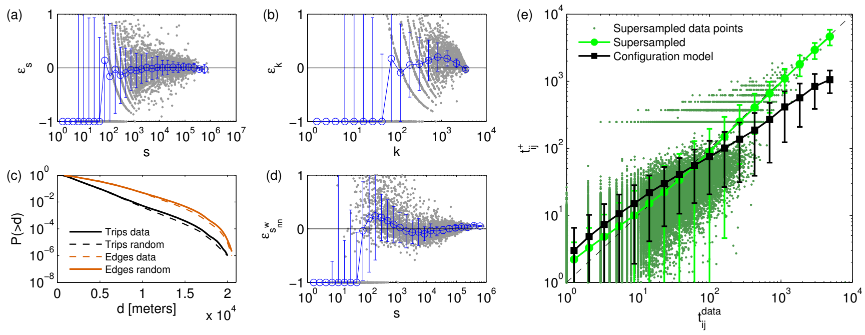

Figure 10.1: **Supersampling results.** Main network differences between real data accumulated over a year and Supersampled model from real one month data with $f = 0.1$ subsampling. **(a)-(d)** Relative error $((\langle x \rangle - \hat{x})/\hat{x}$ with $\hat{x}$ being a magnitude measured from the aggregated yearly network) between reconstructed network using supersampling and original data for outgoing degrees (a), strengths (b) and average neighbor strength (d) (similar results for incoming direction not displayed). The complementary cumulative distribution function of both edge lengths and trips lengths (c) is also shown. **(e)** Comparison between empirical $\{\hat{t}_{ij}\}$ values and model prediction over a single run. Configuration model expectation from a single run using the full year data set is also shown for comparison. All results averaged over 100 repetitions of the model, error bars represent standard deviations on the log-binned data and raw data is shown in the background. For visual clarity, panel e) only shows a random subsample of 1% of the raw data in background.

nor limitations of the model to reproduce perfectly the entire yearly data set.

The second order effects induced by the seasonality of recorded data can also be seen in the performance of our methodology under extreme levels of subsampling (using around 1% of the sample monthly data to feed the model or less). In these circumstances, the model is still able to produce a good prediction of the empirical data, yet it reproduces better the accumulated yearly mobility rather than the monthly one since the inherent seasonal variations of traffic between certain intersection pairs are smoothed by the aggregation procedure.

Furthermore, in the event that enough historical data were available, we could achieve even better results by computing the collection $\{\langle p_{ij} \rangle_\tau\}$ with an appropriate τ period (depending on the granularity of the data) and approximating $p_{ij}^\infty \simeq \langle p_{ij} \rangle_\tau$ for the group of *trusted* trips (equation (10.4)). Such a procedure, which may be extended to overcome the minor limitations imposed by the seasonality of the data and other improvements related with the presence of non-active nodes could be derived to perfect the method. Figure 8.1 and Figure 8.2 in earlier chapters are generated using precisely this proposed methodology.

10.1.3 *Extension of the static supersampling in time*

The above method has been used here to reconstruct a static (aggregated) sample of a mobility process represented as an OD from a limited dataset. However, given that we have detailed knowledge of the temporal dynamics of trip generation in the city of NY, an extension can be derived to distribute trips in time if needed: Taking into account the exponential nature of inter-events times between trips (Figure 7.5) and the change of its average among different hours (Figure 7.4), together with the static picture provided by the entire collection of $\{p_{ij}^\infty\}$ given by (10.3) and the close-to-gaussian statistic of daily trip generation we could readily implement the proposed Algorithm 4.

Using the static procedure to infer $\{p_{ij}^\infty\}$, we could use Algorithm 4 to distribute trips in time, taking as input the collections $\{p_{ij}^\infty\}$ and the average and standard deviation of number of trips generated per day type and hour (see Figure 7.4) and an inflation factor f. In a nutshell, for each day we generate $\hat{T}_d$ trips Gaussianly (see Figure 7.4 A and B) with average and standard deviation empirically introduced but augmented by a factor f. We then distribute the $\hat{T}_d$ trips among the different hours according to a probability q_h (Figure 7.4 D) and the different intersections according to p_{ij}^∞. Finally, the t_{ij}^h trips allocated per hour in each intersection are distributed in time according to exponential inter-event times with average $3600/(\hat{T}_d p_{ij}^\infty q_h)$. This

Algorithm 4: Trip time and intersection supersampling algorithm.

```
Input:  Number of intersections $N$, City intersection pair
        invariants $\{p^{\infty}_{ij} \forall ij = 1...N^2\}$ (float vector), Weekday
        average and standard deviations of trip generation
        $\{(\overline{T}_{day}, \sigma^2_{T_{day}})_d \forall d = 1...7\}$ (float vector), Hourly
        probability of trip allocation $\{q_h \forall h = 1...24\}$ (float
        vector), number of days to sample $d_{days}$ (int), inflation
        factor $f$ and starting time $\tau_0$.
Output: List of generated trips $(i, j, \tau) \forall t = 1...T$ (vector with
        each entry a 3 float tuple).
begin Initialization
    Set $d = 0$. Set $\vec{L} = \vec{0}$. Set $\tau = \tau_0$;
end
begin Day generation
    while $d < d_{days}$ do
        Generate $\hat{T}_d$ trips according to Gaussian distribution of
        parameters $\overline{T}_{day}, \sigma^2_{T_{day}}$;
        begin Hourly generation
            for $h = 0, 24$ do
                begin Intersection generation
                    for $i = 1, N$ do
                        for $j = 1, N$ do
                            Set $\tau' = \tau$;
                            Generate $t_{ij}$ trips according to
                            Poisson distribution of parameter
                            $\langle t_{ij} \rangle = \hat{T}_d p^{\infty}_{ij} q_h$;
                            begin Trip generation
                                for $t = 1, t_{ij}$ do
                                    Generate a time interval $dt$
                                    according to an Exponential
                                    distribution of parameter
                                    $3600/t_{ij}$;
                                    $\tau' += dt$;
                                    Append $(i, j, t)$ to $\vec{L}$
                                end
                            end
                        end
                    end
                end
                $\tau += 3600$;
            end
        end
        $d += 1$;
    end
end
return $\vec{L}$
```

procedure has been used to assess the impact of taxi sharing in the city of Vienna [5].

10.2 SOLVING THE NON-SPARSITY PROBLEM: OBTAINING NETWORK BACKBONES USING GRAPH FILTERING

The previous example shows how one can exploit temporal stability of trip allocation to effectively reconstruct limited datasets. In this final example, we will try to tackle the opposite problem. We want to extract the most relevant information of our network while reducing as much as possible its numbers of elements.

We have seen that urban taxi mobility Origin-Destination matrices display network features lying *not far away* from the predictions yielded by assuming a null model where the strength of each location is fixed (the MECM studied in detail in Section 6.5). We have further seen that the statistics of each node pair can be very well represented by Poisson statistics (Section 8.1). Combining the two prior facts we can derive a simple filter that allows us to tackle problems that arise by the non sparse nature of the networks under study.

Heavily dense[2] graphs such as OD present important computational and graphical handling challenges to network science practitioners. In particular, extracting qualitative information using data exploration techniques [54] is problematic due to the high number of binary edges present in the data. Also many network algorithms are designed to be efficient in sparse structures and hence their applicability is severely limited when networks possess many binary edges.

The usual form to tackle this problem is to apply different filtering algorithms to the networks under study. A *naive* way to do so is by applying a simple cut-off rule: Just keep the binary edges with occupation number values $\hat{t}_{ij}$ exceeding a given threshold t^{th}. However this has many undesired and uncontrolled effects, since it erases low-occupied binary edges without taking into account their statistical significance. This in turn may alter the topological properties of the networks under study and introduce a typical scale on complex networks which are inherently heterogeneous and whose characteristics usually span multiple scales.

For a very rich discussion on filtering and list of references see [167], including a rigorous comparison with the filter proposed in [162] (see comment in [166]) applied to human migration data.

Many sophisticated variants pivoting around the same principle of filtering applying a cut-off rule have been proposed and shown to be effective in reducing the density on a variety of graphs (most notably human migration flows and trading networks [86]), however, these methodologies are not directly based on statistical plausibility arguments since they do not make any reference to a null model.

Other authors have proposed to follow a different line of attack: Procedures have been proposed which filter out binary edges according to a given null model [162, 143] with a tunable confidence parameter α (associated to a probability) which determines the *severity* of the filter. Their only divergence lies in the choice of null model.

[2] In the context of this section, *dense* loosely refers to a network having a large number of binary edges.

In the two cases mentioned both the binary and *weighted* structure of networks is considered, yet [143] focuses on the observed occupation number distribution at a global scale while [162] puts the emphasis at the local level of the nodes. [162] considers a null model that assumes a uniform allocation of total occupation among the k outgoing (and incoming) binary edges of a node while [143] uses naive bayesian models taking into account a fixed weight and degree distribution.

All the previously mentioned techniques do preserve the heterogeneous scale of the filtered networks and provide a sort of simplified (sparse) *backbone* of a network, with can be handled with more ease, and where network features can be more easily studied. The obvious problem with this approach is that such features heavily depend on the null model considered.

When talking about *network backbones*, it is worth emphasizing that our primary objective in the present case is to obtain a resulting network where the most relevant edges are maintained and which obviously contains distinctively less binary edges and events than the original one. In consequence, our focus is not placed in obtaining a graph *as sparse as possible* where all nodes remain connected, because there is no reason to suppose that some (unimportant) nodes should remain connected to the network at the end of the process. Evidently, what we mean by "most relevant edges" remains to be defined, and in our case, we will identify those binary edges with the ones having exceedingly more and less trips than what would be prescribed by the considered null model.

In our case, we do know that most of the characteristics encountered in urban mobility data are close to those generated by the MECM. Hence, we consider that the binary structure (node degrees) is rather a *consequence* of the strength distribution rather than an additional feature to be taken into account, and consequently the null model must be modified.

10.2.1 *The binary edge Poisson filter*

The proposed filter is inspired by the one proposed in [162] and [143], yet using a different null model and rejection method for the binary edges. Another example of similar filters can be found in [66].

The filtering we propose is based on assuming certain statistics for the node pair occupation numbers t_{ij}. The parameters determining the statistics of the null model may be varied, and in this example I will use the Poisson form corresponding to the non-binary MECM due to the fact that its solution is analytical and it is specially helpful for the case at hand. But it must be emphasized that any statistics derived from the examples of non binary maximum entropy ensembles presented in Part ii can be used for the null model, according to the needs of the practitioner, as long as the complete form of the probability distribution for the occupation of each pair of nodes is known.

For a given confidence level $0 \leqslant \alpha \leqslant 1$ and for every considered node pair ij with expected null model statistics $q_{\text{null}}(t_{ij})$, the filter works as depicted in Figure 10.2-A: It computes the pair of values $(t_{\min}, t_{\max})$ at minimal distance from the mean that fulfil the condition

$$P(t_{\min}, t_{\max}) \equiv \sum_{t_{\min}}^{t_{\max}} q_{\text{null}}(t_{ij}) \geqslant \alpha. \tag{10.5}$$

Then, if the empirical $\hat{t}_{ij}$ lies within the interval $[t_{\min}, t_{\max}]$, it is removed, otherwise it is kept[3]. α in this case can be considered as the complementary of a *p-value*: It is equivalent to the probability to discard a binary edge, assuming that it is generated by statistics corresponding to $p_{\text{null}}(t_{ij})$.

Due to the integer value of the occupation number statistics, the equivalence is not exact for small values of $\langle t_{ij} \rangle|_{\text{null}}$ (which are the most commonly found in real data due to the inherent skewness of the considered distributions). This can be seen in Figure 10.2-B, where we test the filter on network instances produced precisely by the null model. The average empirical rejection probability $\overline{P(t_{\min}, t_{\max})}$ is shown to converge to the expected confidence level α as the sampling is increased (hence Poisson statistics converge to Gaussian statistics). We must additionally take into account that, due to the non smooth integer nature of the statistics, the manner of choosing the values of the bounds $t_{\min}, t_{\max}$ is not unique. In the present case, I have chosen an incremental method starting from the closest integer value t to the expected mean $\langle t_{ij} \rangle|_{\text{null}}$ as described in Algorithm 5.

10.2.2 *Testing the filter on the null model*

The filtering algorithm is implemented in [10] and freely available.

To test the effectiveness of the method, I have applied it first on samples generated using the assumed null model (with self loops), for which,

$$\langle t_{ij} \rangle|_{\text{null}} = \frac{\hat{s}_i^{\text{out}} \hat{s}_j^{\text{in}}}{\hat{T}} \tag{10.6}$$

and I have taken as input the original strength sequences of the NY, SI, SF and VI taxi datasets. The skewness of the occupation number distribution favours small average values of $\hat{t}_{ij}$ (and $\langle t_{ij} \rangle|_{\text{null}}$) and thus α (and also $\overline{P(t_{\min}, t_{\max})}$) can only be considered lower bounds to the empirical fraction of removed binary edges $1 - f_E = 1 - \hat{E}/\hat{T}$ in any practical situation (Figure 10.2-C).

[3] Obviously, variants of the method can be envisaged, for instance, subtracting the contribution of the null model to the remaining edges, i. e.keeping the values $t_{ij} - t_{\min}$ if $t_{ij} < t_{\min}$ and $t_{ij} - t_{\max}$ conversely. However, this procedure leads to signed networks with positive and negative edges, and hence will only be considered for visualization purposes.

Algorithm 5: Graph filtering algorithm for node pair occupation number selection.

```
Input: Empirical occupation number $\hat{t}_{ij}$, expected average
       occupation number $\langle t_{ij} \rangle|_{null}$, confidence level $\alpha$.
Output: Filtered occupation number $t$.
begin Initialization
    Set $p = 0$;
    if $|\langle t_{ij} \rangle|_{null} - \lceil \langle t_{ij} \rangle|_{null} \rceil| \leqslant |\langle t_{ij} \rangle|_{null} - \lfloor \langle t_{ij} \rangle|_{null} \rfloor|$ then
        Set $t = \lceil \langle t_{ij} \rangle|_{null} \rceil$, $t_{min} = t$, $t_{max} = t + 1$;
        Set $k = -1$.
    end
    else
        Set $t = \lfloor \langle t_{ij} \rangle|_{null} \rfloor$, $t_{min} = t - 1$, $t_{max} = t$;
        Set $k = -1$.
    end
end
begin Bound search
    while $p < \alpha$ do
        if $k > 0$ then
            $t = t_{max}$;
            $t_{max} = t_{max} + 1$;
            if $t_{min} > -1$ then
                $k = -k$.
            end
        end
        else
            $t = t_{min}$, $t_{min} = t_{min} - 1$;
            $k = -k$.
        end
        $p = p + p_{null}(t)$.
    end
end
if $t_{min} < 0$ then
    $t_{min} = 0$.
end
begin Selection
    if $\hat{t}_{ij} \in [t_{min}, t_{max}]$ then
        $t = 0$.
    end
    else
        $t = \hat{t}_{ij}$.
    end
end
return $t$
```

A last step prior to applying our filtering method to the four taxi datasets is necessary in order to choose an appropriate confidence level α for each case (cities display wild differences in sampling). Figure 10.2-D shows the average surviving degree ($\bar{k} = \hat{E}/\hat{N}_{>0}$) evolution with $1-\alpha$ after filtering on instances of the null model. We observe a smooth decreasing behaviour only interrupted when the filter returns an empty net (which happens for 3 normal σ in the case of VI and SF, 4σ for SI and 5σ for NY)[4]. Once reached this point, we can be sure that whatever remains after applying the filter to real data contains absolutely no traces of any MECM-like edge[5].

10.2.3 *Testing the filter on the empirical Taxi datasets*

Having validated our filtering algorithm on a synthetic sample and studied its effects, I proceed to apply it to our four taxi datasets. Results of the application of the filtering procedure are displayed in Table 10.2. Following a standard approach, we have chosen confidence levels inspired by z values of a standardized Gaussian distribution. For the sparser datasets (SF and VI) we have chosen a level of three sigmas $\alpha_{3\sigma} = 0.997$, for the SI dataset a level of four sigmas $\alpha_{4\sigma} = 0.99994$ and for the significantly denser dataset of NY a level of five sigmas $\alpha_{5\sigma} = 0.9999994$, which are all displayed as vertical lines in Figure 10.3-D.

Obviously, when applying the filter to real data the decrease in the number of binary edges is by no means as pronounced as for the MECM, yet, a distinctive number of binary edges are removed even at small levels of confidence as shown in Figure 10.3 (50% in the worst case at $\alpha = 0.68$ confidence level), fact which validates our previous insights (Section 8.2) that the observed OD networks display close features to those of a purely random network with fixed strengths. Additionally, even if the decrease in the number of binary edges is pronounced, still a macroscopic part of the network is recovered, as seen by the fraction of remaining connected nodes. Considering the resulting average degree of the filtered datasets, in all cases we obtain $\bar{k} \sim \mathcal{O}(1-100)$ for the confidence level chosen, fact which eases considerably their analysis. Finally, it is worth noting that even if the reduction of binary edges is important, in all cases the minimal value of the occupation number distribution is $\hat{t}_{\text{min}} \simeq 1$, in stark difference with the situation attained would we apply a thresholding filter: In this case, to obtain an analogous reduction on the number of edges we would need to apply a threshold $t_{\text{thres}} > \hat{t}_{\text{min}}$ which would bias

[4] To quantify the severity of the filter, we use the standard σ quantification referring to the α level corresponding to a two-tailed normal gaussian cumulative probability function.

[5] As usual in statistics involving hypothesis testing, the α parameter is somewhat arbitrary. In here we only propose a *rule of thumb* methodology that we seem fit to apply in the general case.

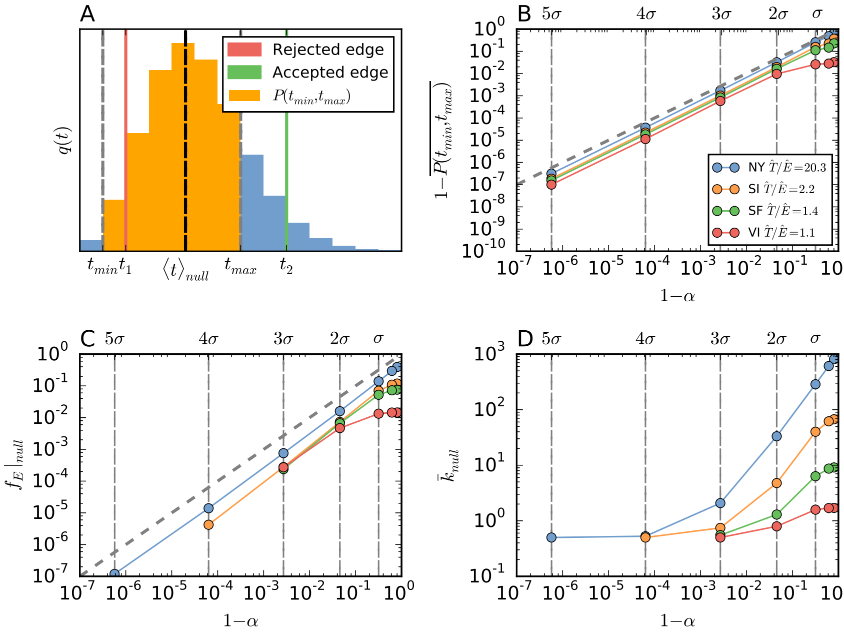

Figure 10.2: **Testing the graph filter on the null model. (A)** Schematic example of filtering procedure: For a given confidence level α, the pair (t_{min}, t_{max}) determines the maximum bounds such that $P(t_{min}, t_{max}) \geqslant \alpha$ for the given null model determined by $\langle t_{ij} \rangle|_{null}$. In this case, the binary edge with t_1 occupation is rejected by the filter while t_2 is accepted. **(B)** Complementary average empirical rejection probability $1 - \overline{P(t_{min}, t_{max})}$ compared to the expected one $1 - \alpha$, as sampling is increased both converge but always $\alpha \leqslant \overline{P(t_{min}, t_{max})}$. **(C)** Empirical acceptance probability (fraction of surviving binary edges f_E) as a function of the complementary confidence level $1 - \alpha$. α can only be considered as a lower bound to the real rejection probability due to the discrete nature of the statistics, the limited sampling and the skewness of the occupation number distribution. **(D)** Choosing the right confidence level: Average surviving degree $\bar{k} = \hat{E}/N_{>0}$ for the filtering applied to null model samples using a single run. Vertical lines marking the different gaussian σ confidence levels have been added for clarity.

DATASET	$1-\alpha$	E	f_E	f_T	f_N	$\hat{t}_{MIN}$	t_{THRES}	OVERLAP
NY	$6 \cdot 10^{-7}$	$6.5 \cdot 10^5$	0.09	0.44	0.99	1	54	0.51
SI	$6 \cdot 10^{-5}$	$1.3 \cdot 10^5$	0.03	0.22	0.85	1	11	0.56
SF	$3 \cdot 10^{-3}$	$4.3 \cdot 10^3$	0.01	0.07	0.32	3	12	0.36
VI	$3 \cdot 10^{-3}$	$5 \cdot 10^3$	0.02	0.07	0.33	3	4	0.67

Table 10.2: **Filtering parameters and global results for the different empirical datasets.** The table shows the fractions of *surviving* binary edges (f_E, also total number of edges shown E), events (f_T) and nodes (f_N) after filtering with confidence level α has been applied. It also show the minimal nonzero value of the observed occupation number sequence after filtering $\hat{t}_{min}$ and the value that would need to be applied in order to obtain a reduction of f_E using a *naive* threshold filter. *Overlap* represents the fraction of surviving edges after applying the filter that would also be present if we applied a *naive* thresholding approach to obtain the same number of edges.

and alter the structure of our empirical network obtaining limited overlap between edges present in one or the other network (see last columns in Table 10.2).

Additional plots showing the effect of the filter on the datasets of SI,SF and VI can be found in Section D.2.

To explore on the effects of the filtering on network features we focus on the case with the higher sampling, that of NY (similar results obtained for the other datasets, see Section D.2). As shown in Figure 10.4, the heterogeneous nature of the networks under study is never lost with increasing values of α. The distribution of occupation numbers (Figure 10.4-E) remains broad and the distribution of node strengths (Figure 10.4-A) remains very stable after the initial change due to the elimination of the majority of *random* trips (which normally display small occupation values, hence the bump that progressively appears in their distribution as the filter becomes more strict). In contrast, macroscopic observables depending on the binary edges such as the node degrees (Figure 10.4-B) or the disparity (Figure 10.4-C) are altered. This causes the network features to progressively lose their MECM-like characteristics and approach those hinted in our previous analysis (Section 8.2): The positive slope of the assortativity profile (Figure 10.4-D) becomes pronounced (small nodes are over-connected among them) and the relation between degree and strength ceases to be linear. One can see, however, that the filter does not alter the node-hierarchy of the network. In Table 10.3, we measure the spearman rank-correlation coefficient between node attributes for the remaining nodes after filtering and their original values, and we see how their strength value ordering is not greatly altered in either case (albeit with different intensity for each city[6]).

[6] Since the number of nodes is not maintained specially for the smaller datasets, the rank comparison in this case becomes difficult.

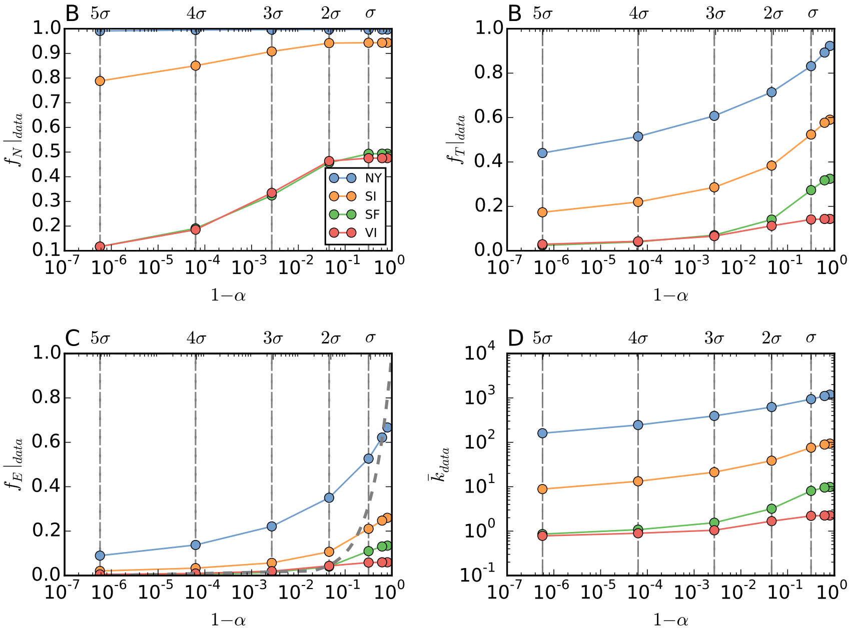

Figure 10.3: **Applying the graph filter on the empirical datasets. (A-C)** Empirical node f_N, binary edge f_E and event f_T acceptance probability (fraction of surviving elements) as a function of the complementary confidence level $1-\alpha$. **(D)** Average surviving degree ($\bar{k} = E/N$) for the empirical datasets. Vertical lines marking the different σ levels and an identity dashed line for the f_E plot have been added for clarity.

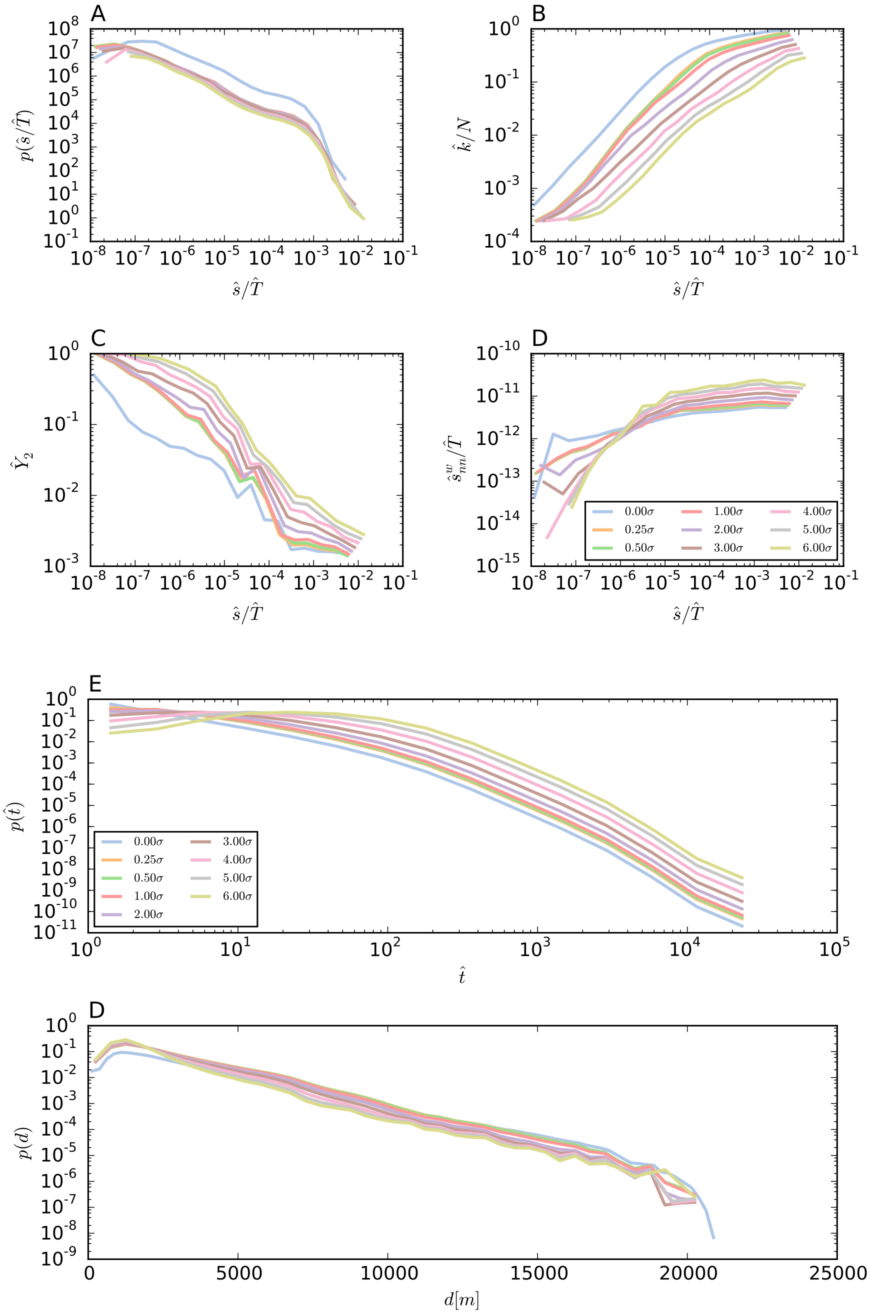

Figure 10.4: **Graph filter sensitivity to confidence level α for the case of NY.** Empirical network features after filtering for different confidence levels α for the case of NY. **A-D** Node related features preserving heterogeneity: The strength distribution (A) and (increasing) assortativity profile (D) remain widely stable after the first level of filtering, while binary edge related magnitudes such as the degree (B) and disparity (C) suffer wider changes. **E,F** Edge related features: The existing trip distance distribution (F) is mostly unaffected by the filter and neither is the tail of the existing occupation number distribution (E).

DATASET	ρ_s	ρ_k	ρ_{Y_2}	$\rho_{s^w_{nn}}$	$\varepsilon_{\bar{d}}$
NY	0.99	0.96	0.85	0.94	-0.21
SI	0.92	0.86	0.77	0.41	-0.21
SF	0.74	0.59	0.55	0.45	0.05
VI	0.55	0.34	0.27	0.26	-0.11

Table 10.3: **Filtering parameters and topological results for the different empirical datasets.** The table shows the spearman correlation ρ between the strengths s, degrees k, disparities Y_2 and average neighbor strength s^w_{nn} of the surviving nodes after filtering and the original data. Only the outgoing direction is shown as results in the incoming direction are quantitatively equal. Also the relative difference $\varepsilon_{\bar{d}} = (\bar{d}_f - \bar{d})/\bar{d}$ for the average trip cost $\bar{d} = \sum_{ij} \hat{t}_{ij} d_{ij} / \sum_{ij} \hat{t}_{ij}$ is shown.

Finally, while the shape of the distribution of trip lengths remains largely unaltered, the average length of the trips significantly decreases (up to a -20%) fact which shows the favour of empirical data towards short trips[7] as seen by the spatially cohesive regions obtained in our earlier modularity analysis (see Section 8.3).

10.2.4 *Exploiting the symmetry of the filter to extract relevant features from the empirical networks: Under and over used trips*

An important feature of our filtering procedure is that by construction it is symmetrical with respect to edge removal, because it uses a two-tailed statistical test. Hence from its result, one may decide to generate the mobility network of under-used trips or conversely the network of unexpectedly (according to the null model) over-used trips. Both pictures provide different information about the transportation phenomenon at study. Obviously, due to the implicit conservation rule in the MECM (the total number of trips and the strength of each node is conserved) both networks are correlated: Over-used trips must be compensated with under-used ones. However, there exist still degrees of freedom for each of the two networks, so studying them separately is still interesting to identify possible sociological factors that inhibit trips among two particular pairs of locations for instance. In general, due to the asymmetry of the Poisson distribution for small occupation numbers, we expect the number of under-used trips to be smaller than that of over-used ones. However, this difference is expected to balance as sampling is increased (as we reach a fully connected network, however difficult this might be). For NY the

[7] The only case where this is not true is SF, but in this case the presence of the airport distinctively isolated from the rest of the nodes, which concentrates a large share of the (unexpected) traffic, compensates this fact.

proportion of under-used trips after filtering relative to the total is 0.26 and it decreases for SI (0.1) and SF (0.1) until it disappears for VI.

From the obtained networks, one may want to classify their edges according to importance. However, with regards to "importance" of a given edge, one may identify it with their "unexpectedness" (according to a null model) or from their resulting occupation. We take normalized z-scores of each surviving edge to quantify the former level of importance while the edge residual occupation $\Delta_t = \hat{t}_{ij} - \langle t_{ij} \rangle_{\text{null}}$ to express the latter.

10.2.4.1 *Under-used trips*

In this case, both the values of z and Δ_t convey similar information as they are very correlated (spearman correlation rank $\rho > 0.7$ in all cases among both variables) while being very broadly distributed (see additional Figure D.10 in Section D.2). Also, contrary to the case of over-used trips, no general direct relation can be made among either indicator and the strength of the nodes at the start and end of these trips for all the datasets. This may indicate that exogenous (sociological, city dependent) factors determine these "anomalies". However, the very existence of this trips with non zero, yet under-expected, values is a noteworthy fact even at small levels of sampling (SF).

10.2.4.2 *Over-used trips*

For over-used trips, however, the z-score and Δ_t metrics provide different information as they are related in a non-trivial way. Focusing on the types of nodes they connect, Figure 10.5 shows the average values of both metrics as function of the origin and destination node original strength (prior to filtering). We observe how the bulk of unexpected trips lie in the diagonal $\hat{s}^{in} \sim \hat{s}^{out}$ (due to the assortativity of the network), with trips among small nodes favouring large z-scores while trips among top-visited locations having large residuals. Large z-scored edges correspond to random taxi trips taken by users for particular, non-generalizable purposes. It remains to be seen if this circumstance is only particular to taxi modes of transport or could be detected in other source of mobility data, but these kind of trips would correspond to the detected "random" trips mentioned in [118] (see discussion in Section 8.2), whose contribution grows with city size (population and thus sampling) with respect to the other kinds of flows. Edges with large residuals connect mainly hubs, but also pairs of intermediate locations display large number of residuals (termed "integrated" flows in [118]), which are probably caused by sociological factors which cannot be accounted for using a simple tconstrained framework like the entropic models earlier proposed.

The overall observations may lead the reader to believe the existence of a clear anti-correlation among both metrics, however, their

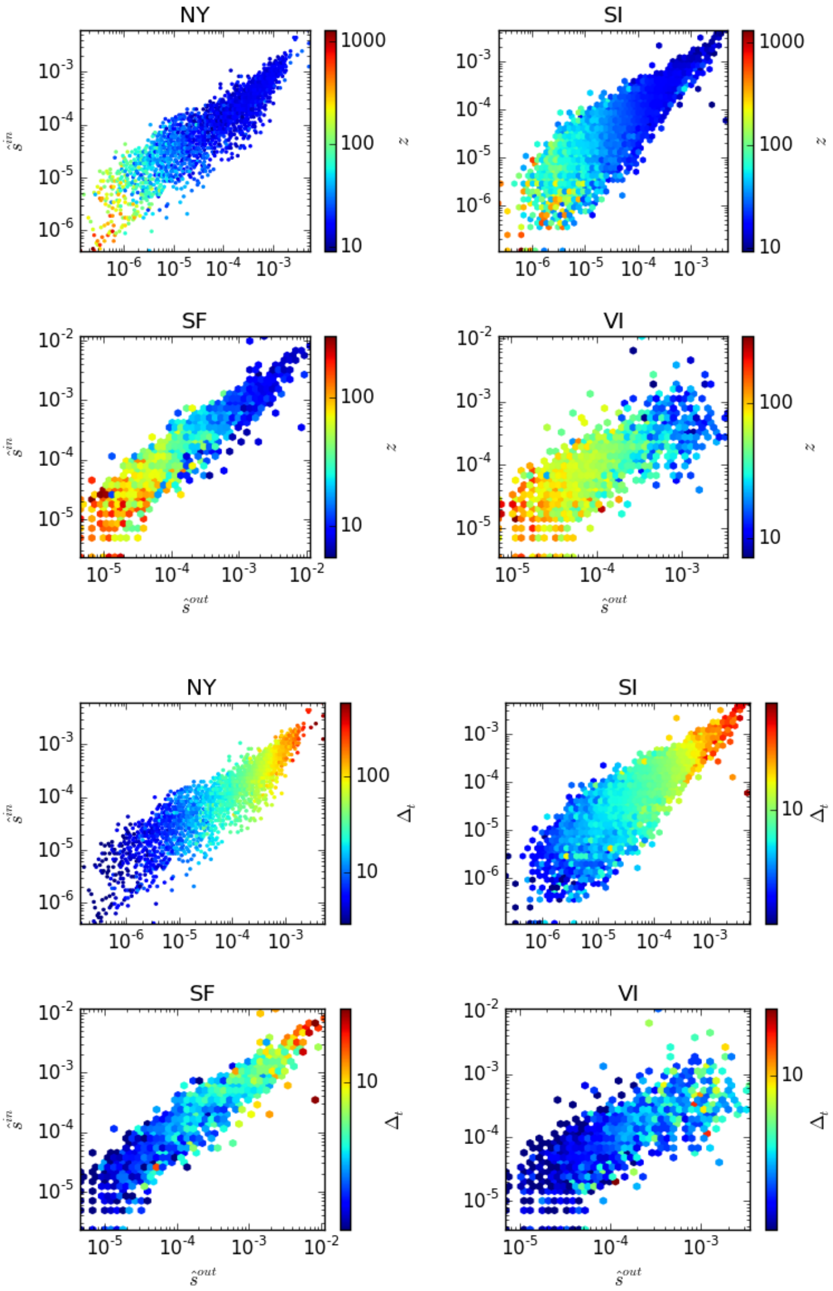

Figure 10.5: **Over-used trips z-scores and residuals allocation.** Average z-score (top) and residual (bottom) as function of the non-filtered strength of the origin and destination nodes whom they connect is shown (bottom). The fact that hubs tend to be connected by large residuals while scattered locations display large z-score values is apparent.

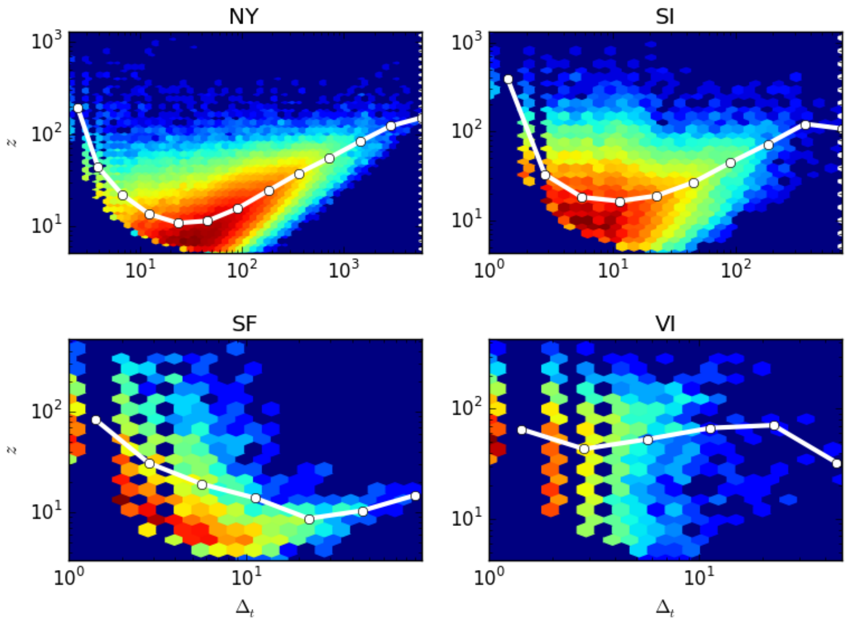

Figure 10.6: **Over-used trips z-scores and residuals correlation.** Correlation among z-scores and residuals Δ_t of edges surviving after filtering the empirical datasets. Overprinted white lines correspond to average values using log-binning.

relation is more complicated and non-monotonous, see Figure 10.6. For cities with large sampling, among the trips with the largest residuals we find also some of the trips with the largest z-scores.

To finish this section, we proceed to show a visual representation of the filtered networks under study. Even if the filter is successful at obtaining the *backbones* of the considered networks, visualizing graphs with $\mathcal{O}(10^3)$ binary edges is a complicated endeavour. For this reason, and for visualization purposes, we have applied first the proposed filters with the chosen confidence levels to each of the subgraphs of over-used trips and only then have we applied a thresholding procedure keeping only the top 100 binary edges with largest residual Δ_t. The result is shown in Figure 10.7. It must be stressed once again that even choosing such a small subgroup of edges, the overlap between doing so choosing edges with the largest residuals compared to the edges with the top occupation (choice that would be made using a standard thresholding approach) is only partial (0.77 NY, 0.84 SI, 0.68 SF and 0.92 for VI). Visualizations are shown in Figure 10.7.

Obtaining informative, unbiased network visualizations is complicated and represents and active area of research in the growing field of Data-visualization [54].

Obtaining sound and unbiased conclusions from network visualization is a complicated issue because one needs to carefully choose what to represent. In the present case, my intention is focused on highlighting several aspects of the networks under study: We can observe that the strength distribution is widely distributed (node size is proportional to outgoing strength). Airports concentrate a large por-

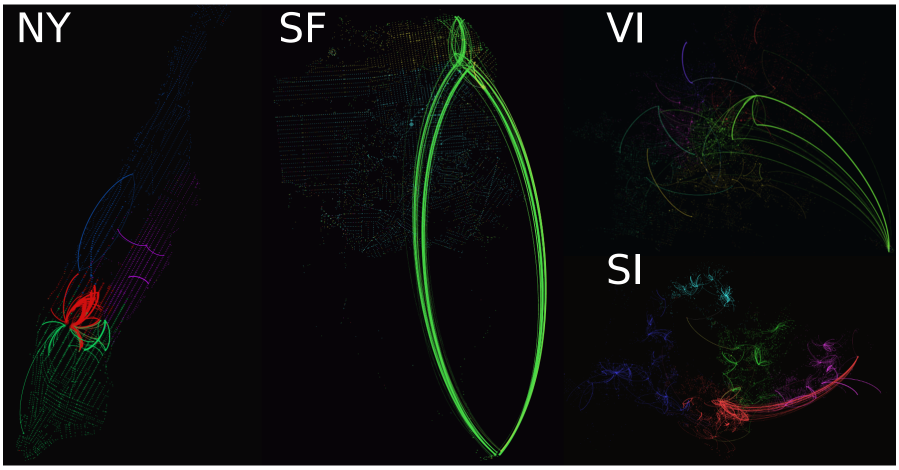

Figure 10.7: **Network representation of the four filtered datasets.** Network representation displaying the top 100 edges with top residuals Δ_t (over-used trips). Edge thickness is proportional to residual value and edge colors is a blending of the colors of parent and child nodes being connected, node size to outgoing strength (original network) and node color to community assigned by modularity maximization on the original network. Nodes are placed according to their original geolocalized coordinates.

tion of the total traffic but now large nodes appear not only concentrated around the city center. This is specially evident in the case of Singapore (and also Vienna), where a *polycentric* [**?**] structure (nodes distributing trips among close locations which in turn are connected to other larger nodes and the city centre) is apparent. We can also further see that setting apart airports, the allocation of highly used binary edges seems to be really dependent on context specific properties of the nodes (sociological importance) and hence no apparent pattern with relation to node strength nor distance is visually appreciated. This is one of the main reasons why *gravity-like* inspired models (partially) fail when applied to urban environments, as already pointed out in Chapter 9.

As a final remark, we must acknowledge the difficulty of assessing the effectiveness of a filtering procedure on a given dataset. By construction, any procedure will provide output results, whose plausibility cannot be evaluated using a non-arbitrary criterion. The effectiveness of the filtering method must be instead checked with external sources of data and knowledge from the studied process.

In the present section, we have provided an example of application of the knowledge developed in previous parts of this thesis regarding the underlying statistical structure of the considered datasets and their associated null models. Although I have particularized to a given example, the applicability of the method is general and compatible with any of the models studied in Part ii. We here provide statistical-based procedure which, in contrast to *thresholding* methods, allows to detect and quantify statistically relevant data features using a starting hypothesis (null model) based on a maximum entropy principle. We thus aim to contribute to solve the problem of assessing the impact that data thresholding techniques [78] have on complicated structures and multi-dimensional datasets represented using graphs.

10.3 WRAPPING UP: APPLICATIONS OF MAXIMUM ENTROPY MODELS TO KNOWLEDGE EXTRACTION FROM EMPIRICAL DATA

In this final section of the thesis we have shown two particular applications of the theory developed combined with the data-exploration insights obtained in previous chapters. Others could be devised, but in this particular case we have tackled two problems that have also recurrently mentioned in the text, specially in the introductory part concerning data analysis (Chapter 7).

A. **Sampling:** We have proposed and implemented a procedure that exploits the temporal stability of inter-node statistics $\left\{p_{ij}^{\infty}\right\}$ (shown in Chapter 8) together with the high effectiveness of Wilson's doubly constrained exponential gravity model to model urban mobility (discussed in Chapter 9) to reconstruct a picture of a mobility process using only limited samples of aggregated data. The procedure has been tested on the NY dataset (for which a *ground truth* is available since all yearly trips are recorded) and its effectiveness shown up to extreme levels of subsampling. Furthermore, an extension to apply what we call *supersampling procedure* to richer data sets where temporal information is available has been proposed and the limitations of the method have been discussed.

B. **Feature extraction and network visualization:** Inspired by the existing literature on graph-trimming, we have developed a filtering procedure that allows to separate the contribution of *random* trips given by the structure of the city (or any of the entropic null models developed in Part ii) from the distinctive, *non-random* unexpected features of urban mobility. Given that, as we had seen in Section 8.2, an important part of taxi trips can be explained by the city layout, the application of such a filter using as base the MECM model is highly effective. On the one hand it allows to significantly reduce the number of trips and binary edges of the networks, thus easing its analysis in terms of computation costs. On the other hand, it allows to extract features of the four considered datasets and to study separately the under-expected and over-expected trips. The filter has been used visualize the datasets.

The success of both our reconstruction and simplification methods, even using very small amounts of data, points out the composite structure of the network of urban mobility: Taxi displacements are characterized by a core of unexpected trips coupled with trajectories generated at random but conditioned by the structural constraints of the city such as population distribution and mobility costs. At the same time, the core of unexpected trips can be split among a non-negligible groups of over-expressed trips among scattered locations generated by particular user necessities distributed randomly in the spatial layout of the city and a group of edges with large occupation connecting preferably hubs (either local or global) and reinforcing the assortative nature of the network of displacements.

This chapter provides complementary tools to those used in Chapter 8 for the study of mobility networks and closes the part of the thesis devoted to the analysis of empirical data. Analyzing non-binary network structures is a complicated task, but the methods here used are fully general and could be extended to many different cases of non binary network studies (not necessarily related to mobility). The addition of entropic models based on constraint hypothesis helps in extracting knowledge from such complicated structures. Also, the flexibility of our modelling framework allows to focus on testing different aspects of the topology and also to visualize them in an informative manner.

Finally, the implications concerning the success of the applied supersampling methodology are two-fold: On the one hand, for the particular case of urban mobility, the stationarity of the temporal patterns could be exploited to save space and effort in recording mobility data. On the other hand, our method opens the possibility of efficiently scale up data from reduced fleet of vehicles in cases where a full knowledge of the system is needed.

Part IV

CONCLUSIONS

A research never truly ends, it is just merely delimited in time due to personal constraints. Nevertheless, for any work done, lessons must be extracted from it and an evaluation on what has and has not been achieved must be performed. This final part recapitulates all the research done in this thesis, reviews possible criticism and sketches future interesting research directions, based on my personal opinion and experience.

11 CONCLUSIONS AND PERSPECTIVES

One never notices what has been done; one can only see what remains to be done.

— Marie Sklodowska Curie [58]

11.1 OBJECTIVE REVIEW AND CONTRIBUTIONS

The analysis of urban mobility data has been our case study to exemplify the possibilities that network science can offer with regards to data analysis. More particularly, we have shown that non binary networks, which are particular types of graph where the interactions among elements are graded, provides an interesting framework to do so.

However, by choosing such tools, we have quickly realized how non binary networks are complicated structures to analyze, as their degrees of freedom are considerately larger than those of binary networks (where interactions among elements are dichotomic). In the first part of this work, our aim has thus been to provide tools to study them in a systematic and consistent manner.

In order to create recipes to guide this systematic approach, we have started our study focusing on theoretical foundations for the generation of non binary network models with flexible prescribed properties. The objective is to use them to generate null models mimicking properties of real data, to assess the structural impact (and possible causality) these properties have on other measurable metrics of the system under study.

Our work has first placed into context our research, circumscribing it on non-sparse networks with integer weights. For these structures, the need for defining asymptotic limits and a relevant scaling variable different to what is commonly used for ensemble descriptions of sparse binary networks has been discussed. For the case at hand, we have set a convenient scaling variable related to the number of recorded events used to connect nodes.

Drawing inspiration from classical statistical mechanics, we have argued that using a maximum entropy principle to tackle this challenge is a promising approach, yet important considerations must be made about the structure of the systems under study in order to allow for a precise ensemble treatment of the problem.

We have classified non binary networks which can be obtained by aggregation of different layers of information into three distinct general groups, depending on the distinguishability of the events forming them (Multi Edge, Weighted and Binary networks). This taxonomy naturally includes in our framework previous work done by other authors and allows to clearly establish differences among their starting hypotheses.

For all the considered cases, we have explored ensembles with hard and soft constraints, circumscribing our analysis to a particular (yet quite usual) form for the considered constraints. This analytical study has lead us to realize that given our starting hypotheses and our definition of asymptotic limits, the only types of ensembles likely to be observed in real data would be the ones composed by distinguishable elements (ME case). Furthermore, this case is the only one with well defined asymptotic properties and for which the treatment for both soft and hard constraints leads to equivalent results. Even for this ME case, the only total harmony among ensembles is obtained for constraints expressed as linear forms of the occupation of events in each state (pair of nodes), which is the case that can be naturally related to the Maxwell-Boltzmann statistics studied in classical physical systems.

Motivated by this fact, we have explored the similitude and differences among network ensembles and the corresponding usual equilibrium statistical mechanics ones. In doing so, we have derived the Maxwell-Boltzmann statistics of occupation of discrete energy levels for the Grand Canonical ensemble in a completely consistent way from the point of view of particle distinguishability. Also, we have identified two different (yet related) conditions that clarify the concept of *ensemble* equivalence among soft and hard constrained ensembles. We have seen how equivalence of asymptotic ensemble entropies is a weaker condition than strict vanishing of constraints fluctuations relative to the scaling variable for the soft constrained ensemble. In this sense, we have shown how for the ME case with constraints depending on the binary occupation of events in each state the former condition is met while the later is not, while the ME case with linear constraints meets both conditions and thus is the only situation that leads to strict ensemble equivalence among the cases studied in this thesis.

We have consequently focused on this specific case, where we have extended the previous work by other authors into a completely tunable null model that includes many constraints that may be of interest in real world applications. These include fixing event cost statistics, strength distribution, node strength correlations, community structure, occupation number distribution and also quantities depending on the binary structured of the networks.

Once the limits of our theoretical work have been explored, and with practical applications in mind, we have proceeded to put it in practice in order to generate samples of our proposed models for the three earlier mentioned ME, W and B cases. In this sense, we have provided all necessary ingredients to simulate the variety of null models studied and additionally openly made accessible a software package to do so for a majority of the considered examples of constraints. We have further provided an effective way to obtain analytic expectations for network related magnitudes to complement those obtained by simulation, and we have made use of them to experimentally test our prediction of ensemble equivalence for the Multi Edge case with linear constraints. In order to highlight the importance of choosing the right model for each system, we have studied the particular ensemble of non binary networks where the strength of each node is fixed. This case has also served to study the strong impact that fixing this feature, which tends to be heterogeneously distributed in empirically observed networks derived from human activities, has on overall topological indicators.

With the appropriate set of tools developed in the first part of this thesis, we have faced the analysis of urban mobility data. To begin, a general overview on the available data sources (taxi displacements in New York, Singapore, San Francisco and Vienna) has been made, and we have discussed their limitations, as well as the filtering procedure used in each case to *clean* them. We have performed a general analysis on the temporal aspects of data to justify the aggregation of all temporal data in a static picture, thus disregarding dynamic effects. The taxi trips have been shown to suffer from hourly, weekday and seasonal fluctuations (in decreasing order of importance), yet we have argued that, if sufficient data were available, a detailed analysis using the same proposed methodologies could be done particularizing to each case (on a hourly, week-day, month, etc... basis). The analysis of these effects has shown that they are common to the four studied datasets. We have then defended that the diverging geographical layouts of the different cities studied justify the network approach taken to study urban mobility to try and isolate geographical-dependent related factors from overall common, "universal", traits.

Taking the non binary network point of view, we have confirmed that datasets of taxi displacements among locations display statistical properties compatible with a description in terms of maximum entropy Multi Edge network ensembles with linear constraints, assuming a fixed (quenched) quantity p_{ij}^{∞} that characterises each intersection pair ij. The distribution of urban activities, encoded in the observed trips emerging and entering each node, is highly heterogeneous in the four studied datasets. In order to assess their influence on the observed mobility, we have made use of the earlier studied Multi Edge configuration model (MECM). We have found that the concentration of activities in the urban context plays an important role in shaping the observed mobility, albeit distances and an array of sociological factors account also for variations among empirical data and the MECM. This study has allowed us to uncover common network mobility features present in the four studied datasets and also relating them to others analyzing different, yet similar, sources of data. Those common features include exponential decay of trip lengths, heterogeneous strength distribution, moderately assortative profile of node strengths, large tails in the distribution of trip occurrences, over-expression of trips among low strength valued nodes and also the presence of spatially cohesive communities of over-connected nodes.

Once the main topological characteristics of empirical datasets have been studied, we have performed an exhaustive review of existing proposed models of urban mobility from a theoretical point of view, discussing also issues related to their practical implementation to the urban context. We have implemented the models and tested their quality to reproduce taxi urban mobility. To do so, we have first presented indicators commonly used to assess their quality and we have also proposed new ones aiming to provide a complete view on model performance with the addition of entropy measures and network topology related indicators. The practical implementation of the models has allowed us to show ME entropic models as the best performing ones, balancing explanatory power, ease of use and analytical treatment. This procedure has also made patent that despite the usually large amounts of data available for these kind of studies, the non-sparse nature of the studied networks provokes a strong dependence of model indicators on sampling, specially (and sadly) those related with entropic quantities studied in the earlier part of the thesis. This fact is aggravated by the heterogeneity of the distribution of occupation number values (trips among locations), which justifies the need of simulation and benchmark measures in every empirical study to control for sampling effects.

We have then focused on analyzing the best performing model, which has been Wilson's original formulation of the gravity law of transportation, a maximum entropy model enforcing the total strength of each node and the average perceived cost of trips, assumed to be proportional to euclidean distances. Despite capturing to a good extent the urban mobility process, even this constrained model cannot fully account for the observed patterns of connections, highlighting the limitation of simple models to capture complicated, human-dependent activities. The main sources of discrepancy for all datasets being the unexpected long tail of the occupation number distribution (some particular intersection pairs ij displaying very large occupation values) and the over-represented weight of trips connecting nodes with low values of strength.

Summing up all the above observations, in the last chapter of this thesis we have tried to produce useful applications to apply to the field of urban mobility problems.

Taking into account the ability of a strength-constrained and cost constrained model to correctly describe the majority of trips, combined with the stability of intersection pair ij specific values (independent of the sampling) and the flexibility of the framework earlier developed, we have proposed a partially data-driven model that is highly effective at reconstructing mobility processes from limited samples. This model allows to reconstruct a mobility process from (very) limited samples, and has been tested on the NY dataset, exemplifying how it could be used to generate traffic predictions based on historical or partial data.

We have also tackled the problem of extracting *noise* from observations of urban mobility datasets by devising a filtering methodology, inspired in previous work by other authors but making use of our developed null models. Such a filter can be used to subtract from empirical data the contribution of any of the earlier developed null models. In particular, given the large contribution detected by the strength distribution in explaining the observed mobility patterns, we have exemplified its use on our four datasets taking as example the MECM. As a novelty, the ability to produce samples of our proposed null models has allowed us to study the effect of the filter on random instances of the model, which has helped us in tuning the parameters to use when applying it to the empirical datasets. The application of the filter has confirmed our earlier characterization of the empirical networks. Additionally, being our filtering procedure symmetric, it has further allowed us to divide the remaining trips in over-used and under-used trajectories, which can be then studied separately. We have produced also visualizations of the datasets to exemplify how the proposed filter methodology, based on statistical arguments and an underlying null model, is different to thresholding procedures.

To conclude, and connecting with the objectives set at the beginning of this journey, and with the (artificial) debate exposed in the introduction between empirical knowledge, theory and data it is worth summarizing the conclusions drawn from our work. First of all, even if our datasets are very large, we have seen with the help of an appropriate theory, how, paradoxically, they are still not dense enough to provide complete information about the mobility process. At the same time, our theory, as flexible as it can be, can never fully account for the observed phenomenological patterns, being them dependent on a large variety of uncontrolled social factors. Yet, the merging of both in a symbiotic cycle can provide us with useful tools to apply in real world situations, as exemplified in the last chapter of this thesis.

11.2 POSSIBLE CRITICISM AND FUTURE RESEARCH PERSPECTIVES

Before concluding this work, it is a healthy exercise to ask ourselves about the limitations of the study just presented and put forward new ideas for future research.

Concerning the first part of the Thesis, the obvious critique one can make is that the framework presented only explores a limited subset of all the possible maximum entropy non binary network ensembles. The techniques here proposed, while general, can only fully be exploited for moderately to large graphs if one is able to analytically sum the involved partition functions and their associated saddle point equations (which we have only been able to do for some cases of linear compositions of state ij independent functions). Also, as we have seen, considering more involved types of aggregation mechanisms to generate networks would imply the calculation of more complicated degeneracy factors[1]. Despite all this, exploring in this framework novel degeneracy terms (to study different aggregation network techniques for instance) is an interesting perspective.

Another critique one may do is that no completely new techniques are presented or developed. While this is partially true, and I have mainly adapted the methodologies from other problems (notably binary network ensembles) and the earlier (partial) work of other authors on the field (most prominently Wilson), the global structure given here has allowed to formalize and expand the treatment of the problem, while unifying and most specially clarifying earlier work under a common framework. The most relevant example of this is the identification of several types of non binary networks and the study of their sharp differences even under the same set of constraints. Another important aspect that has been clarified are the conditions for equivalence among hard and soft constrained examples.

A discussion on the issue of continuous weights is sketched in Chapter 3.

Concerning future research from the point of view of stablishing the theoretical foundations of non binary network ensembles, I feel that two important and interesting problems are worth facing: That of the separation between binary and non binary network structure and that of continuous weights. Both issues are clearly joined and in a nutshell facing them amounts to answering two questions. Firstly, *under which (quantifiable) conditions can we separate (uncorrelate) the binary and non binary structure of a graph?* As we have seen, solving this question would allow for the development of a new range of maximum entropy network ensembles with scaling parameters N (number of nodes) and T (events). Secondly, *is there a natural way of discretizing weights?* In order to expand our theory to networks with continuous weights, we must search for a graph equivalent of Planck's constant

[1] A prominent example we tried to explore was that of book randomization: If one creates networks from a text by joining consecutive words and wants to randomize it, one must take as degeneracy a quantity known as chromatic number [102], whose calculation is already a complicated endeavour (it is a NP problem).

h, while exploring in detail the consequences of thresholding connections due to resolution limits of our data recording devices.

Finally, I feel that more research is needed to understand and establish a range of null models that serve as benchmarks for testing proposed models on non binary networks related phenomena. In every study concerned with the analysis of a topology using empirical data, the use of an appropriate null model is imperative in order to separate correlation from causality provoked by the structure itself. While randomization might be seen as sufficient tool, further theoretical work is needed to understand its relation to analytical models and their possible biases (a variety of rewiring schemas exist). Also, analytical models are important as their flexibility allows to develop a range of applications for data analysis that purely computational methods cannot provide. An interesting expansion for our framework would be to obtain fully closed analytic solutions for the Γ entropies of ME using a micro-canonical formalism for the cases of fixed strength and strength correlations, by application of similar tools as those in [21], that would also allow the study of event switching rewiring algorithms for the case of non-binary networks [145], which would possibly allow to explore the Micro Canonical ensemble in an unbiased manner.

The limitations of the dataset have been discussed in Chapter 7.

We now focus on the second part of this work. In the first place, possible concerns can be raised with regards to data issues about the "universality" of the conclusions drawn from four limited datasets of taxi displacements. Those are totally valid concerns, furthermore considering that the datasets have been filtered using a particular procedure and that their sampling is limited. However, many common traits have been identified with other studies using other types of data (mainly call detail records and geolocalized data from online social media). For this reason, future research would be to analyze different datasets (on urban mobility) representing other modes of transports to establish what are common traits and what are not with regards to taxis using our systematic approach. That would also allow to confirm to which extent our detected features are "universal" (we have seen already that differences are observed across the studied datasets). In particular, an interesting corpus of data to consider is that of shared bicycle systems[2] since studying the effect of cost relating distance and physical geography (terrain steepness) would be an interesting project. Along this lines, studying the performance of our models using progressively larger areas, starting at the urban level and going to regional and national level would allow to investigate further on the relation between distance and user cost perception.

Closing the data related issues, an important (unresolved) problem that must be faced is that of the MAUP. We must study and try to obtain a non-arbitrary methodology to aggregate continuous points

[2] See for instance https://en.wikipedia.org/wiki/Bicycle-sharing_system.

into discrete structures for the case of networks, specially for the urban context.

The limitations of existing mobility models have been discussed in Chapter 9.

From the urban mobility modelling point of view, two main research directions can be sighted. On the one hand, we have seen that pairs of trips among remote locations are over-represented and these could be added in the form of stochastic perturbations to the constrained models proposed in this thesis. Another improvement would be the possibility of studying in depth non-isotropic opportunity based models or cost perception (where users focus not on radial areas for choice of destination but in arbitrary ones). Finally, exploration of embedding of mobility networks into hyperbolic spaces using simple rules would be a good idea to try and detect possible social/infrastructural system weaknesses inhibiting traffic among certain locations.

A synthetic example of possible application of multiplex detection is shown in Section D.5.

For the applications part, an interesting project would be to try and find datasets behaving approximately like the studied W or B cases[3], which would allow for the detection of possible truly multiplex maximum entropy structures and to infer their number of layers M and/or their typical structure and fluctuations[4]. Concerning the urban mobility aspects, our data-driven proposed model could be implemented for the design of efficient (predictive) taxi dispatching systems.

To conclude, I must emphasize that this work is incomplete as long as the tools hereby developed are not used in conjunction with other researchers to extract conclusions about the urban environment. To understand the coupling among the city layout, cost perception, land used and psychological factors the participation of a diverse team of individuals is called for. Hence, an obvious research perspective is to devote efforts to publicise the work done and to prompt other disciplines in using these models, indicators, applications and software tools, and also to raise objections to the statements drawn from my analysis of the data.

11.3 CLOSING PERSONAL REMARKS

A PhD dissertation is not only a research process but in general also serves as an introduction to the scientific world for young researchers. This has indeed been my case, and I cannot conclude this work without stating some personal (non strictly scientific) opinions derived from my experiences in this time in the Complexity Science community.

[3] We have argued that due to the mathematical properties displayed by the W case, it is unlikely that empirical observations may exhibit these properties, yet, one cannot be completely sure whether this model would help to describe a particular dataset.

[4] This is specially interesting for the case of neural studies for example, where different samples of the same process are available and could serve in characterizing an *average healthy brain* model and its associated ensemble [186].

A first philosophical question is related to the fuss on data related studies. While there is little doubt that large quantities of data have enriched our understanding and tools to explore problems from an empirical point of view, I feel little discussion is devoted to the social implications about data ownership and the limitations it imposes to scientific research. Big data is held by private corporations, and in general, can only be accessed by selected research groups for particular questions, casting doubts on reproducibility of obtained results and stating a "de facto" hierarchical structure in the research community. We should not renounce to gain knowledge from these platforms, but the scientific community should neither resign from the important role it can have as policy actor in leading this debate and raising questions about the social utility, perils and implications that private tracking of personal activities has on the society.

A possible alternative way to gather social data for scientific studies is studied and discussed in [9], and a practical example is implemented in [6].

Another philosophical issue is related to confirmation bias and the never ending race of scientific publishing. Not only must we ask where the data we use comes from and whether our findings will be reproducible, but also our attention must be doubled with respect to empirically obtained results. What helps fellow researchers is not only good story-telling and communication but also transparency. More discussion should be devoted on what works and what does not, on limitations as well as strengths of models, on availability of implemented software and on good and constructive peer-reviewing process. All in all, a certain lowering of expectations is needed: Charting unknown territory is very difficult, so doing it collectively should make it simpler.

All this requires time, sure, but also would imply a reduction in the number of published papers. And this is good news, for it will ease literature research and thus reduce the problem and fear of constant rediscovery of models[5] (it will also simplify the assessment of research quality). It also requires enforcing interdisciplinary work: A real dialogue among open minded scientists from different disciplines to share tools and merge methodologies under a common framework. Funding institutions mention recurrently this idea, but in practice very few incentives are laid for interdisciplinary research: Presenting results in *boundary* fields is complicated, takes time, a lot of effort and is commonly received with scepticism, but is the only way in tackling many important upcoming research challenges.

Notwithstanding all of the above, many interesting initiatives, tools, methodologies, reviewing structures, data and code sharing platforms and practices are emerging. And obviously I do not mean to question the fact that excellent research is being produced under the current framework, yet I feel that being science founded in the (positive) crit-

[5] In this sense, I have tried to cite all the material from which I have drawn inspiration for this work. Given that the problem under study has been studied from many fields, I apologize in advance if I have missed important references unbeknown to me.

ical spirit of people, these and other political related questions concerning research should be voiced and discussed in the open, specially by researchers in early-stages of their careers.

12 RESUM DE LA TESI EN CATALÀ

L'idioma és la columna vertebral d'una cultura, l'únic instrument [...] que [n']assegura la seva preservació, [...] una experiència sense repetició possible. [...] Veiem el món amb els nostres ulls i l'interpretem amb la nostra parla.

— Manuel de Pedrolo [64]

12.1 INTRODUCCIÓ

L'abaratiment i popularització de tecnologies de geolocalització d'alta precisió han propiciat l'aparició, en poc temps, d'enormes bases de dades a escales temporals i espacials molt precises que permeten l'estudi de dinàmiques de moviment de persones en molts ambients. Això possibilita l'anàlisi de fenòmens de mobilitat en entorns delimitats i densos con les ciutats.

Típicament, l'anàlisi de la mobilitat se centra en la representació de matrius d'origen i destinació (OD) que recullen els trajectes recollits entre totes les parelles de destinacions que hi ha presents a les dades. Aquestes matrius, es poden considerar un exemple de xarxes complexes on cada localització és un node que està connectat a d'altres nodes amb enllaços, que es graduen d'acord al nombre de trajectes enregistrats entre les diferents destinacions.

L'estudi general de les anomenades "xarxes complexes" ha experimentat un gran auge en els darrers temps. Aquesta popularitat, es deu, en part, al fet que les estructures de xarxa són intuïtivament simples de representar, tractar i analitzar. Majoritàriament, els estudis dedicats a xarxes s'han centrat en aquelles més simples per les quals les interaccions entre nodes es poden codificar de manera dicotòmica: O bé hi ha interacció i s'estableix un enllaç (de valor 1) o no n'hi ha, amb la conseqüent manca d'enllaç (valor 0).

Una part important de l'estudi de xarxes complexes rau en desenvolupar algoritmes i tractaments que ens permetin generar models amb propietats prefixades. Aquests models es poden llavors emprar per comparar si la generació d'estructures mantenint certes propietats duen a prediccions que s'adiuen a dades empíriques. Per les xarxes binàries, aquests tipus de models estan molt ben treballats. Per al cas de xarxes pesades (com les OD de mobilitat), múltiples desafiaments resten sense resoldre a l'hora d'enfocar el seu anàlisi ja que la inclusió d'una nova "dimensió" quantificant la fortalesa dels enllaços en complica l'aparent simplicitat. En particular, la principal necessitat per tot estudi empíric basat en xarxes pesades és la disposar de models nuls flexibles que permetin aïllar els diferents factors que afecten les observacions d'una manera sistemàtica.

El nostre objectiu en la primera part d'aquesta tesi ha estat el de desenvolupar un marc teòric que permeti l'anàlisi d'aquestes estructures de manera sistemàtica per poder analitzar la influència dels diferents possibles factors que determinen la observació d'un fenomen que es pugui modelitzar amb una xarxa pesada. Tot seguit, en la segona part, apliquem la teoria desenvolupada per exemplificar-ne l'ús en un estudi de mobilitat en entorns urbans emprant dades de desplaçaments en taxi per 4 ciutats diferents.

12.2 TEORIA DE COL·LECTIVITATS APLICADA A XARXES NO-BINÀRIES

En aquesta part de la tesi, emprant eines manllevades de la mecànica estadística de sistemes físics en equilibri, apliquem un principi de màxima entropia per a resoldre el problema matemàtic de generar xarxes amb certes propietats prefixades.

12.2.1 *Caracterització general del marc teòric*

En primer lloc, analitzem les diferents maneres que hi ha d'entendre una xarxa pesada d'acord amb el mecanisme darrere de la seva construcció. En particular, identifiquem 3 possibles casos rellevants en funció o no de la distingibilitat dels esdeveniments que formen la xarxa (que anomenem xarxes pesades -W-, agregades -B- o multienllaç -ME-) i n'estudiem les diferencies. També establim de manera precisa el problema matemàtic a resoldre, els seus límits asimptòtics i les possibles discrepàncies respecte la formulació mecanico-estadística usual.

Per cadascun dels tres casos, considerem dos tipus de col·lectivitats, la Micro-canònica (MC - on totes les propietats que volem fixar es compleixen de manera exacta) i la Gran Canònica (GC - on les propietats de les xarxes obtingudes es mantenen només en promig sobre la col·lectivitat). En particular, estudiem els casos on les propietats prefixades depenen de funcions lineals del pes dels enllaços entre els nodes o de funcions que depenen de l'estructura dicotòmica de la xarxa. De tots els casos estudiats, constatem que sols un, aquell on els esdeveniments que formen la xarxa són completament distingibles (ME), té propietats matemàtiques convenients per als límits asimptòtics establerts. Aquest cas es correspon de manera natural a la formulació per estudiar processos de mobilitat.

12.2.2 *Generació pràctica de xarxes*

Emprant la teoria desenvolupada, hem creat eines teòriques, computacionals i de programari obert per a generar xarxes pesades dels tres tipus estudiats anteriorment, tot exemplificant-ne l'ús per a una sèrie de casos concrets. En particular, desenvolupem un estudi complert del principal model que usarem per establir la influència de l'estructura urbana sobre els fenòmens de mobilitat, el model configuracional no binari amb múltiples connexions (MECM). Per a aquest model, exemplifiquem com seguint la nostra teoria hom pot generar prediccions exactes per observables sense necessitat de simulació i demostrem la predicció teòrica de com el seu tractament emprant els dos tipus de col·lectivitats estudiades (GC i MC) porta a resultats equivalents.

12.3 anàlisi de xarxes de mobilitat urbana

Amb les eines teòriques desenvolupades en la tesi, procedim a encarar l'estudi de sets de dades de mobilitat humana en entorns urbans. Per fer-ho, explorem els possibles desafiaments que planteja l'anàlisi de dades empíriques d'aquests fenòmens a gran escala, el possible paper que hi poden jugar els models nuls i les eines associades que es poden desenvolupar. També realitzem un estudi crític dels principals models de generació de prediccions de mobilitat existents i la limitació de la seva aplicació a l'entorn urbà.

12.3.1 *Característiques rellevants de les xarxes empíriques de mobilitat*

En primer lloc presentem les dades emprades, corresponents a desplaçaments en taxi per 4 ciutats (Nova York, Viena, San Francisco i Singapur) i en discutim les fortaleses i limitacions, així com el procediment emprat per filtrar-les. Tot seguit, les representem mitjançant xarxes pesades OD i mostrem com les seves propietats espacio-temporals són altament regulars. Aquestes, es poden descriure de manera molt satisfactòria emprant un formalisme de xarxa com els estudiats en la primera part de la tesi.

La conveniència del punt de vista de xarxa per a aquest anàlisi es justifica per l'observació de trets comuns als quatre sets de dades estudiats, a pesar de les seves notables diferències geogràfiques. En particular, observem una estructura general molt heterogènia amb diferencies notables entre el tràfic acumulat pels diversos nodes. Emprant com a model nul el MECM, observem en tots els casos divergències comunes respecte a aquest: El patró de connectivitat tendeix a reforçar les connexions entre nodes de tràfic semblant i a més, es pot observar una estructura modular on es formen comunitats espacialment coherents de nodes. Finalment, observem una sobre-representació respecte al model nul de trajectes entre nuclis petits i també l'existència d'enllaços particulars que acumulen una porció anormalment gran del tràfic total.

12.3.2 *Crítica i avaluació de models existents de generació de demanda de tràfic*

Un cop observades característiques comunes en els sets de dades estudiats, procedim a fer un anàlisi dels diferents models proposats en la literatura per descriure el moviment humà.

Per tal de fer una comparació sistemàtica de models, introduïm diferents indicadors (alguns d'establerts en la literatura i d'altres de nous) per a quantificar la qualitat de cada model des d'un punt de vista global. Tot seguit, per als principals apropaments al problema, en presentem les principals virtuts i procedim a aplicar-los per a avaluar la seva precisió pel cas urbà.

Observem com clarament el model millor per a descriure de manera aproximada les observacions es correspon amb un model de màxima entropia com els estudiats en la primera part de la tesi, on es fixa l'atractivitat de cada localització (el tràfic que acumula) però també el cost mig dels trajectes avaluat en termes de distància euclídia entre localitzacions, anomenat model de Wilson. A pesar del bon rendiment del model entròpic, constatem que aquest model encara falla considerablement per descriure el tràfic entre les localitzacions més transitades, així com el tràfic entre localitzacions menys utilitzades.

12.4 APLICACIONS

Per finalitzar l'estudi pràctic, en aquesta part procedim a introduir possibles aplicacions de la teoria que explotin al mateix temps les característiques empíriques observades en les dades i les virtuts i febleses dels models entròpics per a capturar-ne les propietats.

12.4.0.1 *Model híbrid per la predicció de demanda*

Des del punt de vista de modelatge, hem explotat el fet que el nostre treball teòric és altament flexible combinat amb la gran estabilitat espacio-temporal de les dinàmiques de trajectes observada en el nostre estudi empíric per a resoldre el problema de l'extrapolació de sets de dades reduïts. En particular, hem proposat un model híbrid que explota les qualitats del model entròpic de Wilson per capturar la majoria de moviments amb la capacitat d'inferir aproximadament el pes dels trajectes més emprats utilitzant sets reduïts de dades històriques. Aquest model, s'ha demostrat capaç de reproduir de manera acurada escenaris de mobilitat emprant per al seu "entrenament" petites quantitats de les dades originals. Per a aquest estudi hem emprat les dades de Nova York, ja que és el set de dades més extens i que està complert (conté tots els trajectes de taxi registrats a la zona sense excepció).

Degut al seu èxit, aquest model es pot emprar fàcilment per a generar prediccions basades en dades històriques o per a estudiar fenòmens urbans que requereixin un coneixement aproximadament complert dels fenòmens de mobilitat que es produeixen a la ciutat.

12.4.0.2 *Filtratge i extracció de l'estructura bàsica de xarxes de mobilitat*

Finalment, hem explotat la capacitat analítica dels nostres models nuls per a dissenyar una metodologia que permet explorar de manera efectiva les desviacions entre dades i els models proposats. Basant-nos en el treball d'altres autors, hem proposat un mètode de filtratge que permet eliminar la contribució d'un model nul sobre les dades empíriques. La capacitat de generar instàncies dels models nuls permet estudiar l'efecte del filtre sobre aquests i determinar per cada cas d'estudi els paràmetres adients per la seva configuració. A més, aquest filtre permet estudiar de manera separada els enllaços pels quals els trajectes observats queden sota la predicció dels models i els casos on aquests queden per sobre, fet que pot portar a la identificació de problemes urbanístics o sociològics particulars. A més, aquest procediment elimina gran quantitat d'enllaços "no rellevants" i permet una visualització aproximada de les xarxes estudiades.

12.5 CONCLUSIONS

Aquesta tesi s'ha desenvolupat amb un objectiu principalment pràctic en ment: El de proporcionar eines i exemples d'utilització de models nuls de xarxes pesades per a l'estudi de dades empíriques. En aquest sentit, s'han desenvolupat marcs teòrics que han permès la generació computacional de xarxes amb propietats prefixades que han estat implementats en programaris de lliure accés. Aquest desenvolupament teòric també ha permès desenvolupar eines per l'anàlisi de dades empíriques, tant a nivell teòric com d'aplicacions de modelització d'una banda i de visualització i "simplificació" de xarxes de l'altra.

Finalment, un producte secundari interessant de tot aquest estudi ha estat la de refinar alguns aspectes de l'estudi de la mecànica estadística d'equilibri de sistemes físics habituals per al cas de la col·lecitivitat Gran Canònica de sistemes de partícules distingibles.

En global, el desenvolupament d'aquesta tesi mostra que el treball teòric i la modelització matemàtica són, avui en dia, eines més necessàries que mai per a extreure coneixement fiable dels extensos sets de dades que les tecnologies mòbils posen a la nostra disposició. L'aplicació d'aquests pel cas de mobilitat urbana vol exemplificar-ne la importància per a un cas particular i contribuir a ressaltar la necessitat del treball interdisciplinari per a resoldre els desafiaments científics del present i del futur.

Part V

APPENDICES

As it is usually said, *the devil is in the detail*. This part provides additional material that support the statements in the main text, and contains technical information for the interested reader.

A

MATHEMATICAL DETAILS

A.1 EFFECTIVE DEGENERACY TERM FOR DISTINGUISHABLE PARTICLES IN THE GRAND-CANONICAL ENSEMBLE

The correct counting of configurations in a Grand Canonical ensemble is a controversial issue spanning more than a century (see [178, 126, 179, 187] for details and extended discussion), ever since Gibbs used it to establish the relation in classical statistical mechanics between the Canonical and Grand-Canonical Ensembles of an ideal gas. In the present case, we provide a complete computation of both the complete degeneracy term $\mathcal{D}(\vec{T}, F)$ and the effective degeneracy term $\tilde{\mathcal{D}}(\{t_{ij}\})$ for the case of a GC ensemble of distinguishable particles considering linear constraints. The treatment given with respect to a *reservoir* of particles here is fully equivalent to the one provided in [3] using the technique of copying identical systems, and both conform the usual way this problem is tackled in classical equilibrium statistical mechanics [135]. We show the case for a monoplex ($M = 1$) for simplicity but the extension to general number of (distinguishable) layers M is straightforward and follows from the derivation.

Imagine we have a reservoir composed of F particles such that $F \geqslant T$. Now we select T (distinguishable) particles out of these F particles and distribute them in the occupation levels $\vec{T} = \{t_{ij}\}$. In this case, the degeneracy of a given configuration $\vec{T}$ is:

$$\mathcal{D}(\vec{T}, F) = \binom{F}{T} \mathcal{D}(\vec{T}) = \binom{F}{T} \frac{T!}{\prod_{ij}^{L} t_{ij}!} = \frac{1}{\prod_{ij}^{L} t_{ji}!} \frac{F!}{(F-T)!}. \tag{A.1}$$

Alternatively, we may consider that from the group of F particles, we select first t_{11}, then from the remaining $F - t_{11}$ we select t_{12} and we do that recursively up to t_L. In this case, we are led to the same expression:

$$\begin{aligned} \mathcal{D}(\vec{T}, F) =& \binom{F}{t_1} \binom{F - t_{11}}{t_{12}} \ldots \binom{F - \sum_{1}^{(ij)-1} t_{kl}}{t_{ij}} & \text{(A.2)} \\ =& \frac{1}{\prod_{ij}^{L} t_{ij}!} \frac{F!}{(F - \sum_{ij}^{L} t_{ij})!} = \binom{F}{T} \frac{T!}{\prod_{ij}^{L} t_{ij}!}. & \text{(A.3)} \end{aligned}$$

The degeneracy term depends both on the size of the reservoir of events and on the actual adjacency matrix considered (in fact, one could consider an adjacency matrix with $L + 1$ levels, being the other level the *void* where the $F - T$ particles are located). The first term is the way of selecting T out of F distinguishable events while the second term refers to the microscopic allocation of the T events in the different levels. The degeneracy term diverges (as expected), since any configuration T can be exchanged with the (infinite elements of the reservoir) and give rise to a different configuration.

$$\begin{aligned}\lim_{F\to\infty}\mathcal{D}(\vec{T},F) =& \frac{1}{\prod_{ij} t_{ij}!}\lim_{F\to\infty}\frac{F!}{(F-T)!} \\ =& \frac{1}{\prod_{ij} t_{ij}!}\lim_{F\to\infty}F(F-1)..(F-T-1)=\infty.\end{aligned} \tag{A.4}$$

Even if this degeneracy term diverges once the limit for infinite F is taken, the probabilities are well defined in the Ω space. For the MC ensemble we have,

$$\mathcal{P}(\vec{T},F) = \frac{\binom{F}{T}\frac{T!}{\prod_{ij} t_{ij}!}\delta_{\hat{T},T}\delta_{\vec{C},\hat{\vec{C}}}}{\mathcal{Z}_{MC}} \tag{A.5}$$

$$= \frac{\delta_{T,\hat{T}}\delta_{\vec{C},\hat{\vec{C}}}}{\prod_{ij} t_{ij}!}\frac{1}{\sum_{T=0}^{F}\delta_{T,\hat{T}}\sum_{\{\vec{T}|\sum t_{ij}=T\}}\frac{\delta_{\vec{C},\hat{\vec{C}}}}{\prod_{ij} t_{ij}!}} \tag{A.6}$$

The prior expression not depend on F (due to the Kroenecker delta's involved), and hence is independent of the reservoir or number of system copies, .

For the C ensemble a similar thing happens,

$$\begin{aligned}\mathcal{P}(\vec{T},F) =& \frac{\binom{F}{T}\frac{T!}{\prod_{ij} t_{ij}!}\delta_{\hat{T},T}\prod_{ij} z_{ij}^{t_{ij}}}{\mathcal{Z}_{C}} \\ =& \left(\frac{z_{ij}^{t_{ij}}}{t_{ij}!}\right)\frac{1}{\sum_{T=0}^{F}\delta_{T,\hat{T}}\sum_{\vec{T}|\sum t_{ij}=T}\left(\frac{z_{ij}^{t_{ij}}}{t_{ij}!}\right)} \\ =& \left(\frac{z_{ij}^{t_{ij}}}{t_{ij}!}\right)\frac{1}{\frac{1}{\hat{T}!}\left(\sum_{ij} z_{ij}\right)^{\hat{T}}}.\end{aligned} \tag{A.7}$$

Finally, for the GC ensemble we have,

$$\begin{aligned}\mathcal{P}(\vec{T},F) =& \frac{\binom{F}{T}\frac{T!}{\prod_{ij} t_{ij}!}z_T^T\prod_{ij} z_{ij}^{t_{ij}}}{\mathcal{Z}_{GC}} \\ =& \left(\prod_{ij}\frac{(z_T z_{ij})^{t_{ij}}}{t_{ij}!}\right)\frac{\binom{F}{T}T!}{\sum_{T=0}^{F}\binom{F}{T}z_T^T\sum_{\vec{T}|\sum t_{ij}=T}T!\left(\prod_{ij}\frac{z_{ij}^{t_{ij}}}{t_{ij}!}\right)} \\ =& \left(\prod_{ij}\frac{(z_T z_{ij})^{t_{ij}}}{t_{ij}!}\right)\frac{\binom{F}{T}T!}{\sum_{T=0}^{F}\binom{F}{T}(z_T\sum_{ij} z_{ij})^T} \\ =& \left(\prod_{ij}\frac{(z_T z_{ij})^{t_{ij}}}{t_{ij}!}\right)\frac{\binom{F}{T}T!}{(1+z_T\sum_{ij} z_{ij})^F}.\end{aligned} \tag{A.8}$$

In this case, the statistics are not independent on F, since the distinguishability of events does induce correlation between all events in the system. However, the statistics are still well defined. In particular, introducing auxiliary fields $\{e^{h_{ij}}\}$ we can construct a generating function for the statistics of t_{ij} and obtain their moments by derivation. In doing so, we obtain,

$$K(\{h_{ij}\}) = \ln \mathcal{Z}_{GC}(\{h_{ij}\}) = F \ln(1 + z_T \sum_{ij} z_{ij} e^{h_{ij}}) + C. \qquad \text{(A.9)}$$

Where C includes all degeneracy terms that do not incorporate nor depend on $\{h_{ij}\}$. By derivation, we have:

$$\begin{aligned}
\langle t_{ij} \rangle &= \partial_{h_{ij}} K(\{h_{ij}\})|_{\{h_{ij}\}=\vec{0}} = F \frac{z_T z_{ij}}{1 + z_T \sum_{ij} z_{ij}} \leqslant F \\
\sigma^2_{t_{ij}} &= \partial^2_{h_{ij}} K(\{h_{ij}\})|_{\{h_{ij}\}=\vec{0}} = \langle t_{ij} \rangle \left(1 - \frac{\langle t_{ij} \rangle}{F} \right) \\
\sigma^2_{t_{ij},t_{kl}} &= \partial_{h_{ij},h_{kl}} K(\{h_{ij}\})|_{\{h_{ij}\}=\vec{0}} = -\frac{\langle t_{ij} \rangle \langle t_{kl} \rangle}{F} \\
\frac{\sigma^2_{t_{ij}}}{\langle t_{ij} \rangle^2} &= \frac{1}{\langle t_{ij} \rangle} - \frac{1}{F} \geqslant 0 \qquad \frac{\sigma^2_{t_{ij} t_{kl}}}{\langle t_{ij} \rangle \langle t_{kl} \rangle} = \frac{1}{F}.
\end{aligned} \qquad \text{(A.10)}$$

Note that in the limit $F \to \infty$ the marginal statistics become independent (the relative fluctuations among different pairs ij, kl decay to zero with a rate $\mathcal{O}(F^{-1})$). Also note that the occupation numbers are well defined (and cannot exceed in any case the total number of events F available for allocation). In any case, the average values depend on F due to the closing constraint equation,

$$\hat{T} = \sum_{ij} \langle t_{ij} \rangle = F \frac{z_T \sum_{ij} z_{ij}}{1 + z_T \sum_{ij} z_{ij}} \implies z_T \sum_{ij} z_{ij} = \frac{\hat{T}/F}{1 - \hat{T}/F}. \qquad \text{(A.11)}$$

Inserting this relation in the average occupation numbers we have,

$$\langle t_{ij} \rangle = \hat{T} \frac{z_{ij}}{\sum_{ij} z_{ij}} \qquad \text{(A.12)}$$

and for the partition function and the probabilities in this ensemble we have,

$$\begin{aligned}
\mathcal{Z}_{GC} &= \left(1 + \frac{\hat{T}}{F - \hat{T}} \right)^F = \left(\frac{F}{F - \hat{T}} \right)^F \\
\mathcal{P}(\vec{T}, F) &= \frac{F!}{(F - T)! \prod_{ij} t_{ij}!} \left(1 - \frac{\hat{T}}{F} \right)^{F-T} \prod_{ij} \left(\frac{\langle t_{ij} \rangle}{F} \right)^{t_{ij}}.
\end{aligned} \qquad \text{(A.13)}$$

Which is multinomial allocation of F events in L states which on average have $\{\langle t_{ij} \rangle\}$ occupation and the void state with $F - \sum \langle t_{ij} \rangle = F - \hat{T}$ average occupation.

Considering the infinite reservoir limit for the partition function, we have:

$$\lim_{F\to\infty}\mathcal{Z}_{GC}=\lim_{F\to\infty}\left(1-\frac{\hat{T}}{F}\right)^{-F}=e^{\hat{T}}=e^{\sum_{ij}\langle t_{ij}\rangle}=\prod_{ij}e^{\langle t_{ij}\rangle}. \tag{A.14}$$

And for the probabilities (using Stirling's approximation for the factorial),

$$\lim_{F\to\infty}\mathcal{P}(\vec{T},F)=\prod_{ij}\frac{\langle t_{ij}\rangle^{t_{ij}}}{t_{ij}!}e^{-T}\lim_{F\to\infty}\left(\frac{F-\hat{T}}{F-T}\right)^{F-T}=\prod_{ij}\frac{\langle t_{ij}\rangle^{t_{ij}}}{t_{ij}!}e^{-\hat{T}}= \\ =\prod_{ij}e^{-\langle t_{ij}\rangle}\frac{\langle t_{ij}\rangle^{t_{ij}}}{t_{ij}!}. \tag{A.15}$$

The prior expression obviously does not depend on F and corresponds to a set of independent Poisson random variables with mean $\langle t_{ij}\rangle$.

Considering from the beginning an effective degeneracy term $\tilde{\mathcal{D}}(\vec{T})=\prod_{ij}(t_{ij}!)^{-1}$ allows to obtain absolutely the same results and simplifies enormously the calculations. However, the treatment we have given the problem here, is more elegant and allows a clear discrete and distinguishable treatment of the problem.

We must however notice that the entropies in the Γ space (equation (3.4)) will not be intensive in the *high sampling limit*. The reason for this can be understood in the MC ensemble: Considering the complete degeneracy term which includes the reservoir, if such a term is infinite (as we have justified), then the probabilities in (3.3) for obtaining a single configuration of the network vanish leading to infinite Γ entropies. A possible solution to unify the entire framework is to note that since this effective degeneracy term is not an integer number, one can always multiply by a constant[1] like $\hat{T}!$,

$$\tilde{\mathcal{D}}(\vec{T})=\frac{\hat{T}!}{\prod_{ij}t_{ij}!}=\prod_{ij}\frac{(\hat{T}!)^{1/L}}{t_{ij}!} \tag{A.16}$$

to use as degeneracy term instead of the true value in Table 3.1). This has no effect on both the MC and C ensembles ($\hat{T}=\sum_{ij}t_{ij}$ exactly for each configuration with non-zero probability). In doing so, no reference is needed to the reservoir and the entropies in the Γ space

[1] In fact, we can multiply the degeneracy by any constant and the probabilities in the Ω space will remain unchanged.

become sample independent and equivalent to those in the MC and C ensembles for linear constraints:

$$\lim_{\hat{T}\to\infty}\frac{S^{\Gamma}}{\hat{T}} = -\lim_{\hat{T}\to\infty}\sum_{ij}\frac{\langle t_{ij}\rangle}{\sum_{ij}\langle t_{ij}\rangle}\ln\frac{\langle t_{ij}\rangle}{\sum_{ij}\langle t_{ij}\rangle} = -\sum_{ij}p_{ij}^{\infty}\ln p_{ij}^{\infty}. \tag{A.17}$$

The substitution between the variable T and the value $\hat{T}$ can be understood relating to the closing condition that $\langle T\rangle = \hat{T}$. This (macroscopic) condition imposes limitations on the number of configurations to be counted, which impose a coupling between the Lagrange multiplier z_T and the size of the reservoir considered F. For the case with linear constraints such relation can be inverted and an effective degeneracy term can be obtained, but this is not the case when binary constraints are considered (or the binary distinguishable case is considered). Hence, in these cases, we shall accept the substitution $T \simeq \langle T\rangle = \hat{T}$ as an approximation in order to be able to obtain an effective factorization of degeneracy that allows the GC partition function to be computed. This approximation amounts to consider that the relative difference between both is negligible compared to the size of the reservoir F. A final additional reason for accepting this approximation is that the saddle point approximation of the MC partition function leads to the same approximated statistics of the GC ensemble computed using the substitution, see Chapter 5.

The treatment given in this appendix to the Grand Canonical ensemble of distinguishable events is fully general and may also be used in other statistical mechanics related problems. It is also fully consistent with the assumption of particle distinguishability (throughout the entire calculation) and valid for any sampling in $\hat{T}$.

Note that one could be tempted to use for the GC partition function an alternative expression (as done in [73]):

$$\mathcal{Z}'_{GC} = \lim_{F\to\infty}\sum_{T=0}^{F}(z_T\mathcal{Z}_C)^T = \sum_{T=0}^{\infty}\left(z_T\sum_{ij}z_{ij}\right)^T = (1 - z_T\sum_{ij}z_{ij})^{-1}. \tag{A.18}$$

Yet this expression is not consistent because it *mixes* a distinguishable treatment in the calculation of the Canonical partition function which leads to $\left(\sum_{ij}z_{ij}\right)^T$ with an indistinguishable treatment for the remaining part of the calculation depending on T. Also, this leads to correlated statistics at the occupation number level $\{t_{ij}\}$ and contradicts the numerical experiments (see Section 6.5.2.2).

A.1.1 *An important note concerning the application of this result to classical statistical mechanics*

All the considerations given here to the Grand-Canonical ensemble of distinguishable particles can be applied to classical statistical mechanics of state separable Hamiltonians, taking into account the limitations highlighted in Section 3.3.4, since the constraints considered in this case are linear on the occupation number values (energy $E = \sum_i n_i \varepsilon_i$ and number of elements $N = \sum_i n_i$).

The analogy to systems with discrete energy levels $\{\varepsilon_i\}$ is direct[2], thus complementing the usual description in terms of occupation of states for the Grand-Canonical ensemble, adding the case of fully distinguishable particles without resorting to the famous Gibbs correction factor $N!$ (or the classical limit). Obviously, in nature, the particles populating energy levels are either bosons or fermions, and are never fully distinguishable, but if we were to treat with fully distinguishable particles, the correct counting would be the one provided here (valid in any range, not only in the classical limit).

Also, note that the equivalence of both the bosonic and fermionic description of these systems does not depend on the temperature of the system, but rather on the average occupation of its states. In the classical limit, the average occupation of each state is very small, $\langle t_{ij} \rangle \ll 1 \, \forall ij$, hence the sampling is very limited and the distinguishability or not of particles makes no difference, all converging to Poisson distributed variables (due to the law of small numbers or rare events).

Concerning the statistical mechanics of classical particles, the above discussion motivates the need to slightly generalize the usual textbook expression of the Grand Canonical ensemble (see for instance equation 13, p.94 in [135]).

$$\mathcal{Z}_{GC} = \lim_{F \to \infty} \sum_{N}^{F} \mathcal{D}(F,N) \mathcal{Z}_C^{(N)}. \tag{A.19}$$

$\mathcal{D}(F,N)$ being a degeneracy factor that counts in how many ways we can select N particles out of F that give rise to the same Canonical partition function for N elements $\mathcal{Z}_C^{(N)}$. This factor depends on the specifics of each system and is subtle to compute. Let's review the examples in [135], p.96-100 for instance.

For a set of classical harmonic oscillators, the partition function is described in terms of collective normal modes of vibration, which are non-degenerate for the system, not on particles. This means that $\mathcal{D}(F,N) = 1$ in this case and we recover the usual expression. For the case where quantum harmonic oscillators are considered, again

[2] See Chapter 6 in the classical reference book [135] for instance

the description is made in terms of energy eigenvalues allocated to particles (which are non-degenerate).

$$\mathcal{Z}_{GC} = \lim_{F\to\infty} \sum_{N}^{F} \left(\mathcal{Z}_C^{(1)}\right)^N z^N = \frac{1}{1 - z\mathcal{Z}_C^{(1)}}. \tag{A.20}$$

In contrast, for a set of independent, non-interacting, strictly distinguishable particles (ideal gas), our complete calculation would be needed (the particles being independent mean that $\mathcal{Z}_C^{(N)} = (\mathcal{Z}_C^{(1)})^N$).

$$\mathcal{Z}_{GC} = \lim_{F\to\infty} \sum_{N}^{F} \binom{F}{N} \left(\mathcal{Z}_C^{(1)}\right)^N z^N = \exp(z\mathcal{Z}_C^{(1)}). \tag{A.21}$$

However, for the Canonical partition function calculated in terms of energy states of particles in a box, those energy states are calculated not taking into account the distinguishability of particles, but taking already the classical limit ($n_i \ll 1 \, \forall n_i \implies n_i! \simeq 1$ where n_i is the number of particles in state of energy ε_i), hence again $\mathcal{D}(F, N) = 1$ and no incoherence is found.

$$\begin{aligned}
\mathcal{Z}_{GC} &= \lim_{F\to\infty} \sum_{N}^{F} \sum_{E|\sum_i n_i = N} e^{-\beta E} \sum_{\{n_i\}|\sum n_i\varepsilon_i = E, \sum_{n_i} = N} 1 \\
&= \lim_{F\to\infty} \sum_{N}^{F} \sum_{E|\sum_i n_i = N} e^{-\beta E} \mathcal{Z}_{MC}^{(N\text{ indist})} \\
&= \lim_{F\to\infty} \sum_{N}^{F} \sum_{\{n_i\}|\sum_i n_i = N} \prod_i e^{-\beta n_i \varepsilon_i} \\
&\simeq \lim_{F\to\infty} \sum_{N}^{F} \sum_{E|\sum_i n_i = N} e^{-\beta E} \frac{\mathcal{Z}_{MC}^{(N\text{ dist})}}{N!} \\
&= \lim_{F\to\infty} \sum_{N}^{F} \frac{1}{N!} \sum_{\{n_i\}|\sum_i n_i = N} \frac{N!}{\prod_i n_i!} \prod_i e^{-\beta n_i \varepsilon_i} \\
&= \lim_{F\to\infty} \sum_{N}^{F} \frac{1}{N!} \left(\mathcal{Z}_C^{(1)}\right)^N = \exp(z\mathcal{Z}_C^{(1)})
\end{aligned} \tag{A.22}$$

Note that in this last example, the classical limit is already taken while computing the canonical (or micro-canonical) partition function, hence even if the calculation is made in terms of indistinguishable elements, the final result identical to Maxwell-Boltzmann poisson statistics should not come as a surprise.

A.2 ACTIVITY DRIVEN MODEL WITH $m = 1$ AND ITS EQUIVALENCE TO THE ME NON BINARY CONFIGURATION MODEL

In this short section I show how the model appearing in [138] is a particular case of the MECM.

This model has been developed to explain the dynamics of face to face interactions [51], for which diverse datasets are available at the project `www.sociopatterns.org`.

In their model, the authors consider a fixed activity a_i for each node defined as,

$$a_i = \frac{\hat{s}_i}{\sum \hat{s}_i} = \frac{\hat{s}_i}{\hat{T}} \leqslant 1 \tag{A.23}$$

which is a pre-defined (quenched) node property. At each time step, we choose a node with probability $p = a_i$ to be activated and it *shots* m edges to randomly considered nodes with probability $1/(N-1)^{-1}$. I. e., the probability per unit time Δt to have an event joining i and any other node is $m\frac{a_i}{(N-1)}\Delta t$ and hence at time τ the number of events joining two any given nodes is a Poisson distribution with mean $m\tau\frac{a_i}{(N-1)}$. For the ME configuration model we get Poisson edge statistics for any given node with mean $\langle t_{ij} \rangle = x_i y_j$ where the Lagrange multipliers $\{x, y\}$ are related with the outgoing and incoming strength values of each node. However, the probability per unit time of incoming links is uniform across nodes hence $y_j = K$, $\langle t_{ij} \rangle = Kx_i$ and $x_i \propto \langle s_i^{out}(\tau) \rangle$. Since the incoming probability of connection per unit time between any pair of nodes is uniform $m/(N-1)$, we have that $\langle s^{in} \rangle (\tau) = m\tau$ and finally,

$$\langle t_{ij} \rangle (\tau) = \mu_i(\tau) = m\tau x_i = m\tau a_i \tag{A.24}$$

One clearly sees that the resulting strength distribution of this network will be a Poisson homogeneous distribution for the incoming case and exactly equal to the activity distribution for the outgoing case. If one disregards the directionality of links and considers the network after τ time, one has that,

$$\langle s_i \rangle(\tau) = \langle s_i^{out} + s_i^{in} \rangle = m\tau(a_i + \sum_i a_i) = m\tau(a_i + \frac{1}{N-1}) \tag{A.25}$$

and moreover,

$$\begin{aligned}\langle k_i(\tau|a_i) \rangle &= (N-1)(1 - e^{-\frac{\tau m}{N-1}}) + \sum_i (1 - e^{-\tau a_i}) = \\ &= (N-1)(2 - e^{-\frac{\tau m}{N-1}}) - \sum_i e^{-m\tau a_i}\end{aligned} \tag{A.26}$$

And one gets exactly the same results as obtained in the mentioned paper with the addition of obtaining an explicit form for the degree distribution at time τ.

A.3 TAYLOR APPROXIMATION OF ME NON-BINARY CONFIGURATION MODEL ENSEMBLE METRICS

I gather here the expressions for the standard deviation of all the metrics considered in Section 6.5.

$$\sigma^2_{\mathcal{P}(t)} \simeq \frac{Q_1}{\langle E\rangle^2}\left(1-\frac{Q_2}{Q_1}+\frac{1}{\langle E\rangle Q_1}\left[1-2\frac{R_1}{Q_1}\right]-\frac{R_2}{\langle E\rangle^2 Q_1}\right) \tag{A.27}$$

$$\sigma^2_{Y_2} \simeq \hat{T}_1^{-2}\left(\frac{a_3\hat{s}_i^3+a_2\hat{s}_i^2+a_1\hat{s}_i+a_0}{(\hat{s}_i+1)^4}\right) \tag{A.28}$$

$$\sigma^2_{s^w_{nn}} \simeq b_1\frac{1}{\hat{s}_i}+b_0+b_2\hat{s}_i^2+b_3\hat{s}_i^3. \tag{A.29}$$

where E is the number of present edges and I have defined the following notation:

$$q_{ij}(t) \equiv e^{-\langle t_{ij}\rangle}\left\langle t_{ij}\right\rangle^t/t!\langle t_{ij}\rangle = \hat{s}_i\hat{s}_j/\hat{T} \tag{A.30}$$

$$Q_1 = \sum_{ij} q_{ij}(t) \qquad R_1 = \sum_{ij} q_{ij}(t)q_{ij}(0) \tag{A.31}$$

$$Q_2 = \sum_{ij} (q_{ij}(t))^2 \qquad R_2 = \sum_{ij} (1-q_{ij}(0))^2 \tag{A.32}$$

$$a_3 = -4\left(\hat{T}_2^2-\hat{T}_1\hat{T}_3\right) \qquad b_2 = 2\hat{T}_2/\hat{T}_1^3 \tag{A.33}$$

$$a_0 = 2\hat{T}_1^4-2\hat{T}_2\hat{T}_1^2 \qquad b_3 = -2/\hat{T}_1^2 \tag{A.34}$$

$$a_2 = 2\left(\left[\hat{T}_1^2-5\hat{T}_2\right]\hat{T}_2+4\hat{T}_1\hat{T}_3\right) \tag{A.35}$$

$$a_1 = \hat{T}_1^4+2\hat{T}_2\hat{T}_1^2+4\hat{T}_3\hat{T}_1-7\hat{T}_2^2 \tag{A.36}$$

$$b_1 = \left(\hat{T}_1\left[\hat{T}_2+\hat{T}_3\right]-\hat{T}_2^2\right)\hat{T}_1^{-2} \tag{A.37}$$

$$b_0 = \left(-\hat{T}_2^2+\hat{T}_1\left[\hat{T}_2+3\hat{T}_3\right]-\hat{T}_4\right)\hat{T}_1^{-3} \tag{A.38}$$

with $\hat{T}_n \equiv \sum_i \hat{s}_i^n$. The calculations leading to these results are admittedly tedious, but follow directly from (6.4) and are of no particular difficulty beyond algebraic manipulation and careful reordering of the sums. As an illustrative example, consider the disparity $Y_2(s_i)$ for node i, defined as

$$Y_2(s_i) = \sum_j \frac{t_{ij}^2}{s_i^2} \tag{A.39}$$

Identifying $x \equiv \sum_j t_{ij}^2$ and $y \equiv s_i^2$, (6.4) can be readily applied. In order to approximate $\langle Y_2(s_i)\rangle$ and σ_{Y_2}, we need to compute $\langle x\rangle, \langle y\rangle, \left\langle x^2\right\rangle, \left\langle y^2\right\rangle$ and $\langle xy\rangle$. Let me show in full detail, as an illustrative example, how to compute $\left\langle x^2\right\rangle$ in this case. First, we expand x^2 as follows,

$$x^2 = \sum_{j,k} t_{ij}^2 t_{ik}^2 = \sum_{j,k\neq j} t_{ij}^2 t_{ik}^2 + \sum_j \left[t_{ij}^4+2t_{ij}^2t_{ii}^2\right]+t_{ii}^4, \tag{A.40}$$

so that when the ensemble average is taken, all products factorize (they correspond to different pairs of values ij, which are independent),

$$\langle x^2 \rangle = \sum_{j,k\neq j} \langle t_{ij}^2 \rangle \langle t_{ik}^2 \rangle + \sum_{j} \left[\langle t_{ij}^4 \rangle + 2 \langle t_{ij}^2 \rangle \langle t_{ii}^2 \rangle \right] + \langle t_{ii}^4 \rangle \quad \text{(A.41)}$$

Finally, since the variables t_{ij} are Poisson-distributed, we can compute their moments ($\left\langle t_{ij}^2 \right\rangle = \langle t_{ij} \rangle (1 + \langle t_{ij} \rangle)$), and using that $\langle t_{ij} \rangle = \hat{s}_i \hat{s}_j / \hat{T}$, and after some algebra, we get to

$$\langle x^2 \rangle = \frac{\hat{T}_2^2 \hat{s}_i^4}{\hat{T}_1^4} + \left(\frac{2\hat{T}_2}{\hat{T}_1^2} + \frac{4\hat{T}_3}{\hat{T}_1^3} \right) \hat{s}_i^3 + \left(\frac{6\hat{T}_2}{\hat{T}_1^2} + 1 \right) \hat{s}_i^2 + \hat{s}_i \quad \text{(A.42)}$$

The rest of the terms can be computed in a similar vein, leading to

$$\langle x \rangle = \frac{\hat{T}_2 \hat{s}_i^2}{\hat{T}_1^2} + \hat{s}_i \qquad \langle y \rangle = \hat{s}_i^2 + \hat{s}_i \quad \text{(A.43)}$$

$$\langle xy \rangle = \frac{\hat{T}_2 \hat{s}_i^4}{\hat{T}_1^2} + \left(\frac{5\hat{T}_2}{\hat{T}_1^2} + 1 \right) \hat{s}_i^3 + \left(\frac{4\hat{T}_2}{\hat{T}_1^2} + 3 \right) \hat{s}_i^2 + \hat{s}_i \quad \text{(A.44)}$$

$$\langle y^2 \rangle = \hat{s}_i^4 + 6\hat{s}_i^3 + 7\hat{s}_i^2 + \hat{s}_i \quad \text{(A.45)}$$

Finally, inserting (A.42)-(A.45) into (6.4) and some simplification leads to the desired result, (6.12) and (A.28).

A.4 NETWORK METRICS

Throughout the present thesis many network related properties are used. In this appendix, they are listed with their corresponding formula and a short explanation of what they account for.

A. **Strength s:** Total number of incoming or outgoing events (trips) to a given location (node),

$$s_i^{out} = \sum_j t_{ij} \qquad s_j^{in} = \sum_i t_{ij}. \quad \text{(A.46)}$$

B. **Degree k:** Total number of incoming or outgoing binary connections (connections with non-zero occupation),

$$k_i^{out} = \sum_j \Theta(t_{ij}) \qquad k_j^{in} = \sum_i \Theta(t_{ij}). \quad \text{(A.47)}$$

C. **Node strength or degree asymmetry Δ_i^s, Δ_i^k:** Relative difference between outgoing and incoming strength (can also be computed for degrees). This value is strictly bounded ($1 \leqslant \Delta_s \leqslant -1$) and values close to 1 indicate majority of outgoing connections and conversely values close to -1 majority of incoming connections, with zero values indicating balanced nodes.

$$\Delta_i^s = \frac{s_i^{out} - s_i^{in}}{s_i^{out} + s_i^{in}} \qquad \Delta_i^k = \frac{k_i^{out} - k_i^{in}}{k_i^{out} + k_i^{in}}. \tag{A.48}$$

D. **Node disparity** Y_{2i}**:** Measures the concentration of occupation numbers across binary connections of a given node. It is strictly bounded ($1 \geqslant Y_{2i} \geqslant 1/k_i$). Values close to unity indicate strong concentration of events on few binary links,

$$Y_{2i}^{out} = \frac{\sum_j t_{ij}^2}{\left(\sum_j t_{ij}\right)^2} \qquad Y_{2i}^{in} = \frac{\sum_i t_{ij}^2}{(\sum_i t_{ij})^2}. \tag{A.49}$$

E. **Average weighted neighbor strength:** Measures the tendency of nodes to be connected to other nodes by more or less occupied edges. A increasing trend is known as *assortative* profile and a decreasing trend is known as *dissassortative*. Both indicate the over occupation of links connecting nodes to other higher (lower or equal) strength nodes. The uncorrelated profile for Multi-Edge networks is flat.

$$s_{nn}^{w}|^{out} = \frac{\sum_j t_{ij} s_j^{in}}{\sum_j t_{ij}} \qquad s_{nn}^{w}|^{in} = \frac{\sum_i t_{ij} s_j^{out}}{\sum_i t_{ij}}. \tag{A.50}$$

F. **Conditioned average occupation number on origin destination strength product:** Measures the average occupation of links connecting nodes with outgoing and incoming strength product lying in a given binning.

$$\overline{t_{ij}|_{s^{out}=s, s_j^{in}=s'}}(ss') = \frac{\sum_{ij} t_{ij} \delta_{s_i^{out}s} \delta_{s_j^{in}s'}}{\sum_{ij} \delta_{s_i^{out}s} \delta_{s_j^{in}s'}}. \tag{A.51}$$

G. **Conditioned connection probability on origin destination strength product:** Measures the average binary connection probability of links connecting nodes with outgoing and incoming strength product lying in a given binning.

$$\overline{\Theta(t_{ij})|_{s^{out}=s, s_j^{in}=s'}}(ss') = \frac{\sum_{ij} \Theta(t_{ij}) \delta_{s_i^{out}s} \delta_{s_j^{in}s'}}{\sum_{ij} \delta_{s_i^{out}s} \delta_{s_j^{in}s'}}. \tag{A.52}$$

H. **Network modularity** Q**:** Measures the degree to which a given node partition of the network into C labelled groups $\{c_i | \forall i = 1...C\}$ accumulates over-expected intra-connecting events according to a given null model.

$$Q(\hat{\vec{T}}, \left\langle \vec{T} \right\rangle_{null})) = \frac{1}{\hat{T}} \sum_{ij} \left(\hat{t}_{ij} - \langle t_{ij} \rangle_{null}\right) \delta_{c_i c_j}. \tag{A.53}$$

B

NUMERICAL DETAILS

B.1 CONSTRAINT EQUATION SOLVING PROCEDURE FOR ALL EXPLICIT EXAMPLES CONSIDERED

We provide the explicit details for the solving of the saddle point equations here on a one-by-one case basis, reviewing all the examples presented in this thesis. Note however that the examples which involve the solution of a one-dimensional equation will not be discussed here (noted 1D and 2D in table 6.1), as it is generally an easy problem to be solved with standard methods.

B.1.1 *Linear constraints*

I provide here either the average values $\langle t_{ij} \rangle$ or z_{ij} which are the single parameters (besides the number of layers M) which determine the different occupation statistics (Poisson, Negative Binomial, Binomial).

A. Fixed $\hat{T}$: $z_{ij} = z$. In all cases the direct analytical solution is to use $\langle t_{ij} \rangle = \langle t \rangle = \hat{T}/L$.

B. Fixed $\hat{T}, \hat{C}$: $z_{ij} = ze^{-\gamma d_{ij}}$.

- ME: Solve $\hat{C}/\hat{T} = \sum_{ij} d_{ij} e^{-\gamma d_{ij}} / \sum_{ij} e^{-\gamma d_{ij}}$ by standard 1-D methods. Then, $z = \hat{T}/\sum_{ij} e^{-\gamma d_{ij}}$.
- W: Solve by brute force optimization with $z \geqslant 0, \gamma > 0$ and $ze^{-\gamma d_{ij}} < 1$.
- B: Solve using Algorithm 2 with node balacing for

$$z^{(n+1)} = \frac{\hat{T}}{\sum_{ij} \frac{e^{-\gamma d_{ij}}}{1+z^{(n)}e^{-\gamma d_{ij}}}}, \tag{B.1}$$

  and general constraint search on γ.

C. Fixed $\hat{\vec{s}}$: $z_{ij} = x_i y_j$.

- ME: Direct analytical solution $x_i = \hat{s}_i^{\text{out}}/\hat{T}^{1/2}, y_j = \hat{s}_j^{\text{in}}/\hat{T}^{1/2}$.
- W: Solve by brute force optimization with $x_i, y_j \geqslant 0$ and $x_i y_j < 1 \, \forall ij$.
- B: Solve using Algorithm 1 with

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{\text{out}}}{M \sum_j \frac{y_j^{(n)}}{1+x_i^{(n)} y_j^{(n)}}} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{\text{in}}}{M \sum_i \frac{x_i^{(n+1)}}{1+x_i^{(n+1)} y_j^{(n)}}}. \end{cases} \tag{B.2}$$

D. Fixed $\hat{\vec{s}}, \overline{\hat{t}_{ss'}}$: $z_{ij} = x_i y_j e^{\alpha_{ss'}}$. In all cases the direct analytical solution is to use $\langle t_{ij} \rangle = \sum_{ss'} \overline{\hat{t}_{ss'}} \delta_{ss_i^{\text{out}}} \delta_{ss_j^{\text{in}}}$ without needing to solve z_{ij}.

E. Fixed $\hat{\vec{s}}, \hat{C}$: $z_{ij} = x_i y_j e^{-\gamma d_{ij}}$.

- ME: Solve using Algorithm 2 with node balacing for

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{\text{out}}}{M \sum_j y_j^{(n)} e^{-\gamma d_{ij}}} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{\text{in}}}{M \sum_i x_i^{(n+1)} e^{-\gamma d_{ij}}}, \end{cases} \tag{B.3}$$

  and general constraint search on γ.
- W: Solve by brute force optimization with $x_i, y_j \geqslant 0, \gamma > 0$ and $x_i y_j e^{-\gamma d_{ij}} < 1 \, \forall ij$.
- B: Solve using Algorithm 2 with node balacing for

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{\text{out}}}{M \sum_j \frac{y_j^{(n)} e^{-\gamma d_{ij}}}{1 + x_i^{(n)} y_j^{(n)} e^{-\gamma d_{ij}}}} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{\text{in}}}{M \sum_i \frac{x_i^{(n+1)} e^{-\gamma d_{ij}}}{1 + x_i^{(n+1)} y_j^{(n)} e^{-\gamma d_{ij}}}}, \end{cases} \tag{B.4}$$

  and general constraint search on γ.

F. Fixed $\hat{T}, \overline{\hat{t}_{uu}}$: $z_{ij} = z e^{\alpha_u \delta_{u_i u} \delta_{u_j u}}$. In all cases the direct analytical solution is to use $\langle t_{uu'} \rangle = \overline{t_{uu}} \delta_{uu} + \frac{\hat{T} - \sum_u \overline{t_{uu}} \mathcal{N}_u^2}{L - \sum_u \mathcal{N}_u^2} (1 - \delta_{uu})$.

G. Fixed $\hat{\vec{s}}, \overline{\hat{t}_{uu}}$: $z_{ij} = x_i y_j e^{\alpha_u \delta_{u_i u} \delta_{u_j u}}$.

- ME: Solve using Algorithm 1 with

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{\text{out}}}{\frac{\overline{\hat{t}_{u_i u_i}} \mathcal{N}_u^2}{\sum_q x_q^{(n)} \delta_{u_q u_i}} + M \sum_q y_q^{(n)} (1 - \delta_{u_q u_i})} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{\text{in}}}{\frac{\overline{\hat{t}_{u_j u_j}} \mathcal{N}_u^2}{\sum_q y_q^{(n)} \delta_{u_q u_j}} + M \sum_q x_q^{(n+1)} (1 - \delta_{u_q u_j})}, \end{cases} \tag{B.5}$$

  and once the algorithm has been balanced, apply:

$$e^{\alpha_u} = \frac{\overline{\hat{t}_{uu}} \mathcal{N}_u^2}{M \sum_q x_q \delta_{u_q u} \sum_{q'} y_{q'} \delta_{u_{q'} u}}. \tag{B.6}$$

- W: Solve by brute force optimization with $x_i, y_j, e^{\alpha_u} \geqslant 0$ and $x_i y_j e^{\alpha_u \delta_{u_i u} \delta_{u_j u}} < 1 \, \forall ij$.
- B: Solve using Algorithm 1 with

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{\text{out}}}{\frac{\overline{\hat{t}_{u_i u_i}} \mathcal{N}_u^2}{\sum_q x_q^{(n)} \delta_{u_q u_i}} + M \sum_q \frac{y_q^{(n)}}{1 + x_i^{(n)} y_q^{(n)}} (1 + \delta_{u_q u_i})} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{\text{in}}}{\frac{\overline{\hat{t}_{u_j u_j}} \mathcal{N}_u^2}{\sum_q y_q^{(n)} \delta_{u_q u_j}} + M \sum_q \frac{x_q^{(n+1)}}{1 + x_q^{(n+1)} y_j^{(n)}} (1 + \delta_{u_q u_j})}, \end{cases} \tag{B.7}$$

and once the algorithm has been balanced, reapply it with:

$$(e^{\alpha_u})^{(n+1)} = \frac{\overline{\hat{t}_{uu}}\mathcal{N}_u^2}{M\sum_q x_q \delta_{u_q u}\sum_{q'}\frac{y_{q'}}{1+y_{q'}x_q(e^{\alpha_u})^{(n)}}\delta_{u_{q'}u}} \tag{B.8}$$

to obtain e^{α_u}.

B.1.2 *Binary constraints*

In all these scenarios, z is obtained in the following form, once $\hat{t}^+ = \hat{T}/\langle E\rangle$ is known:

- ME: In this case we have an analytical solution $z(\hat{t}^+) = W(-\hat{t}^+ e^{-\hat{t}^+}) + \hat{t}^+$ (equation (4.27)).
- W: In this case one needs to invert by standard 1-D methods the equation (4.26) $\hat{t}^+ = M\frac{z}{1-z}\frac{1}{(1-(1-z)^M)}$.
- B: In this case one needs to invert by standard 1-D methods the equation (4.26) $\hat{t}^+ = M\frac{z}{1+z}\frac{1}{(1-(1+z)^{-M})}$.

As for $\tilde{z}_{ij}$ or $\langle\Theta(t_{ij})\rangle$, we have:

A. Fixed $\hat{T}, \hat{E}$: $\tilde{z}_{ij} = \tilde{z}$. In all cases the direct analytical solution is to use $\langle\Theta(t_{ij})\rangle = \hat{E}/L$.

B. Fixed $\hat{T}, \hat{\hat{k}}$: $\tilde{z}_{ij} = v_i w_j$. In all cases solve using Algorithm 1 with

$$\begin{cases} v_i^{(n+1)} = \dfrac{\hat{k}^{out}}{\sum_j \frac{\mu_c y_j^{(n)}}{1+\mu_c x_i^{(n)} y_j^{(n)}}} \\ w_j^{(n+1)} = \dfrac{\hat{k}^{in}}{\sum_i \frac{\mu_c x_i^{(n+1)}}{1+\mu_c x_i^{(n+1)} y_j^{(n)}}}. \end{cases} \tag{B.9}$$

and $\mu_c(z)$ from (4.36).

$$\mu_c = \begin{cases} \text{ME:} & e^{Mz} - 1 \\ \text{W:} & (1-z)^{-M} - 1 \\ \text{B:} & (1+z)^M - 1. \end{cases} \tag{B.10}$$

B.1.3 *Linear and Binary constraints*

A. Fixed $\hat{\hat{s}}, \hat{E}$: $z_{ij} = x_i y_j$ and $\tilde{z}_{ij} = \tilde{z}$.

- ME: Solve using Algorithm 2 with node balacing for

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{out}}{M\tilde{z}\sum_j \frac{y_j^{(n)} e^{Mx_i^{(n)}y_j^{(n)}}}{1+\tilde{z}(e^{Mx_i^{(n)}y_j^{(n)}}-1)}} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{in}}{M\tilde{z}\sum_i \frac{x_j^{(n+1)} e^{Mx_i^{(n+1)}y_j^{(n)}}}{1+\tilde{z}(e^{Mx_i^{(n+1)}y_j^{(n)}}-1)}}, \end{cases} \tag{B.11}$$

  and general constraint search on $\tilde{z}$.

- W: Solve by brute force optimization with $x_i, y_j \geqslant 0$, $\tilde{z} > 0$ and $x_i y_j < 1\,\forall ij$.

- B: Solve using Algorithm 2 with node balacing for

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{out}}{M\tilde{z}\sum_j \frac{y_j^{(n)}}{1+x_i^{(n)}y_j^{(n)}} \frac{(1+x_i^{(n)}y_j^{(n)})^M}{\tilde{z}\left((1+x_i^{(n)}y_j^{(n)})^M-1\right)+1}} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{in}}{M\tilde{z}\sum_i \frac{x_i^{(n+1)}}{1+x_i^{(n+1)}y_j^{(n)}} \frac{(1+x_i^{(n+1)}y_j^{(n)})^M}{\tilde{z}\left((1+x_i^{(n+1)}y_j^{(n)})^M-1\right)+1}} \end{cases} \tag{B.12}$$

  and general constraint search on γ.

B. Fixed $\hat{\vec{s}}, \hat{\vec{k}}$: $z_{ij} = x_i y_j$ and $\tilde{z}_{ij} = v_i w_j$.

- ME: Solve using Algorithm 1 with

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{out}}{Mv_i^{(n)}\sum_j w_j^{(n)} \frac{y_j^{(n)} e^{Mx_i^{(n)}y_j^{(n)}}}{1+v_i^{(n)}w_j^{(n)}(e^{Mx_i^{(n)}y_j^{(n)}}-1)}} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{in}}{Mw_j^{(n)}\sum_i v_i^{(n)} \frac{x_j^{(n+1)} e^{Mx_i^{(n+1)}y_j^{(n)}}}{1+v_i^{(n)}w_j^{(n)}(e^{Mx_i^{(n+1)}y_j^{(n)}}-1)}} \\ v_i^{(n+1)} = \dfrac{\hat{k}^{out}}{\sum_j \frac{w_j^{(n)}(e^{Mx_i^{(n+1)}y_j^{(n+1)}}-1)}{v_i^{(n+1)}w_j^{(n)}(e^{Mx_i^{(n+1)}y_j^{(n+1)}}-1)+1}} \\ w_j^{(n+1)} = \dfrac{\hat{k}^{in}}{\sum_i \frac{v_i^{(n+1)}(e^{Mx_i^{(n+1)}y_j^{(n+1)}}-1)}{v_i^{(n+1)}w_j^{(n)}(e^{Mx_i^{(n+1)}y_j^{(n+1)}}-1)+1}}. \end{cases} \tag{B.13}$$

- W: Solve by brute force optimization with $x_i, y_j \geqslant 0$, $\tilde{z} > 0$ and $x_i y_j < 1\,\forall ij$.

- B: Solve using Algorithm 1 with

$$\begin{cases} x_i^{(n+1)} = \dfrac{\hat{s}^{\text{out}}}{M v_i^{(n)} \sum_j w_j^{(n)} \frac{y_j^{(n)}}{1+x_i^{(n)} y_j^{(n)}} \frac{(1+x_i^{(n)} y_j^{(n)})^M}{v_i^{(n)} w_j^{(n)} \left((1+x_i^{(n)} y_j^{(n)})^M - 1 \right) + 1}} \\ y_j^{(n+1)} = \dfrac{\hat{s}^{\text{in}}}{M w_j^{(n)} \sum_i v_i^{(n)} \frac{x_i^{(n+1)}}{1+x_i^{(n+1)} y_j^{(n)}} \frac{(1+x_i^{(n+1)} y_j^{(n)})^M}{v_i^{(n)} w_j^{(n)} \left((1+x_i^{(n+1)} y_j^{(n)})^M - 1 \right) + 1}} \\ v_i^{(n+1)} = \dfrac{\hat{k}^{\text{out}}}{\sum_j \frac{w_j^{(n)} ((1+x_i^{(n+1)} y_j^{(n+1)})^M - 1)}{v_i^{(n)} w_j^{(n)} ((1+x_i^{(n+1)} y_j^{(n+1)})^M - 1) + 1}} \\ w_j^{(n+1)} = \dfrac{\hat{k}^{\text{in}}}{\sum_i \frac{z_j^{(n+1)} ((1+x_i^{(n+1)} y_j^{(n+1)})^M - 1)}{v_i^{(n+1)} w_j^{(n)} ((1+x_i^{(n+1)} y_j^{(n+1)})^M - 1) + 1}} . \end{cases} \tag{B.14}$$

B.2 Explicit saddle point equation solving for non-binary configuration model W case

In the general scenario of solving the constraint equations for the W case, both with linear and binary constraints, the difficulty of the problem depends on the number of constraints and varies with the particular restrictions of the considered case. Even in the simple cases no convergence of the proposed resolution method is assured, yet in this section we provide explicit details on how to try to solve the case where we wish to fix the strength pair of each node. Such an example is implemented in the freely available, open source package ODME [10].

The weighted case includes the restriction that $0 \leqslant x_i y_j < 1 \, \forall \, i, j$ and hence the Likelihood maximization is performed on a non-convex domain. The balancing approach in Algorithm 1 is then not satisfactory, since there is no explicit enforcement for the values $\{x, y\}$ to remain in the domain of the Loglikelihood function one wants to maximize. The scalar function being considered is

$$\mathcal{L}_s^W = K(M, \{\hat{t}_{ij}\}) + M \sum_{ij} \ln(1 - x_i y_j) + \sum_i \hat{s}_i^{\text{out}} \ln x_i + \sum_j \hat{s}_j^{\text{in}} \ln y_j \tag{B.15}$$

with derivatives,

$$\begin{aligned}
\partial_{x_q}\mathcal{L}_s^W &= \frac{\hat{s}^{out}}{x_q} - M\sum_j \frac{y_j}{1-x_q y_j} \\
\partial_{y_q}\mathcal{L}_s^W &= \frac{\hat{s}^{in}}{y_j} - M\sum_i \frac{x_i}{1-x_i y_q} \\
\partial_{x_q x_l}\mathcal{L}_s^W &= -\frac{\delta_{ql}}{x_q^2}\left(\hat{s}_q^{out} + M\sum_j \left(\frac{x_q y_j}{1-x_q y_j}\right)^2\right) \\
\partial_{y_q y_l}\mathcal{L}_s^W &= -\frac{\delta_{ql}}{y_q^2}\left(\hat{s}_q^{in} + M\sum_i \left(\frac{x_i y_q}{1-x_i y_q}\right)^2\right) \\
\partial_{y_q x_l}\mathcal{L}_s^W &= -\delta_{ql}\frac{M}{(1-x_q y_l)^2}
\end{aligned} \tag{B.16}$$

subject to the conditions that,

$$0 \leqslant x_i y_j < 1\, \forall i,j \in \{1...N\}. \tag{B.17}$$

In principle, the problem is concave, and thus finding a solution to the saddle point equations gives the global maximum. Sadly, for real cases this concavity is lost as soon as the explicit domain constraint is included $0 \leqslant x_i y_j < 1$. Hence, there is no general algorithm that can be applied with assured results, and obtaining a solution to the saddle point equations will not be guaranteed in all cases (specially for large N or a very skewed distribution of strengths).

We thus deal with a large scale, non-concave (non-convex), bounded and constrained maximization (minimization) problem.

B.2.0.1 *Preconditioning*

Basically, the difficulty of the problem derives from the form of the strength sequence $\{\hat{s}^{out}, \hat{s}^{in}\}$. The more skewed this distribution is, the more difficult the problem is to solve, for a given fixed N. For *easy* problems, a good way to pre-condition the problem is to first solve the easier, bounded, unconstrained convex problem of finding,

$$\begin{aligned}
&\min\left[-\mathcal{L}_s^W(\mathbf{x})\right] \\
&0 \leqslant x_i < \beta \quad i \in [1...N] \\
&0 \leqslant y_j < \beta^{-1} \quad j \in [1...N] \\
&\beta \in \mathbb{R}^+.
\end{aligned} \tag{B.18}$$

The problem of this method is that it does not consider all the available phase space (see figure B.1): The solution lies in the hyper-volume defined by the axis and $y_{max}(x_{max}) = x_{max}^{-1}$, which is a larger volume than that defined by the axis and $x_{max} \leqslant \beta, y_{max} \leqslant \beta^{-1}$. Usually, a good choice is $\beta = 1$. If the distribution of of strengths is very skewed, the optimal solution most likely lies outside the second area,

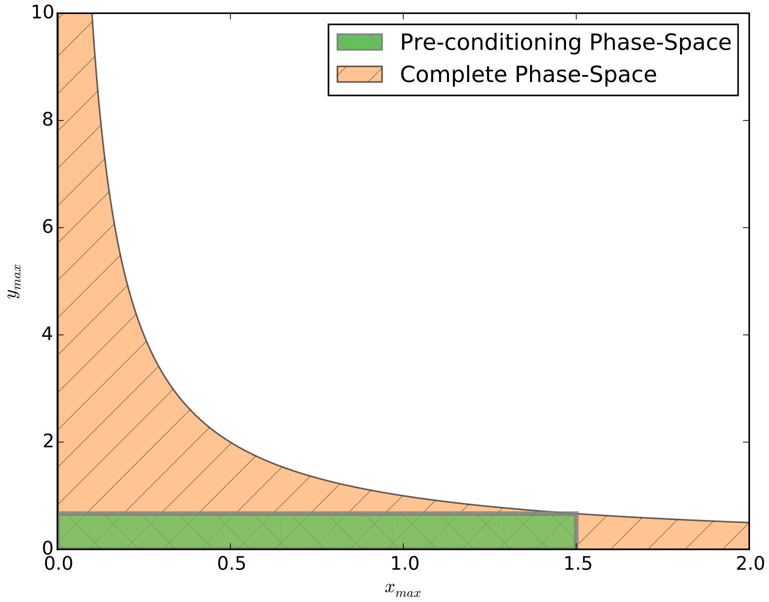

Figure B.1: **Non-convexity of phase space for W case.** Sketch of a plane projection of the phase space hyper volume with $\beta = 1.5$ for the maximization problem in the W case. The preconditioning method looks for solutions inside the area delimited by the green rectangle. The orange area represents the domain of the $\mathcal{L}_s^W$ function, which is clearly non-convex.

but the suboptimal solution within this region serves as preconditioning for the complete maximization problem thanks to the convexity of the function (without considering the domain).

We have implemented this preconditioning procedure using a truncated Newton TNC method [130] from the Scipy suite [98].

B.2.0.2 *Constrained problem*

Since the loglikelyhood function is not defined outside of the domain, we use an interior point method to solve the problem adding L non-linear inequalities of the form $0 \leqslant x_i y_j < 1$ ($L = N^2$ for the case without selfloops and $L = N(N-1)$ otherwise). The implementation is done in CVXOPT [18], but it has an obvious limitation given by the use of memory, which grows very fast with the number of nodes of the given network. Additionally, as early mentioned, the convergence of the algorithm is not assured due to the non-convexity of the complete problem, yet in our case we obtained very satisfactory results for the particular data analyzed.

CASE	$\lvert\Delta C\rvert$	ε_{max}
ME	$1 \cdot 10^{-8}$	$6 \cdot 10^{-14}$
B $(M = 365)$	$1 \cdot 10^{-7}$	$1 \cdot 10^{-13}$
W $(M = 365)$	$7 \cdot 10^{-6}$	$5 \cdot 10^{-12}$
W $(M = 1)$	0.05	$4 \cdot 10^{-8}$

Table B.1: **Results of the maximization problem.** Norms associated to the best solution for the Taxi dataset ($N = 4091$).

B.2.1 *Precision*

For all three cases considered in Section 6.4 (ME and B solved by application of Algorithm 1 and W case by brute force), we analyze the precision of our solving approach by computing the euclidean norm of the absolute error and the maximum of the absolute relative error among the nodes,

$$\begin{aligned} |\Delta C| &= \sqrt{\sum_q (\Delta C_q)^2} \equiv \sqrt{\sum_q (\hat{C}_q - \langle C \rangle_q)^2} \\ \varepsilon_{max} &= \max \frac{|\Delta C_q|}{\hat{C}_q}. \end{aligned} \tag{B.19}$$

The resulting values for each example are reported in table B.1.

B.3 MICROCANONICAL ENSEMBLE NETWORK GENERATION FOR MECM

In this section for the sake of clarity we consider directed networks with equal sequence of outgoing and incoming strengths, hence $\{x_i = y_i\}$ and the indices (out) and (in) are dropped.

The simulations presented in Section 6.5 have been performed using the MC ensemble by applying the well-known configuration model schema [30, 42, 193] (this method was also used in [159]).

The generating algorithm for directed networks is based on the idea of taking each strength sequence and adding $\hat{s}_i^{out}$ and $\hat{s}_j^{in}$ "stubs" to each node, that are later chosen at random in pairs (one from each direction) and joined. The only modification applied with respect to the original formulation is to link stubs permitting multiple connections between nodes (accepting also self-edges).

B.3.1 *Power-law distributed strength sequences*

Since in the non binary ensemble approach, the high sampling limit is defined in terms of $\hat{T} = \sum_i \hat{s}_i$, to test our analytical predictions we use power-law synthetic distributions generated with tunable graph-average strengths

$$\bar{s}(\gamma, s_{min}, s_{max}) = \frac{1}{\sum_{s_{min}}^{s_{max}} \mathcal{N}_s s^{-\gamma}} \sum_{s_{min}}^{s_{max}} \mathcal{N}_s s_i^{1-\gamma}. \tag{B.20}$$

For the case of $\gamma > 2$, we have applied a minimum cut-off s_{min} on the strength sequence, while for the case $\gamma \in (1, 2]$ a maximum cut-off s_{max} has been applied to limit the average strength of the sequence (which would be unbounded in the case of infinite sampling). The source code allowing the generation for both directed and undirected multi-edge networks in different ensembles and the details of the algorithms used together with the strength sequences used in Section 6.5 are provided in [11] for public use.

B.3.2 *Why to allow self-loops? Unbiased sampling*

When treating with the ME configuration model, we have always dealt with directed networks with allowed self-loops. The reason for this is twofold: On the one hand the resulting saddle point equations for the hidden variables $\{x_i\}$ can be exactly solved analytically while on the other hand the simulation using a stub-rewiring algorithm for power law distributed strengths with $\gamma < 2$ becomes feasible.

On the analytical side, in the case of not allowing self-loops, then the N saddle point equations (one for each node i) take the form ($X \equiv \sum_j x_j$),

$$\hat{s}_i = x_i \sum_{j \neq i} x_j = x_i (X - x_i), \tag{B.21}$$

which correspond to a set of coupled equations and cannot be solved analytically. If nevertheless we chose to use the solutions for the case of self-loops to this case ($x_i = \hat{s}_i \hat{T}^{-1/2}$), we have that $\langle s_i \rangle = \hat{s}_i \left(1 - \frac{\hat{s}_i}{\hat{T}}\right)$ so the relative error committed is $\varepsilon_{\langle s \rangle} = \frac{|\hat{s} - \langle s \rangle|}{\hat{s}} = \frac{\hat{s}}{\hat{T}}$ whose importance depends on the strength of each node but does not vanish in the high sampling limit ($\hat{T}, \hat{s} \to \infty$). Even if the relative error is small, note that the absolute error is $\hat{T}\varepsilon_{\langle s \rangle}$, and depending on the strength sequence chosen, this can become quite large. For non-broad distributed strength sequences with finite mean and standard deviation, this is not a problem even for the worst case scenario, since $\hat{s}_{max}/\hat{T} \ll 1$ independently of the sampling. Considering the case of skewed distributions (the paradigmatic case power law), we have that the condition $\varepsilon_{max} = \frac{\hat{s}_{max}}{\hat{T}}$ needs to be analyzed.

For a power law distribution, the maximum value of a distribution given a sample size can be shown to have a Fréchet distribution, with the ratio ε_{max} scaling as $\frac{\hat{s}_{max}}{\hat{T}} \sim \frac{N^{\frac{2-\gamma}{\gamma-1}}}{\bar{s}}$ and we thus see that the only problem (if we fix the average strength of the distribution $\bar{s}$) will come when $\gamma \in (1,2]$, in which case, we will have to manually set an appropriate value for the ratio $\varepsilon_{max} \equiv \hat{s}_{max}/\hat{T}$ for our calculation to have an acceptable accuracy (bear in mind that since s_{max} is broadly distributed in such case, the choice will not be random nor general). Anyhow, the absolute error committed in the approximation will still grow linearly with $\hat{T}$.

The other problem comes from the simulation side. A configuration approach as the one presented earlier, discarding absolutely the configurations where self-edges are present is not feasible in practice for heterogeneous strength sequences. In the GC ensemble (which is equivalent to the MC in the large event limit $T \to \infty$), the probability to obtain exactly 0 self loops SL ($t_{ii} = 0 \, \forall i = 1, N$) while sampling an ensemble that allows them is,

$$P(SL = 0) = \prod_i e^{-\frac{\hat{s}_i^2}{\hat{T}}} = \exp\left(-\frac{\overline{s^2}}{\bar{s}}\right). \tag{B.22}$$

which can be seen to be problematic for skewed distributions even for relatively low sampling $\bar{T}$. In particular for power law distributions, when approaching the limit $\gamma < 3$, $\overline{s^2}/\bar{s}$ will be a large number and this probability will quickly vanish. Otherwise, if no self-loops are allowed in a stub-matching schema, there is no guarantee of convergence (one can easily find counter-examples with 3 nodes where the rewiring algorithm could fail) and this fact worsens specially as hubs gain importance in the system ($\gamma < 3$).

For the reasons provided above, approaches as the ones done in [159], although approximate, may lead to uncotrolled errors and should be avoided when dealing with networks where no self-loops are allowed. It is true that we do not provide here a proof that the stub matching algorithm samples the phase space of the MC ensemble in an unbiased manner, but based in the coincidence with empirical results, we assume the goodness of our approximation, since the effects of a possible bias are not detected in our analysis.

C DATASETS

In this appendix, all details concerning data sources, acquisition and filtering are provided, as well as an overview of its main features. I have mainly used 4 datasets of Taxi trajectories in this thesis, two of which are openly available. Table C.1 provides an overview of the data.

The New York dataset has been obtained from the New York Taxi and Limousine Commission for the year 2011 via Freedom of Information Law request and is the same as the one used in [149], an open version of the same data for various years is openly available[1]. The San Francisco dataset is freely available from [106][2]. Both the Vienna and Singapore datasets were provided to the MIT SENSEable City Lab by AIT and the Singapore government respectively and are not openly accessible. However, the Vienna data has also been used in other projects[3].

The NY dataset spans over an entire year and collects all Taxi trips generated in the area of New York, while the other datasets span roughly a month and contain records provided by a single operator. The total number of taxis in San Francisco count an official figure of 1494[4] while for Vienna, we do not have information on how many taxis the datasets is constituted by, but there are around 3500 active taxis from information provided by AIT. For Singapore the official figure is 25176[5].

I have applied the same filtering procedure to all the datasets[6]: Only trips performed while the Taxi were occupied by customers were considered in the analysis. From these trips, I have only kept the ones with starting and ending GPS positions within 200 m of an intersection present in the considered area of study. Such an area has been obtained by considering the Manhattan borough (NY), the entire island of Singapore (SI), and both the urban areas of Vienna and San Francisco including the road to the airport (SF and VI). The intersections have been obtained from [132] (see Figure 7.1) considering only primary and secondary level roads and by merging all repeated elements corresponding to every given intersection by hand using [142].

[1] http://publish.illinois.edu/dbwork/open-data/.

[2] http://www.crawdad.org/epfl/mobility/.

[3] See http://casualdata.com/senseofpatterns/ for instance.

[4] Retrieved 2012-10-29 from http://www.sfmta.com/.

[5] From http://en.wikipedia.org/wiki/Taxicabs_of_Singapore).

[6] I would like to acknowledge the help of Dr. Michael Szell and Mr. Aldo Treville in this matter.

DATASET	N_{TAXIS}	$\hat{T}$	ρ_{TAXIS}	UTM ZONE	DATES
NY	13052	146986835	1	18	1/01/2011-31/12/2011
SI	15915	8873029	0.6	48	14/02/2011-13/03/2011
SF	537	435670	0.35	10	17/05/2008-10/06/2008
VI	-	284541	-	33	28/02/2011-31/03/2011

Table C.1: **Information on the different Taxi datasets.** N_{taxis} refers to the total number of different Taxis present in the dataset while ρ_{taxis} to its fraction related to the estimated total city taxis. $\hat{T}$ refers to the total number of recorded trips. Used UTM zones for projection in trip intersection matching are also shown.

All trip coordinates were provided in longitude-latitude pairs using the WGS84 ellipsoid but have been projected to euclidean UTM coordinates using the zones specified in Table C.1.

We have decided to keep the self-loops present in the data for simplicity (albeit their fraction is completely negligible). In the analysis, all trips including both week-ends and week-days are considered, since the pattern for weekly trips shows a continuous increase in the number of trips peaking on Friday and followed by a sudden drop on Sundays (see Figure 7.4).

In the cases where subsampling a dataset has been needed (NY), we have used uniform random subsampling of trips from the datasets.

The availability of data relating each taxi to its trajectories for the NY, SI and SF datasets allows for statistical independence testing of the different vehicles (proves). The analysed data shows that most of the taxis share similar performance in terms of the distribution of trips performed (see Figure C.1 albeit some deviations from a Gaussian behaviour are observed). This, coupled with the fact [139] that individual taxi mobility traces are in large part statistically indistinguishable from the overall population, justifies that their individual traces (corresponding to sets of trips performed by different customers which can be considered as independent events) can be safely aggregated for the analysis.

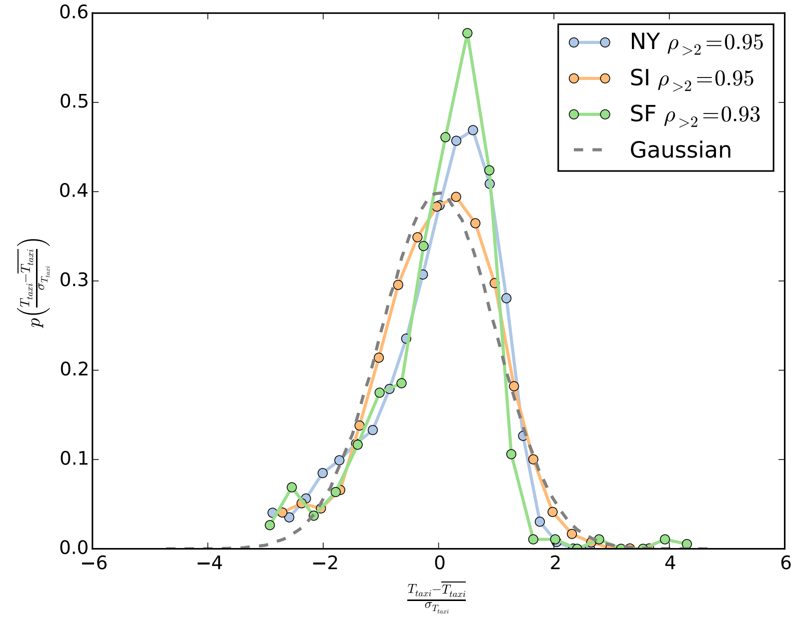

Figure C.1: **Taxi population analysis.** Histogram of the standardized number of trips performed per Taxi for the datasets where data is available compared to Gaussian distribution. $\rho_{>2}$ denotes the fraction of non-outliers in the data lying closer than 2 std from the mean which have been excluded from the histogram.

D ADDITIONAL DATA EXPLORATION AND MODEL VALIDATION

In this appendix we present complementary figures related to model validation and filtering treated in Chapter 9 and Chapter 10.

D.1 ADDITIONAL MODEL VALIDATION GRAPHICS: MOBILITY MODELS COMPARISON TO EMPIRICAL DATA

This section shows additional figures for the Radiation and Sequential gravity model to complement the discussion in Chapter 9.

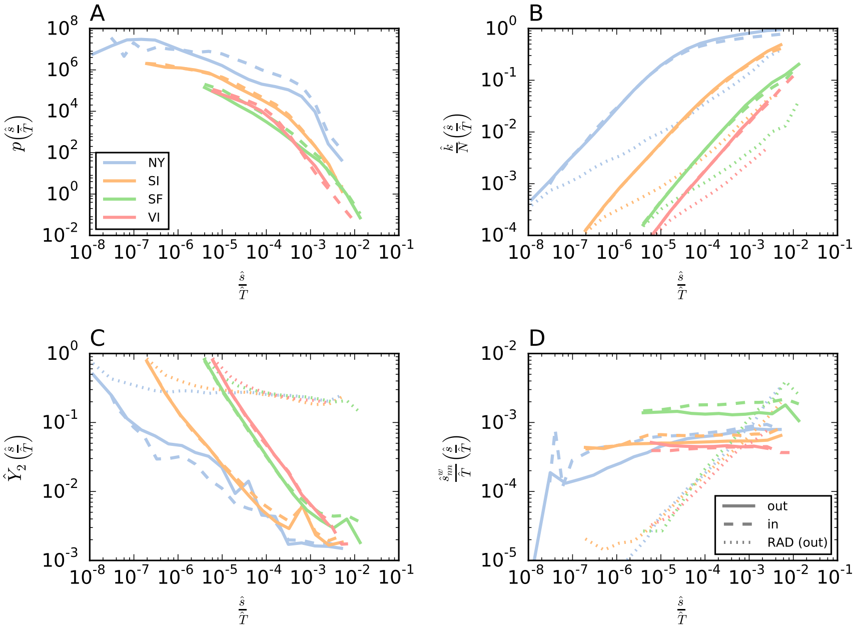

Figure D.1: **Comparison between Taxi empirical data and Radiation model at node level.** Rescaled strength distribution (A) and node related properties as function of strength [degrees (B), disparities (C) and average weighted neighbor strength (D)]. Results show averages over log-binned bins in the x axis. Dotted lines display Radiation model predictions averaged over $r = 10^2$ instances.

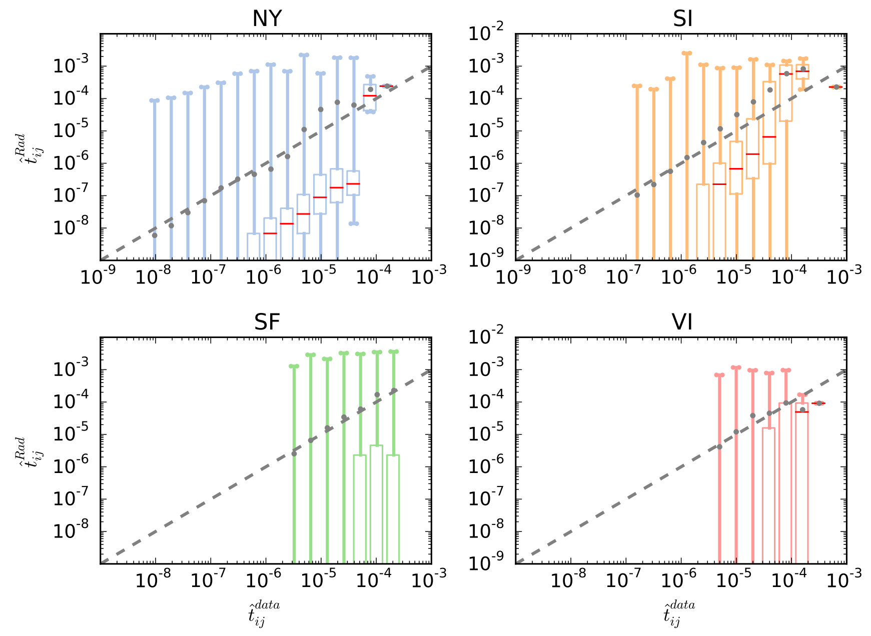

Figure D.2: **Comparison between empirical data and Radiation model at node-pair level.** Box-plot showing correlation between relative scaled occupation number between model predictions and data over a single run. Solid lines mark the [5%, 95%] interval, median marked as red horizontal line and average value with grey dotted point.

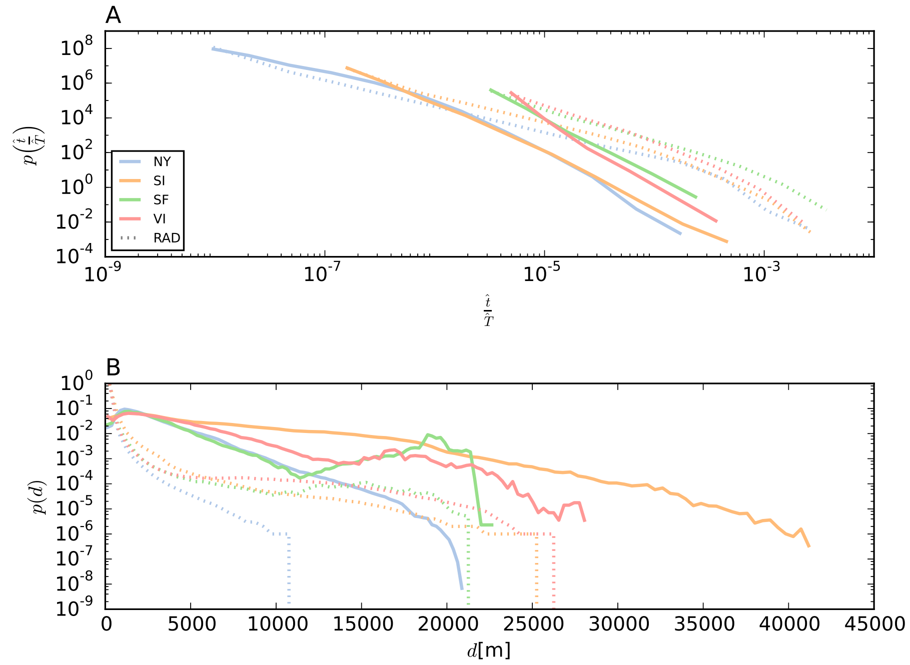

Figure D.3: **Radiation model occupation number and trip length distribution.** Occupation number distribution (A) and trip length distribution (B). Dotted lines correspond to the model while filled lines to empirical data. Logarithm (A) and linear bins (B) have been used respectively.

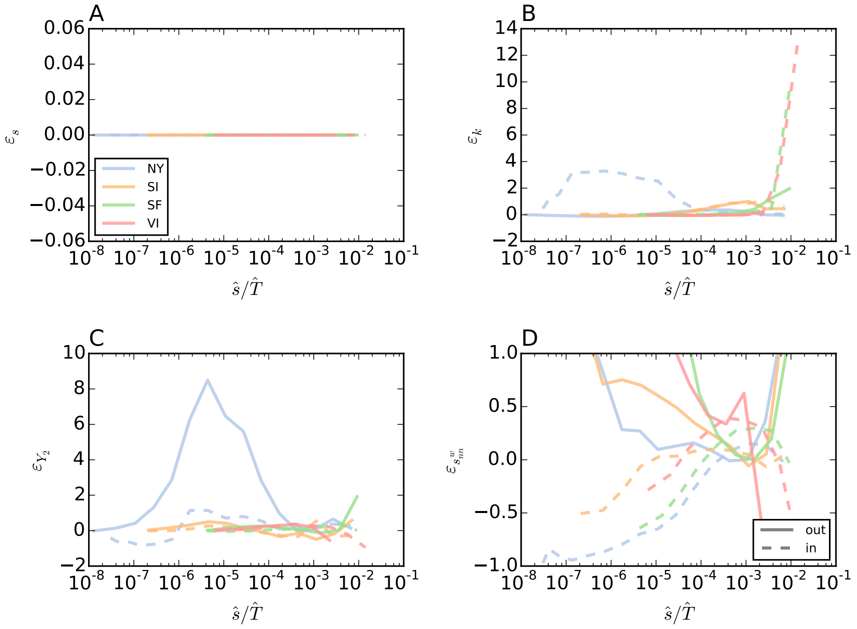

Figure D.4: **Relative difference $\varepsilon = (\hat{x} - \langle x \rangle_{\mathbf{Seq}}) / \langle x \rangle_{\mathbf{Seq}}$ between empirical node properties and Sequential gravity model predictions.** Relative strength difference (A), degree (B), disparity (C) and node neighbor strength correlation (D) averaged using logarithmic binning averaged using $r = 10$ instances of the model.

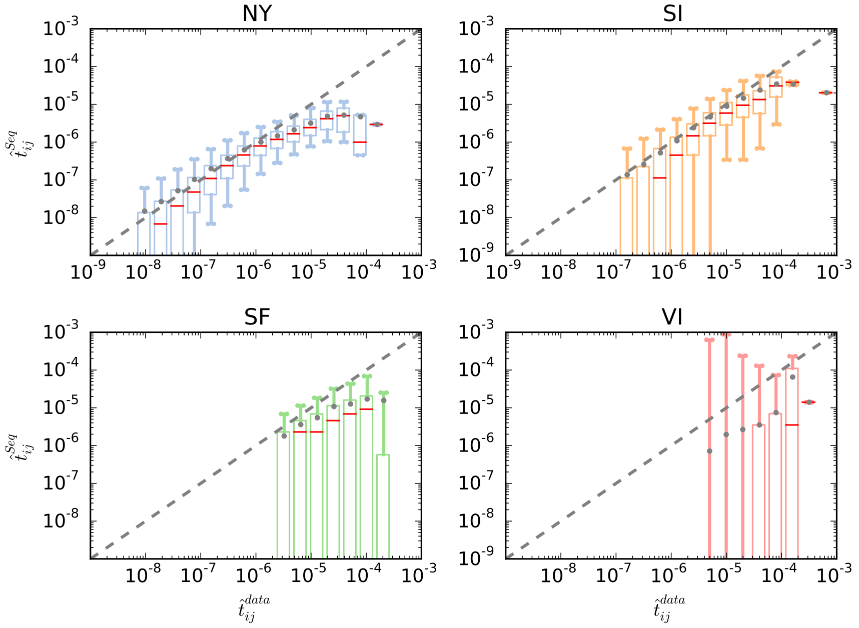

Figure D.5: **Comparison between empirical data and Sequential gravity model at node-pair level.** Box-plot showing correlation between relative scaled occupation number between model predictions and data over a single run. Solid lines mark the [5%, 95%] interval, median marked as red horizontal line and average value with grey dotted point.

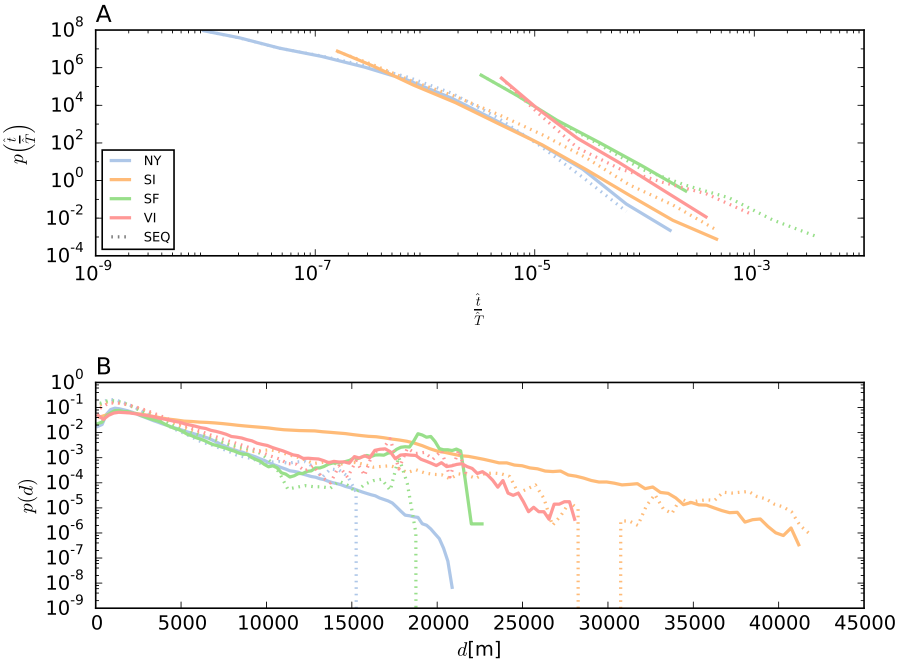

Figure D.6: **Sequential gravity model occupation number and trip length distribution.** Occupation number distribution (A) and trip length distribution (B). Dotted lines correspond to the model averaged over $r = 10$ instances while filled lines to empirical data. Logarithm (A) and linear bins (B) have been used respectively.

D.2 ADDITIONAL FILTER DETAILS

In this section we provide additional information regarding the filter developed in Section 10.2.

D.2.1 *A note on self-loops for visualization*

As shown in Section B.3, the analytical uncorrelated form for the MECM $t_{ij} \propto \hat{s}_i^{out}\hat{s}_j^{in}$ can only be considered if self-loops are allowed in the model (and in the data). In Appendix C, we have argued that we decided to keep self-loops in our treatment due to the fact that they represent a negligible fraction of the total trips, and that (although not usual) they can represent realistic taxi trips (circular trips and short trips with $d_{ij} < 200$ m). From the present analysis, however, the unexpectedness of the large level of self-loops detected becomes patent, as their relative contribution after filtering to the total number of edges grows up to 1% (NY), 4% (SF and SI) and even 15% (VI) only for the over-used trips. For this reason, with respect to visualization, self-loops have been deleted. However, all the results obtained display no qualitative nor quantitative differences when adding them.

D.2.2 *Effect of the filter in empirical data for SI, VI and SF datasets*

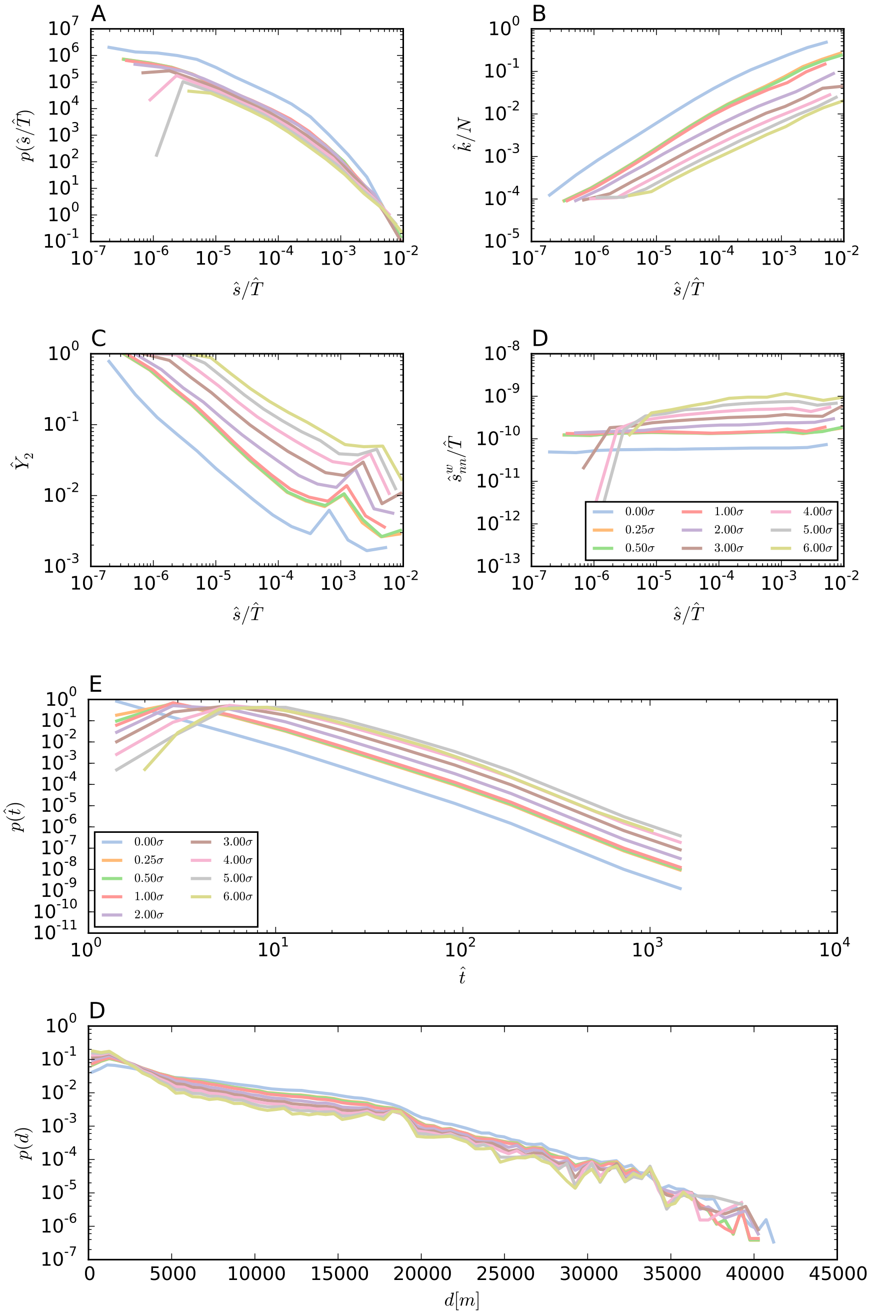

Figure D.7: **Graph filter sensitivity to confidence level α for the case of SI.** Empirical network features after filtering for different confidence levels α for the case of SI. **A-D** Node related features: Strength distribution (A), degree (B), disparity (C) and assortativity profile (D).**E,F** Edge related features: Existing trip distance distribution (F) and existing occupation number distribution (E).

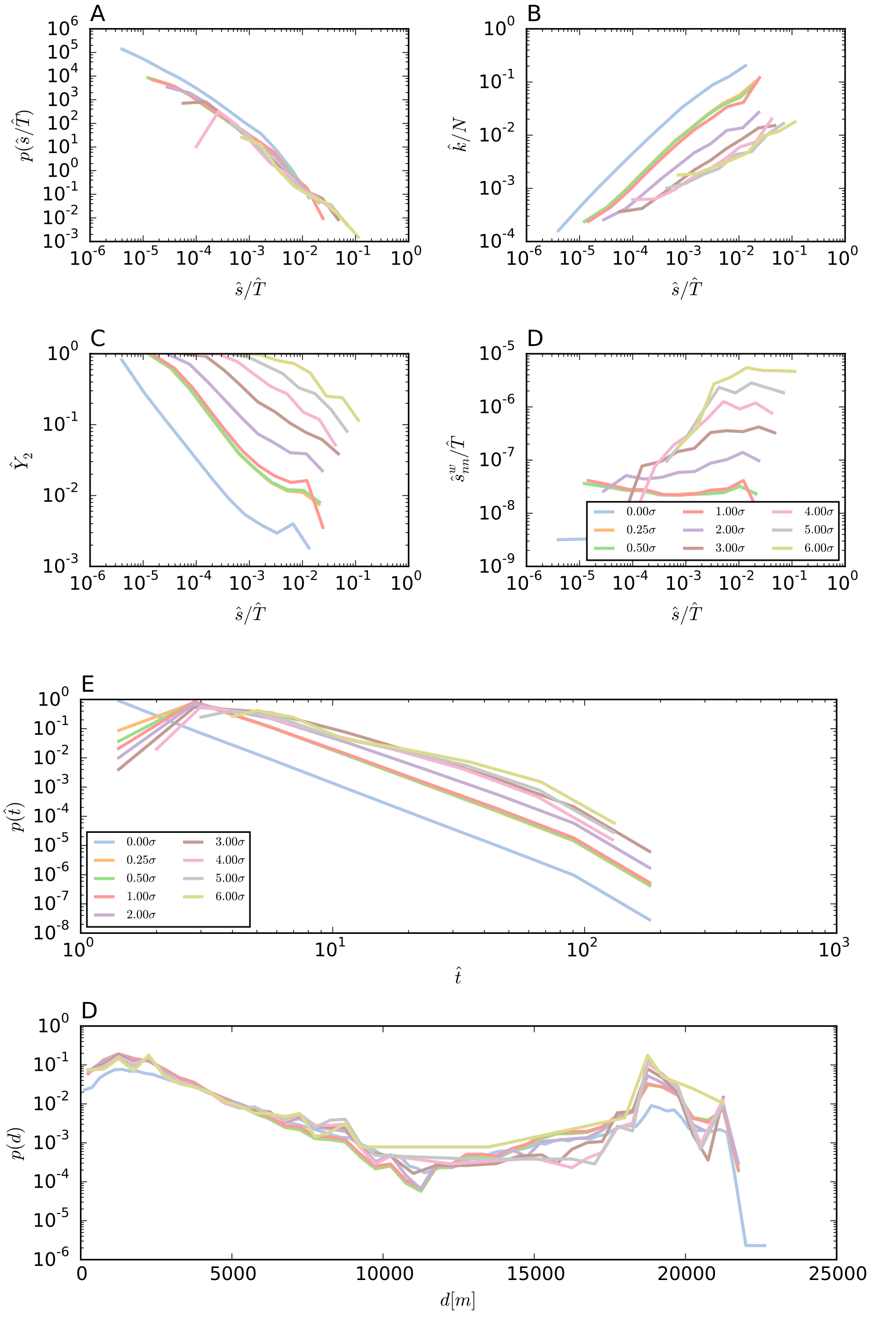

Figure D.8: **Graph filter sensitivity to confidence level α for the case of SF.** Empirical network features after filtering for different confidence levels α for the case of SF. **A-D** Node related features: Strength distribution (A), degree (B), disparity (C) and assortativity profile (D).**E,F** Edge related features: Existing trip distance distribution (F) and existing occupation number distribution (E).

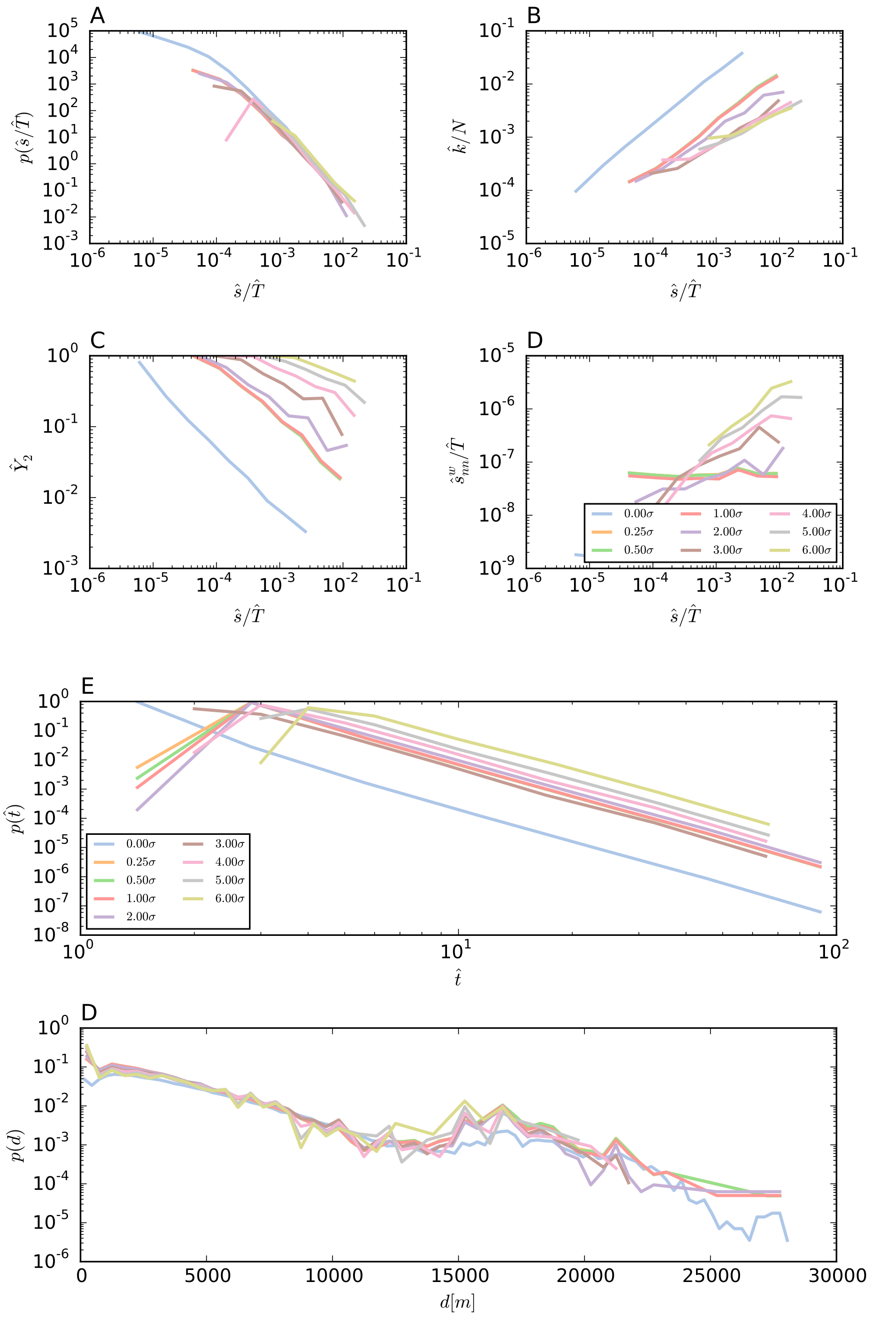

Figure D.9: **Graph filter sensitivity to confidence level α for the case of VI.** Empirical network features after filtering for different confidence levels α for the case of VI. **A-D** Node related features: Strength distribution (A), degree (B), disparity (C) and assortativity profile (D).**E,F** Edge related features: Existing trip distance distribution (F) and existing occupation number distribution (E).

D.2.3 *Under used trip features*

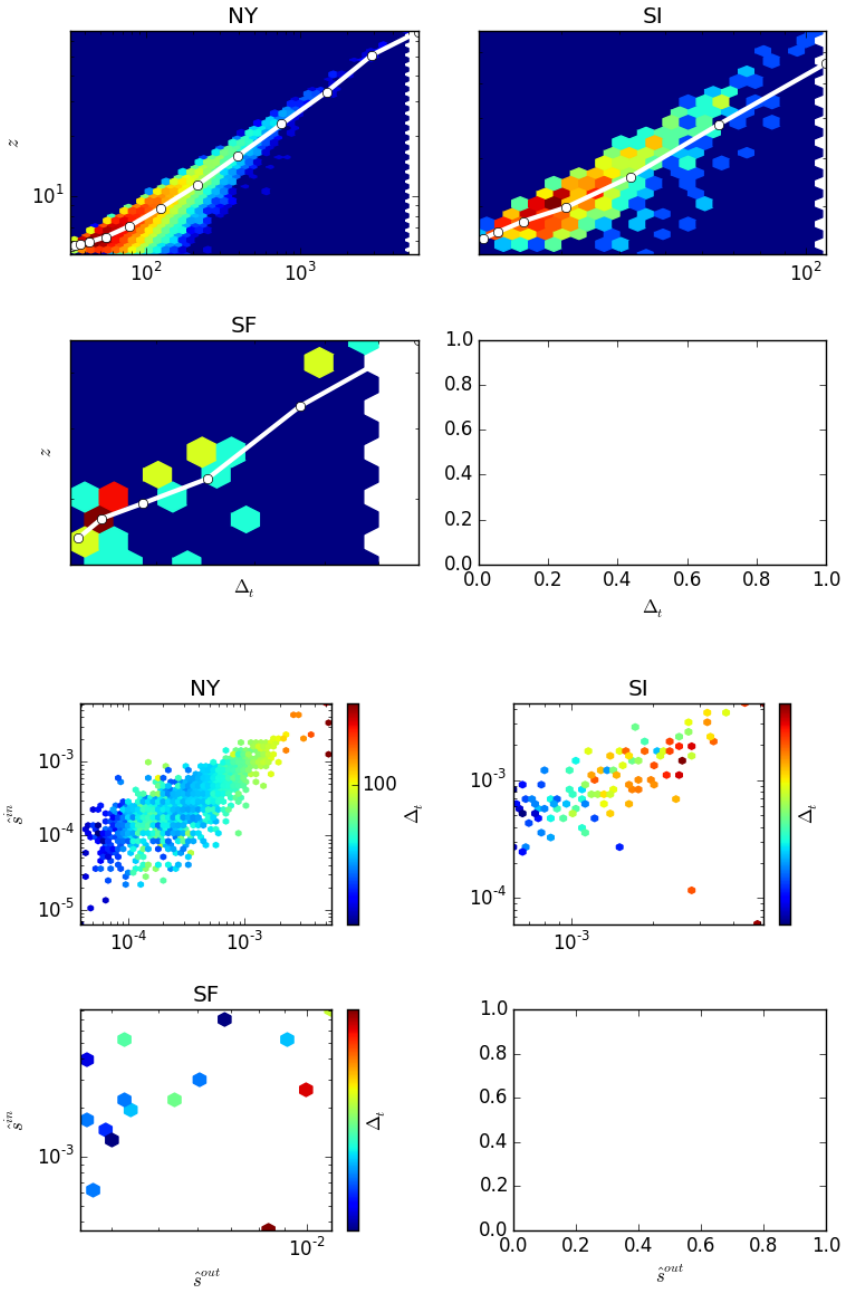

Figure D.10: **Under used trips z-scores and residuals.** Absolute values of z-scores vs residuals Δ_t of surviving edges after filtering the empirical datasets (top). The original, non-filtered strength of the origin and destination nodes whom they connect is shown (bottom). The broad distribution of both values and their correlation is apparent and so is the fact that their allocation among node-importance varies according to specifics of each city. For VI, there are no under-used trips.

D.3 HYPOTHESIS TESTING USING A MULTI EDGE ENHANCED CONFIGURATION MODEL

In Section 8.2 it is shown how empirical data features of the taxi mobility datasets represented as Multi-Edge structures are close to those predicted by a configuration model, the main exception being the node disparities, which are related to the way nodes allocate their strength among existing connections.

We have hypothesized that the relevant observed differences may be caused by the slight differences observed in the node-degree profile (see Figure 8.5). In this section, we use the Multi-Edge enhanced configuration model (MEECM) as an example of hypothesis testing with incremental constraints to test this statement. This model keeps fixed not only the strength of all the nodes but also their degrees.

The procedure to solve the saddle point equations for the MEECM model is implemented in [10].

To proceed, first we need to solve the saddle point equations using the recipes provided in Section B.1. In the present case, for each city, we have 4N Lagrange multipliers, four for each node corresponding to outgoing and incoming strength and degree respectively. The results of the balancing algorithm are provided in Table D.1, using as evaluation metrics the graph average and standard deviation of relative error between constraints.

As we can see, despite the large number of variables ($4N$) to be solved corresponding to the Lagrange multipliers of the degree and strength directed sequence $\{(s^{out}, s^{in}, k^{out}, k^{in})_i \, i = 1, N\}$ the precision of the obtained solution is acceptable on average with average below 6% for strengths and 2% for degrees (see Figure D.11-A,B). However the general precision of the balancing algorithm depends both on the number of variables (depending on N) and the skewness of the degree and strength distribution[1].

We have simulated $r = 1000$ instances of the MEECM model and repeat the plots in Figure 8.5 substituting the reference configuration model by the MEECM.

At the node level, Figure D.11, obviously strength and degree values coincide (thus validating the quality of our solving procedure for the saddle point equations). Disparity values display quantitatively the same differences with respect to the considered model for NY, while for the other datasets the empirical data displays distinctively larger disparity values for large nodes. Concerning the assortativity, since there are less binary connections available, the probability for small nodes to connect to other small nodes increases, and hence the average values of neighbor strengths for small nodes become smaller than in the MECM. However, this trend is not maintained because for hubs (that for SI, VI and SF are not connected to the entire system), the general average occupation of links increases, and hence the

[1] In contrast, the precision obtained for the easier case of only fixing the strength sequence is much higher, see Section B.2.1.

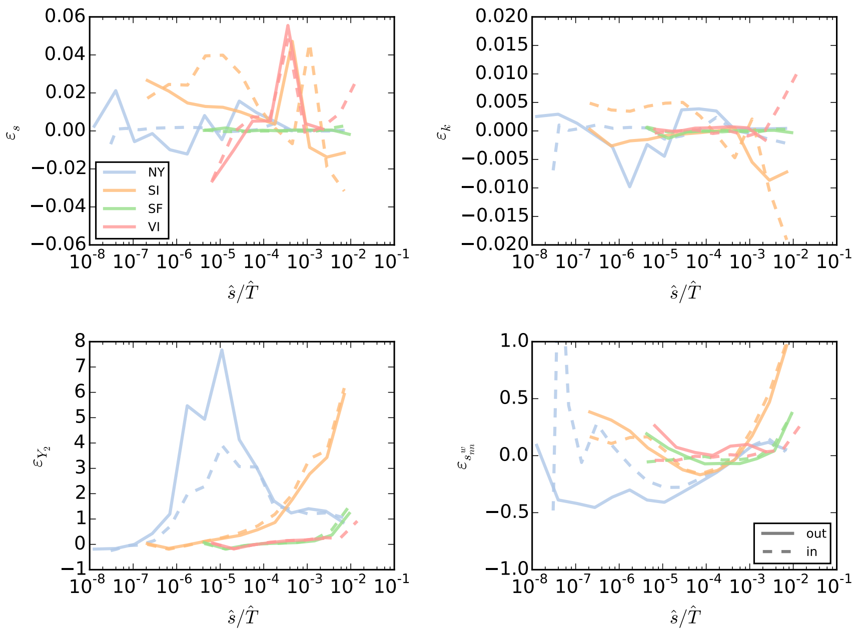

Figure D.11: **Relative difference $\varepsilon = (\hat{x} - \langle x \rangle_{\mathbf{MEECM}}) / \langle x \rangle_{\mathbf{MEECM}}$ between empirical node properties and model predictions.** Relative errors for strengths (A) and degrees (B) are non existent, while more relevant differences are detected for disparities (C) and average neighbor strengths (D).

weight of connections towards small nodes is increased with respect to the MECM, fact which influences the statistic that now displays an inverted U-concave shape across node strengths (this also explains the smaller values of disparity observed for this model for hubs).

At the edge level, Figure D.12, we see that the empirical connection between hubs are still over-occupied with respect to the model, leading to fat tails in the distribution of occupation numbers (not shown). The connections among very small nodes still are not explained. However, (specially for small sampling) the pattern for mid-sized nodes is approximately recovered, yet these nodes still display tight areas of connection distinctively different from the MEECM model across the diagonal $\hat{s}^{\mathrm{out}} \simeq \hat{s}^{\prime\mathrm{in}}$.

In a nutshell, we can thus say that the observed differences in assortativity and disparity profiles between the MECM and empirical data are not due to the smaller value of node degrees, since the addition of 2N extra constraints does not lead to a substantial improvement of the distance between empirical data and model predictions.

DATASET	$\varepsilon_{k^{out}}$ (±STD)	$\varepsilon_{k^{in}}$ (±STD)
NY	0.002 ± 0.115	0.001 ± 0.044
SI	-0.007 ± 0.269	0.003 ± 0.108
SF	0.00008 ± 0.032	-0.0006 ± 0.013
VI	0.053 ± 0.180	0.045 ± 0.308
DATASET	$\varepsilon_{s^{out}}$ (±STD)	$\varepsilon_{s^{in}}$ (±STD)
NY	-0.0001 ± 0.050	0.0002 ± 0.03
SI	-0.0009 ± 0.016	0.003 ± 0.014
SF	-0.0003 ± 0.013	-0.0001 ± 0.010
VI	0.0019 ± 0.009	-0.00001 ± 0.004

Table D.1: **Precision of the maximization problem solutions for the taxi datasets and MEEMC.**

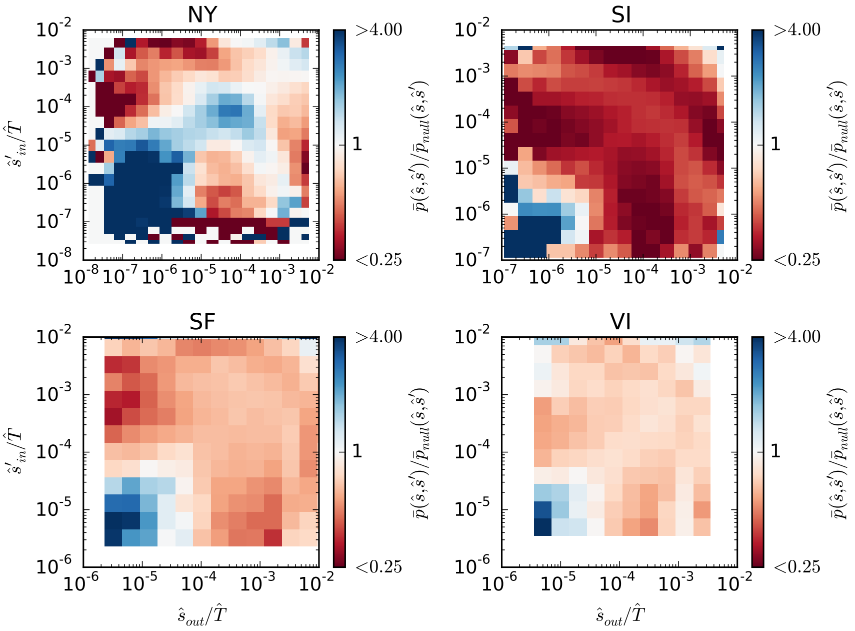

Figure D.12: **Comparison between empirical data and MEECM at edge level.** Relative scaled occupation number as function of starting and ending node strength comparing empirical data and MEECM model over a single run. Both cases are normalized over the bins.

D.4 PITFALLS OF NAIVE EMPIRICAL MODELLING: GRAVITY LAW MULTIVARIATE FITS

In this section we perform a multivariate fit on our data to apply the usual gravity model obtained from multivariate fit to exemplify the weaknesses of the method. For each of the datasets, we fit the relations,

$$\begin{aligned} \ln \langle t_{ij} \rangle_1 &= \alpha_1 \ln \hat{s}_i^{out} + \beta_1 \ln \hat{s}_j^{in} - \gamma_1 d_{ij} + \ln K_1 \\ \ln \langle t_{ij} \rangle_2 &= \alpha_2 \ln \hat{s}_i^{out} + \beta_2 \ln \hat{s}_j^{in} - \gamma_2 \ln d_{ij} + \ln K_2 . \end{aligned} \tag{D.1}$$

We also set $\ln 0 = 0$ to avoid problems related with the presence of self-loops for the power law case. Firstly, thought, to test the sensibility of this analysis to sampling, we also provide a synthetic sample using the MECM (for which we known that $\alpha = \beta = 1$ and $\gamma = 0$) for each dataset. The fits are performed using an Ordinary Least Squares (OLS) implemented in [155] and their results are shown in Table D.2. While the multivariate analysis correctly predicts a very negligible dependence on distance (small γ values) for the MECM samples, a first observed drawback is that both exponents are significantly different from 1. Note also that the R^2 values are not close to the theoretical value of 1 (and the quality of the fit obviously decreases significantly with less sampling).

Concerning the multivariate analysis on the datasets we only display results for the densest case of NY, since in the other cases the sampling is insufficient (the results of the fit are very close to the ones displayed in Table D.2). For the NY case, results are showed in Table D.3, where we also display the CPC values obtained by the model. Taking into account that the MECM model obtains $CPC_{MECM} = 0.65$ using the same inputs and with no fitted parameters, the limitations of the fitting procedure become apparent: they do not even beat the spatially agnostic model.

MODEL	α	β	γ	R^2
NY				
Exp	0.74	0.80	2e-5	0.85
Pow	0.74	0.80	0.02	0.85
SI				
Exp	0.180	0.184	1e-07	0.33
Pow	0.19	0.184	0.005	0.33
SF				
Exp	0.08	0.09	1e-7	0.22
Pow	0.08	0.09	0.006	0.22
VI				
Exp	0.01	0.02	-1e-6	0.06
Pow	0.01	0.02	0.005	0.06

Table D.2: **Multivariate fit of synthetic samples of the MECM.** The adjusted R^2 represent OLS model output results, values transcribed up to the last concordant decimal value in their 95% confidence intervals. The divergence of values $\alpha, \beta \neq 1$ is noteworthy, as well as its dependence on sampling.

MODEL	α	β	γ	R^2	CPC
Exp	0.5	0.5	0.002	0.61	0.65
Pow	0.6	0.5	0.62	0.63	0.60

Table D.3: **Multivariate fit of of the NY dataset.** The adjusted R^2 represent OLS model output results, values rounded up to last concordant decimal value in their 95% confidence intervals. CPC computed with number of trips of the models normalized to match $\sum_{ij} \langle t_{ij} \rangle = \hat{T}$.

D.5 MULTIPLEX DETECTION: ANALYSIS OF INDIVIDUAL EDGE STATISTICS $\{t_{ij}\}$ OVER DIFFERENT SAMPLES OF THE SAME PROCESS

In Chapter 4 the possibility to study the edge related statistics of different network samples generated by the same process is proposed as means to find a correct ensemble to model the system under study. An example of this is shown in Section 8.1, where the study of the rescaled node-pair occupation numbers $\{p_{ij} \equiv t_{ij}/T\}$ is performed for the Taxi datasets using different daily aggregated time slices and coincidence with Poisson statistics is detected.

In the following I investigate numerically the different possible outcomes we could find starting from a given distribution of $\{\langle t_{ij}\rangle\}$ values, comparing the Multi-Edge case (regardless of the number of layers) and Weighted cases with $M = 1$ and $M = 50$ respectively. In doing so, I show how on the one hand that a linear approximation to the values of both $\langle p_{ij}\rangle$ and $\sigma^2_{p_{ij}}$ in the form of (6.4) provides a good asymptotic estimate of the real values, and on the other hand that this prediction can be used to assess the statistics of $\{t_{ij}\}$ and even the number of aggregated layers M of the original system.

Using a linear approximation of the form (6.4), the corresponding values $\langle p_{ij}\rangle$ and $\sigma^2_{p_{ij}}$ for the different cases read,

$$\begin{aligned}
&\text{ME:}\begin{cases} \langle p_{ij}\rangle_{\text{linear}} = \langle t_{ij}\rangle / \langle T\rangle \equiv \hat{p}_{ij} \\ \sigma^2_{p_{ij}}|_{\text{linear}} = \frac{1}{\langle T\rangle}\hat{p}_{ij}(1-\hat{p}_{ij}) \end{cases}\\
&\text{W:}\begin{cases} \langle p_{ij}\rangle_{\text{linear}} = \hat{p}_{ij}\left(1 + M^{-1}(\sum \hat{p}^2_{ij} - \hat{p}_{ij})\right) \\ \sigma^2_{p_{ij}}|_{\text{linear}} = \frac{1}{\langle T\rangle}\hat{p}_{ij}(1-\hat{p}_{ij}) + \frac{\hat{p}^2_{ij}}{M}(1 + \sum \hat{p}^2_{ij} - 2\hat{p}_{ij}). \end{cases}
\end{aligned} \tag{D.2}$$

From which one sees that asymptotically, the relative fluctuations are distinctive in each case and follow,

$$\begin{aligned}
\text{ME:}\quad & \frac{\sigma^2_{p_{ij}}}{\langle p_{ij}\rangle^2} \sim (\langle T\rangle\, \hat{p}_{ij})^{-1} \\
\text{W:}\quad & \frac{\sigma^2_{p_{ij}}}{\langle p_{ij}\rangle^2} \sim M^{-1}
\end{aligned} \tag{D.3}$$

I have simulated $r = 10^7$ different instances of a fixed collection of $L = 99992$ $\{\langle t_{ij}\rangle\}$ quenched values distributed as $p(\langle t\rangle) \sim \langle t\rangle^{-\gamma}$ with $\gamma = 3/2$ (in order to obtain very large $\langle T\rangle = \sum_{ij}\langle t_{ij}\rangle$), generating at each run a collection of random variables $p_{ij} = t_{ij}/\sum_{ij} t_{ij}$.

All the simulations involving random numbers in this Thesis have been performed using the GSL library [89].

I compare $\langle p_{ij}\rangle$ to the results of equations (D.2) in Figure D.13. As one can see, the approximation of $\langle p_{ij}\rangle$ by $\langle p_{ij}\rangle_{\text{linear}}$ is very good (within relative errors of $1/1000$) for the 3 cases while the approximation for σ^2_p fails for small values of $\langle p\rangle$, but this is caused by

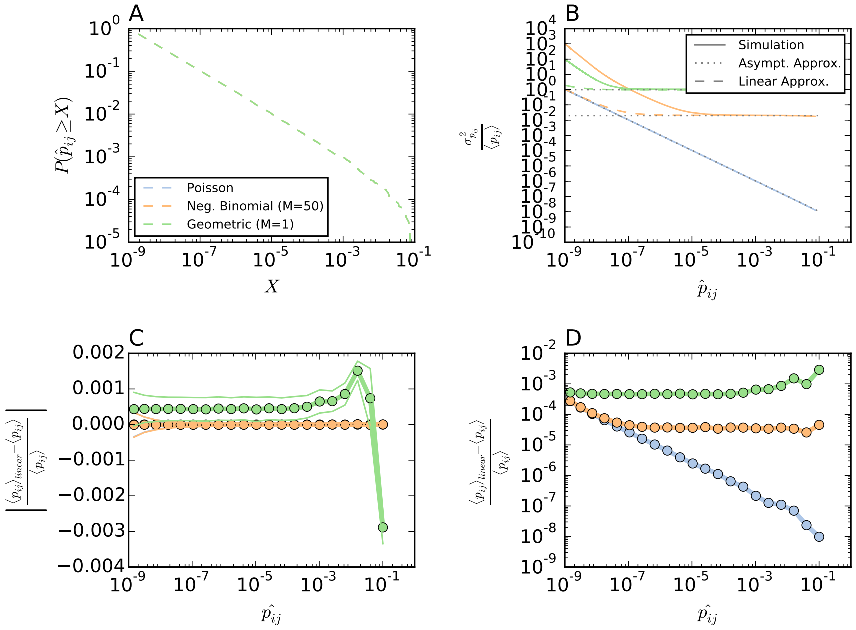

Figure D.13: **Accuracy of linear approximation for p_{ij} moments.** Simulation of $L = 99992$ random variables $p_{ij} = t_{ij}/T$, using a fixed distribution of occupation numbers $\langle t \rangle$ distributed according to $p(\langle t \rangle) \sim \langle t \rangle^{-\gamma}$ with $\gamma = 1.5$ and $\langle T \rangle = 954573797 \simeq 10^9$. Results averaged over $r = 10^7$ repetitions. Dotted and dashed grey lines are set as guides to the eye according to (D.2) and (D.3) respectively. In the negative binomial case we have set $M = 50$.

sparse sampling in the simulation and not by inadequacy of the measured values (the error is reduced as number of reps r increases). The discrepancies are seen for very small values of $\langle p_{ij} \rangle \sim \mathcal{O}((Mr)^{-1})$.

It is important to note, however, that the approximation $\langle p_{ij} \rangle = \hat{p}_{ij}$ does not hold in general (see figure D.14) since for extremely skewed distributions the additional terms are important and thus one needs to apply the complete expression (in the geometric case or the negative binomial case with small number of layers).

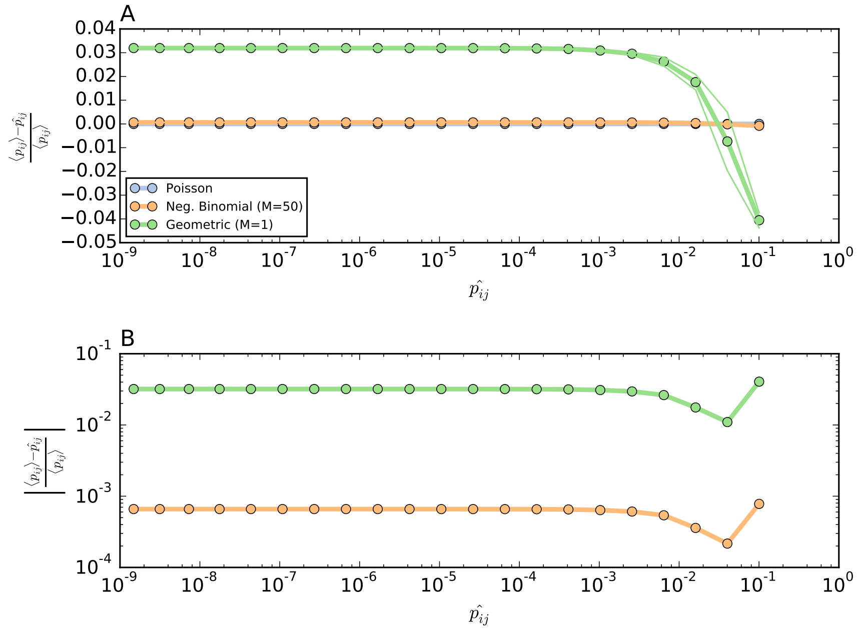

Figure D.14: **Accuracy of raw approximation for p_{ij} average.** Comparison $\hat{p}_{ij}$ vs $\langle p_{ij}\rangle$ with identical simulation parameters as the previous figure. Poisson case not shown as errors are inferior to the machine precision.

PUBLICATIONS

During the development of this Thesis, we have been published several papers. Some contain ideas, figures and derivations that are present in this document. All the software used in this Thesis is also available in the form of public repositories.

LIST OF RELATED PAPERS PUBLISHED DURING THE ELABORATION OF THIS THESIS:

[1] Sagarra, O., Font-Clos, F., Pérez-Vicente, C. J., and Díaz-Guilera, A. (2014). The configuration multi-edge model: Assessing the effect of fixing node strengths on weighted network magnitudes. *EPL*, 107(3):38002.

[2] Sagarra, O., Pérez-Vicente, C. J., and Díaz-Guilera, A. (2013). Statistical mechanics of multi-edge networks. *Phys. Rev. E*, 88(6):062806.

[3] Sagarra, O., Pérez Vicente, C. J., and Díaz-Guilera, A. (2015a). Role of adjacency-matrix degeneracy in maximum-entropy-weighted network models. *Phys. Rev. E*, 92:052816.

[4] Sagarra, O., Szell, M., Santi, P., Díaz-Guilera, A., and Ratti, C. (2015b). Supersampling and network reconstruction of urban mobility. *PLoS ONE*, 10(8):e0134508.

[5] Tachet, R., Sagarra, O., Santi, P., Resta, G., Szell, M., Strogatz, S., and Ratti, C. (2016). Scaling law of urban ride sharing. *Submitted*.

OTHER PAPERS PUBLISHED DURING THE ELABORATION OF THIS THESIS:

[6] Gutiérrez-Roig, M., Sagarra, O., Oltra, A., Bartumeus, F., Diaz-Guilera, A., and Perelló, J. (2015). Active and reactive behaviour in human mobility: the influence of attraction points on pedestrians. *arXiv preprint arXiv:1511.03604*.

[7] Prignano, L., Sagarra, O., and Díaz-Guilera, A. (2013). Tuning synchronization of integrate-and-fire oscillators through mobility. *Phys. Rev. Lett.*, 110(11):1–6.

[8] Prignano, L., Sagarra, O., Gleiser, P. M., and Díaz-Guilera, A. (2012). Synchronization of moving integrate and fire oscillators. *Int. J. Bifurc. Chaos*, 22(07):1250179.

[9] Sagarra, O., Gutierrez-Roig, M., Bonhoure, I., and Perelló, J. (2016). Citizen science practices for computational social science research: The conceptualization of pop-up experiments. *Frontiers in Physics*, 3(93).

LIST OF PUBLIC REPOSITORIES RELATED TO THIS THESIS:

[10] Sagarra, O. (2014). ODME: Origin-Destination Multi-Edge Package.

[11] Sagarra, O. and Font-Clos, F. (2013). Multi Edge Randomizer.

BIBLIOGRAPHY

[12] Ahnert, S. E., Garlaschelli, D., Fink, T. M. A., and Caldarelli, G. (2007). Ensemble approach to the analysis of weighted networks. *Phys. Rev. E*, 76(1):4.

[13] Alhajj, R. and Rokne, J. (2014). *Encyclopedia of social network analysis and mining*. Springer Publishing Company, Incorporated.

[14] Allard, A., Serrano, M., García-Pérez, G., and Boguñá, M. (2016). The hidden geometry of weighted complex networks. *arXiv preprint arXiv:1601.03891*.

[15] Almog, A., Squartini, T., and Garlaschelli, D. (2015). A GDP-driven model for the binary and weighted structure of the International Trade Network. *New J. Phys.*, 17(1):013009.

[16] Anand, K. and Bianconi, G. (2009). Entropy measures for networks: Toward an information theory of complex topologies. *Phys. Rev. E*, 80(4):1–4.

[17] Anand, K., Bianconi, G., and Severini, S. (2011). Shannon and von Neumann entropy of random networks with heterogeneous expected degree. *Phys. Rev. E*, 83(3):036109.

[18] Andersen, M. S., Dahl, J., and Vandenberghe, L. (2013). CVXOPT: A Python package for convex optimization.

[19] Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*.

[20] Annibale, A. and Coolen, A. C. C. (2011). What you see is not what you get: how sampling affects macroscopic features of biological networks. *Interface Focus*, 1(October):26.

[21] Annibale, A., Coolen, A. C. C., Fernandes, L. P., Fraternali, F., and Kleinjung, J. (2009). Tailored graph ensembles as proxies or null models for real networks I: tools for quantifying structure. *J. Phys. A.*, 42(48):485001.

[22] Ansmann, G. and Lehnertz, K. (2011). Constrained randomization of weighted networks. *Physical Review E*, 84(2):026103.

[23] Balcan, D., Colizza, V., Goncalves, B., Hu, H., Ramasco, J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA*, 106(51):21484–21489.

[24] Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.*, 101(11):3747–3752.

[25] Barrat, A., Cattuto, C., Colizza, V., Gesualdo, F., Isella, L., Pandolfi, E., Pinton, J. F., Ravà, L., Rizzo, C., Romano, M., Stehlé, J., Tozzi, A. E., and Broeck, W. (2013). Empirical temporal networks of face-to-face human interactions. *Eur. Phys. J. Spec. Top.*, 222(6):1295–1309.

[26] Barthélemy, M. (2011). Spatial networks. *Phys. Rep.*, 499(1):1–101.

[27] Batty, M. (2013). *The New Science of Cities*. MIT Press.

[28] Bazzani, A., Giorgini, B., Rambaldi, S., Gallotti, R., and Giovannini, L. (2010). Statistical laws in urban mobility from microscopic GPS data in the area of Florence. *J. Stat. Mech. Theor. Exp.*, 2010(05):P05001.

[29] Belik, V., Geisel, T., and Brockmann, D. (2011). Natural human mobility patterns and spatial spread of infectious diseases. *Phys. Rev. X*, 1:011001.

[30] Bender, E. A. and Canfield, E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307.

[31] Bhattacharya, K., Mukherjee, G., and Manna, S. (2007). The international trade network. In *Econophysics of Markets and Business Networks*, pages 139–147. Springer.

[32] Bianconi, G. (2009). Entropy of network ensembles. *Phys. Rev. E*, 79(3):1–10.

[33] Bianconi, G. (2013). Statistical mechanics of multiplex networks: Entropy and overlap. *Phys. Rev. E*, 87(6):62806.

[34] Bianconi, G., Coolen, A. C. C., Vicente, C. J. P., Perez Vicente, C., and Pérez-Vicente, C. J. (2008). Entropies of complex networks with hierarchically constrained topologies. *Phys. Rev. E*, 78(1):016114.

[35] Bianconi, G., Pin, P., and Marsili, M. (2009). Assessing the relevance of node features Definition of Θ. *Proc. Natl. Acad. Sci. U. S. A.*, 106(28):11433–8.

[36] Billings, J. (1874). *Everybody's Friend: Or Josh Billing's Encyclopedia and Proverbial Philosophy of Wit and Humor*. American Publishing Company.

[37] Bleistein, N. and Handelsman, R. A. (1975). *Asymptotic expansions of integrals*. Courier Corporation.

[38] Blondel, V. D., Decuyper, A., and Krings, G. (2015). A survey of results on mobile phone datasets analysis. *Eur. Phys. J. Data Science*, 4(1):1–55.

[39] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theor. Exp.*, 2008(10):P10008.

[40] Boguñá, M. and Pastor-Satorras, R. (2003). Class of correlated random networks with hidden variables. *Phys. Rev. E*, 68(3):36112.

[41] Boguná, M., Papadopoulos, F., and Krioukov, D. (2010). Sustaining the internet with hyperbolic mapping. *Nat. Commun.*, 1:62.

[42] Bollobás, B. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316.

[43] Box, G. E. (1976). Science and statistics. *JASA*, 71(356):791–799.

[44] Box, G. E. and Draper, N. R. (1987). *Empirical model-building and response surfaces*, volume 424. Wiley New York.

[45] Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342.

[46] Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

[47] Burda, Z. and Krzywicki, A. (2003). Uncorrelated random networks. *Phys. Rev. E*, 67(4):46118.

[48] Burger, M. J., Oort, F. G., and Knaap, G. A. (2008). A treatise on the geographical scale of agglomeration externalities and the modifiable areal unit problem. Technical report, ERIM Report Series Research in Management.

[49] Butts, C. T. et al. (2009). Revisiting the foundations of network analysis. *Science*, 325(5939):414.

[50] Cardillo, A., Gómez-Gardeñes, J., Zanin, M., Romance, M., Papo, D., del Pozo, F., and Boccaletti, S. (2013). Emergence of network features from multiplexity. *Sci. Rep.*, 3:1344.

[51] Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J., and Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5(7):e11596.

[52] Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. (2006a). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci. U. S. A.*, 103(7):2015–2020.

[53] Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. (2006b). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci. U. S. A.*, 103(7):2015–20.

[54] Colomer-de Simón, P., Serrano, M. Á., Beiró, M. G., Alvarez-Hamelin, J. I., and Boguñá, M. (2013). Deciphering the global organization of clustering in real complex networks. *Sci. Rep.*, 3.

[55] Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.-P., Sanchez, Á., et al. (2012). Manifesto of computational social science. *Eur. Phys. J. Spec. Top.*, 214(1):325–346.

[56] Coolen, A. C. C., De Martino, A., and Annibale, A. (2009). Constrained Markovian dynamics of random graphs. *J. Stat. Phys.*, 136(6):1035–1067.

[57] Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. (1996). On the Lambert W Function. In *Adv. Comput. Math.*, pages 329–359.

[58] Curie, E. (2001). *Madame Curie: a biography*. Da Capo Press.

[59] de Benedictis, L. and Tajoli, L. (2011). The World Trade Network. *World Econ.*, 34(8):1417–1454.

[60] de Dios Ortuzar, J. and Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons.

[61] De Domenico, M., Nicosia, V., Arenas, A., and Latora, V. (2015). Structural reducibility of multilayer networks. *Nat. Commun.*, 6.

[62] De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., and Arenas, A. (2013). Mathematical formulation of multilayer networks. *Phys. Rev. X*, 3(4):041022.

[63] De Montis, A., Barthélemy, M., Chessa, A., and Vespignani, A. (2007). The structure of interurban traffic: a weighted network analysis. *Env. Plan. B*, 34(5):905–924.

[64] de Pedrolo, M. (1974). *Els elefants son contagiosos. Articles 1962-1972*. Edicions 62, 1 edition.

[65] Del Genio, C. I., Gross, T., and Bassler, K. E. (2011). All scale-free networks are sparse. *Phys. Rev. Lett.*, 107:178701.

[66] Dianati, N. (2016). Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Phys. Rev. E*, 93:012304.

[67] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

[68] Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. (2008). Critical phenomena in complex networks. *Rev. Mod. Phys.*, 80(4):1275.

[69] Erdös, P. and Rényi, A. (1959). On random graphs i. *Publ. Math. Debrecen*, 6:290–297.

[70] Erlander, S. and Stewart, N. F. (1990). *The Gravity Model in Transportation Analysis: Theory and Extensions*.

[71] Fagiolo, G., Reyes, J., and Schiavo, S. (2008). On the topological properties of the world trade web: A weighted network analysis. *Phys. A*, 387(15):3868–3873.

[72] Fedoryuk, M. (1977). *The saddle-point method*. Nauka, Moscow.

[73] Fernandez, A. (2015). Master Thesis: Statistical mechanics of multilayer networks.

[74] Fernandez-Peralta, A. and Toral, R. (2016). Ensemble equivalence for distinguishable particles. *arXiv preprint arXiv:1601.07794*.

[75] Feynman, R. P. (1967). *The character of physical law*, volume 66. MIT press.

[76] Fifield, W. (1964). Pablo Picasso: A composite interview. *Paris Rev.*, 32:66.

[77] Fisher, R. (1938). Presidential Address to the First Indian Statistical Congress, 1938. Sankhya 4, 14-17.

[78] Font-Clos, F., Pruessner, G., Moloney, N. R., and Deluca, A. (2015). The perils of thresholding. *New J. Phys.*, 17(4):043066.

[79] Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.*, 486(3):75–174.

[80] Fronczak, A. and Fronczak, P. (2012). Statistical mechanics of the international trade network. *Phys. Rev. E*, 85(5):056113.

[81] Gallotti, R. and Barthelemy, M. (2014). Anatomy and efficiency of urban multimodal mobility. *Sci. Rep.*, 4.

[82] Garlaschelli, D. and Fisica, D. (2009). The weighted random graph model. *New J. Phys.*, 11(7):73005.

[83] Garlaschelli, D. and Loffredo, M. I. M. (2008). Maximum likelihood: Extracting unbiased information from complex networks. *Phys. Rev. E*, 78(1):1–4.

[84] Garlaschelli, D. and Loffredo, M. I. M. (2009). Generalized bose-fermi statistics and structural correlations in weighted networks. *Phys. Rev. Lett.*, 102(3):2–5.

[85] Gauvin, L., Panisson, A., Cattuto, C., and Barrat, A. (2013). Activity clocks: spreading dynamics on temporal networks of human contact. *Sci. Rep.*, 3:3099.

[86] Glattfelder, J. B. and Battiston, S. (2009). Backbone of complex networks of corporations: The flow of control. *Phys. Rev. E*, 80:036104.

[87] Goh, S., Lee, K., Park, J. S., and Choi, M. (2012). Modification of the gravity model and application to the metropolitan seoul subway system. *Phys. Rev. E*, 86(2):026102.

[88] González, M., Hidalgo, C., and Barabási, A. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

[89] Gough, B. (2009). *GNU scientific library reference manual*. Network Theory Ltd.

[90] Guicciardini, F., Rosini, G., and Botta, C. (1832). *Storia d'Italia*, volume 7. Presso Baudry.

[91] Guimerà, R., Arenas, A., Díaz-Guilera, A., and Giralt, F. (2002). Dynamical properties of model communication networks. *Phys. Rev. E*, 66(2):26704.

[92] Hamming, R. (1962). *Numerical Methods for Scientists and Engineers*. McGraw-Hill, 1 edition.

[93] Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *J. Am. Stat. Assoc.*, 76(373):33–50.

[94] Hufnagel, L., Brockmann, D., and Geisel, T. (2004). Forecast and control of epidemics in a globalized worlds. *Proc. Natl. Acad. Sci. U. S. A.*, 101:15124–15129.

[95] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620.

[96] Jaynes, E. T. (1992). The gibbs paradox. In *Maximum entropy and bayesian methods*, pages 1–21. Springer.

[97] Johnson, S., Torres, J. J., Marro, J., and Muñoz, M. a. (2010). Entropic Origin of Disassortativity in Complex Networks. *Phys. Rev. Lett.*, 104(10):108702.

[98] Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python.

[99] Jung, W.-S., Wang, F., and Stanley, H. E. (2008). Gravity model in the korean highway. *EPL*, 81(4):48005.

[100] Kaluza, P., Kölzsch, A., Gastner, M. T., and Blasius, B. (2010). The complex network of global cargo ship movements. *J. R. Soc. Interface*, 7(48):1093–103.

[101] Kang, C., Liu, Y., Guo, D., and Qin, K. (2015). A generalized radiation model for human mobility: Spatial scale, searching direction and trip constraint. *PLoS One*, 10(11).

[102] Karp, R. M. (1972). *Reducibility among combinatorial problems*. Springer.

[103] Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J., Moreno, Y., and Porter, M. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

[104] Knop, R. A. (2016). Tycho, Kepler, & Newton : a Story in the Progress of Science. http://scienceblogs.com/interactions/2007/03/20/tycho-kepler-newton-a-story-in-1/.

[105] Kölbl, R. and Helbing, D. (2003). Energy laws in human travel behaviour. *New J. Phys.*, 5(1):48.

[106] Kotz, D., Henderson, T., Abyzov, I., and Yeo, J. (2009). CRAWDAD dataset dartmouth/campus (v. 2009-09-09).

[107] Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. (2009). Urban gravity: a model for inter-city telecommunication flows. *J. Stat. Mech. Theor. Exp.*, 2009(07):L07003.

[108] Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguná, M. (2010). Hyperbolic geometry of complex networks. *Phys. Rev. E*, 82(3):036106.

[109] Kurtz, Gary (1977). Movie. star wars: A new hope.

[110] Kuzemsky, A. (2014). Thermodynamic limit in statistical physics. *Int. J. Mod. Phys. B*, 28(09):1430004.

[111] Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34(1):1–14.

[112] Lenormand, M., Bassolas, A., and Ramasco, J. J. (2016). Systematic comparison of trip distribution laws and models. *J. Transp. Geogr.*, 51:158–169.

[113] Lenormand, M., Huet, S., Gargiulo, F., and Deffuant, G. (2012). A universal model of commuting networks. *PLoS ONE*, 7(10):e45985.

[114] Lenormand, M., Picornell, M., Cantú-Ros, O. G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frías-Martínez, E., and Ramasco, J. J. (2014). Cross-checking different sources of mobility information. *PLoS One*, 9(8):e105184.

[115] Liang, X., Zhao, J., Dong, L., and Xu, K. (2013). Unraveling the origin of exponential law in intra-urban human mobility. *Sci. Rep.*, 3.

[116] Liang, X., Zheng, X., Lv, W., Zhu, T., and Xu, K. (2012). The scaling of human mobility by taxis is exponential. *Phys. A*, 391(5):2135–2144.

[117] Lotero, L., Cardillo, A., Hurtado, R., and Gómez-Gardeñes, J. (2014). Several multiplexes in the same city: The role of wealth differences in urban mobility. *arXiv preprint arXiv:1408.2484*.

[118] Louail, T., Lenormand, M., Picornell, M., Cantú, O. G., Herranz, R., Frias-Martinez, E., Ramasco, J. J., and Barthelemy, M. (2015). Uncovering the spatial structure of mobility networks. *Nat. Commun.*, 6.

[119] Louf, R. (2015). Wandering in cities: a statistical physics approach to urban theory. *arXiv Prepr. arXiv1511.08236*.

[120] Lukacs, E. (1970). *Characteristic functions*. Hafner Publishing Company, London.

[121] Maddox, B. (2012). Rosalind franklin: The dark lady of dna.

[122] Mastrandrea, R., Squartini, T., Fagiolo, G., and Garlaschelli, D. (2014). Enhanced reconstruction of weighted networks from strengths and degrees. *New J. Phys.*, 16(4):043022.

[123] Masucci, A. P., Serras, J., Johansson, A., and Batty, M. (2013). Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Phys. Rev. E*, 88(2):022812.

[124] Menichetti, G., Remondini, D., and Bianconi, G. (2014). Correlations between weights and overlap in ensembles of weighted multiplex networks. *Phys. Rev. E*, 90(6):62817.

[125] Molloy, M. and Reed, B. A. (1995). A critical point for random graphs with a given degree sequence. *Random Struct. algorithms*, 6(2/3):161–180.

[126] Nagle, J. F. (2004). Regarding the entropy of distinguishable particles. *J. Stat. Phys.*, 117(5-6):1047–1062.

[127] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.*, 103(23):8577–8582.

[128] Newman, M. E. J. (2010). *Networks: an introduction*. Oxford University Press.

[129] Newman, M. E. J., Clauset, A., and Shalizi, C. R. (2000). Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703.

[130] Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York.

[131] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. (2012). A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS One*, 7:e37027.

[132] OpenStreetMap Collaborators (2004–). OpenStreetMap.org.

[133] Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., and Barabási, A.-L. (2015). Returners and explorers dichotomy in human mobility. *Nat. Commun.*, 6.

[134] Park, J. and Newman, M. E. J. (2004). Statistical mechanics of networks. *Phys. Rev. E*, 70(6):066117.

[135] Pathria, R. K. (1996). *Statistical Mechanics, Second Edition*. Butterworth-Heinemann, 2 edition.

[136] Peixoto, T. P. (2011). Entropy of stochastic blockmodel ensembles. *Phys. Rev. E*, 85(5):1–16.

[137] Peixoto, T. P. (2015). Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups. *Phys. Rev. X*, 5:011033.

[138] Perra, N., Gonçalves, B., Pastor-Satorras, R., and Vespignani, A. (2012). Activity driven modeling of time varying networks. *Sci. Rep.*, 2.

[139] Piorkowski, M., Sarafijanovic-Djukic, N., and Grossglauser, M. (2009). A Parsimonious Model of Mobile Partitioned Networks. In *IEEE Conf. Commun. Syst. Networks*, pages 1–10. IEEE.

[140] Popović, M., Štefančić, H., and Zlatić, V. (2012). Geometric Origin of Scaling in Large Traffic Networks. *Phys. Rev. Lett.*, 109(20):208701.

[141] Pukelsheim, F. and Simeone, B. (2009). On the iterative proportional fitting procedure: Structure of accumulation points and l1-error analysis.

[142] QGIS Development Team (2009). *QGIS Geographic Information System*. Open Source Geospatial Foundation.

[143] Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Information filtering in complex weighted networks. *Phys. Rev. E*, 83:046101.

[144] Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M. C., and Toroczkai, Z. (2014). Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nat. Commun.*, 5.

[145] Roberts, E. and Coolen, A. (2012). Unbiased degree-preserving randomization of directed binary networks. *Physical Review E*, 85(4):046103.

[146] Roberts, E. S. and Coolen, A. C. C. (2014). Entropies of tailored random graph ensembles: bipartite graphs, generalized degrees, and node neighbourhoods. *J. Phys. A*, 47(43):435101.

[147] Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p *) models for social networks. *Soc. Networks*, 29(2):173–191.

[148] Roth, C., Kang, S. M., Batty, M., and Barthélemy, M. (2011). Structure of urban movements: polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6(1):e15923.

[149] Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., and Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareability networks. *Proc. Natl. Acad. Sci. U. S. A.*, 111(37):13290–13294.

[150] Sayer, R. A. (1976). A critique of urban modelling: from regional science to urban and regional political economy. *Progress in Planning*, 6:187–254.

[151] Scellato, S., Musolesi, M., Mascolo, C., Latora, V., and Campbell, A. T. (2011a). Nextplace: A spatio-temporal prediction framework for pervasive systems. In *Proceedings of the 9th International Conference on Pervasive Computing*, Pervasive'11, pages 152–169, Berlin, Heidelberg. Springer-Verlag.

[152] Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. (2011b). Socio-spatial properties of online location-based social networks. *Proc Int AAAI Conf Weblogs Soc Media*, 11.

[153] Schlitt, T. and Coolen, A. C. C. (2011). Tailored graph ensembles as proxies or null models for real networks II: results on directed graphs. *J. Phys. A*, 44(27):21.

[154] Schneider, M. (1959). Gravity models and trip distribution theory. *Pap. Reg. Sci.*, 5(1):51–56.

[155] Seabold, J. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*.

[156] Seidel, J. (1976). A survey of two-graphs. *Colloquio Internazionale sulle Teorie Combinatorie (Rome, 1973)*, 1:481–511.

[157] Serfling, R. (1978). Some elementary results on Poisson approximation in a sequence of Bernoulli trials. *Siam Rev.*, 20(3):567–579.

[158] Serrano, M. A., Boguñá, M., and Vespignani, A. (2007). Patterns of dominant flows in the world trade web. *J. Econ. Interact. Coord.*, 2(2):111–124.

[159] Serrano, M. A. and Boguna, M. (2005). Weighted Configuration Model. *AIP Conf. Proc.*, 776(1):101–107.

[160] Serrano, M. A., Boguna, M., and Pastor-Satorras, R. (2006). Correlations in weighted networks. *Phys. Rev. E*, 74(5 Pt 2):55101.

[161] Serrano, M. Á., Boguñá, M., and Sagués, F. (2012). Uncovering the hidden geometry behind metabolic networks. *Mol. Biosyst.*, 8(3):843–850.

[162] Serrano, M. Á., Boguná, M., and Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.*, 106(16):6483–6488.

[163] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

[164] Simini, F., González, M., Maritan, A., and Barabási, A. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100.

[165] Simini, F., Maritan, A., and Néda, Z. (2013). Human mobility in a continuum approach. *PLoS One*, 8(3):e60069.

[166] Slater, P. B. (2009a). Multiscale network reduction methodologies: Bistochastic and disparity filtering of human migration flows between 3,000+ us counties. *arXiv preprint arXiv:0907.2393*.

[167] Slater, P. B. (2009b). A two-stage algorithm for extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.*, 106(26):E66–E66.

[168] Song, C., Koren, T., Wang, P., and Barabási, A. (2010a). Modelling the scaling properties of human mobility. *Nature Physics*, 6:818–823.

[169] Song, C., Qu, Z., Blumm, N., and Barabási, A. (2010b). Limits of predictability in human mobility. *Science*, 327(5968):1018.

[170] Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34.

[171] Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.*, 15(1):72–101.

[172] Spieser, K., Treleaven, K., Zhang, R., Frazzoli, E., Morton, D., and Pavone, M. (2014). Towards a systematic approach to the design and evaluation of automated mobility-on-demand systems: a case study in singapore. *Road Vehicle Automation*, pages 229–245.

[173] Sporns, O. (2012). *Discovering the human connectome*. MIT press.

[174] Squartini, T., de Mol, J., den Hollander, F., and Garlaschelli, D. (2015). Breaking of ensemble equivalence in networks. *Phys. Rev. Lett.*, 115:268701.

[175] Squartini, T., Fagiolo, G., and Garlaschelli, D. (2011). Randomizing world trade. II. A weighted network analysis. *Phys. Rev. E*, 84(4):046117.

[176] Squartini, T. and Garlaschelli, D. (2011). Analytical maximum-likelihood method to detect patterns in real networks. *New J. Phys.*, 13(8):083001.

[177] Stouffer, S. A. (1940). Intervening Opportunities: A Theory Relating Mobility and Distance. *American Sociological Review*, 5(6):845–867.

[178] Swendsen, R. H. (2011). How physicists disagree on the meaning of entropy. *Am. J. Phys.*, 79(4):342.

[179] Swendsen, R. H. (2015). The ambiguity of distinguishability in statistical mechanics. *Am. J. Phys.*, 83(6):545–554.

[180] Szell, M., Sinatra, R., Petri, G., Thurner, S., and Latora, V. (2012). Understanding mobility in a social petri dish. *Sci Rep*, 2:457.

[181] Thiemann, C., Theis, F., Grady, D., Brune, R., and Dirk Brockmann, D. (2010). The structure of borders in a small world. *PLoS ONE*, 5:e15422.

[182] Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Econ. Geogr.*, pages 234–240.

[183] Touchette, H. (2015). Equivalence and nonequivalence of ensembles: Thermodynamic, macrostate, and measure levels. *Journal of Statistical Physics*, 159(5):987–1016.

[184] Traag, V. A. (2015). Louvain: Community detection module for Python.

[185] United Nations (2014). *World Urbanization Prospects 2014: Highlights*. United Nations Publications.

[186] Van Wijk, B. C., Stam, C. J., and Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. *PLoS One*, 5(10):e13701.

[187] Versteegh, M. A. and Dieks, D. (2011). The gibbs paradox and the distinguishability of identical particles. *Am. J. Phys.*, 79(7):741–746.

[188] Waxman, B. M. (1988). Routing of multipoint connections. *Selected Areas in Communications, IEEE Journal on*, 6(9):1617–1622.

[189] Weiner, E. (1983). Urban transportation planning in the us: An historical overview. Technical report.

[190] Wilson, A. G. (1970). A statistical theory of spatial distribution models. 1(3):253–269.

[191] Wilson, A. G. (2010). Entropy in Urban and Regional Modelling: Retrospect and Prospect. *Geogr. Anal.*, 42(4):364–394.

[192] Wong, D. (2009). *The SAGE handbook of spatial analysis*, chapter The Modifiable Areal Unit Problem (MAUP), pages 105–123. SAGE.

[193] Wormald, N. C. (1980). Some problems in the enumeration of labelled graphs. *Bulletin of the Australian Mathematical Society*, 21(01):159–160.

[194] Yan, X.-Y., Han, X.-P., Wang, B.-H., and Zhou, T. (2013). Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Sci. Rep.*, 3.

[195] Yan, X.-Y., Zhao, C., Fan, Y., Di, Z., and Wang, W.-X. (2014). Universal predictability of mobility patterns in cities. *J. R. Soc. Interface*, 11(100):20140834.

[196] Yang, Y., Herrera, C., Eagle, N., and González, M. C. (2014). Limits of predictability in commuting flows in the absence of data for calibration. *Sci. Rep.*, 4.

[197] Zahavi, Y. and Talvitie, A. (1980). *Regularities in travel time and money expenditures*. Number 750.

[198] Zipf, G. K. (1946). The p 1 p 2/d hypothesis: on the intercity movement of persons. *American sociological review*, 11(6):677–686.

COLOPHON

Text License

Software License

Thesis model and design acknowledgement

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Final Version as of April 7, 2016 (`classicthesis` versió 1.0).