UNIVERSITÀ DEGLI STUDI DI PALERMO
SCUOLA POLITECNICA

Corso di Laurea Magistrale in Statistica e Data Science
Dipartimento di Scienze Economiche, Aziendali e Statistiche

EXPONENTIAL RANDOM GRAPH MODELS FOR NEURAL MICROCIRCUITS

TESI DI LAUREA DI
SALVATORE LATORA

RELATORE
Prof. LUIGI AUGUGLIARO

CORRELATORE
Dott.ssa. PAOLA VITALE

ANNO ACCADEMICO 2021 - 2022

MAGISTRALE

A mio padre e a mia madre, a mia sorella
e a tutta mia famiglia per essermi
stati sempre accanto durante
questo percorso di studi,
spronandomi e incoraggiandomi
a dare sempre il meglio.

Al mio relatore, prof. Luigi Augugliaro,
e correlatore, Dott.ssa Paola Vitale
per la disponibilità e la pazienza
durante la realizzazione della tesi,
fornendomi conoscenze e suggerimenti.

Ai miei colleghi di università per
avermi sempre aiutato nei
momenti di difficoltà, condividendo
con me gioie e fatiche di
questi anni trascorsi insieme.

Ai miei amici, per esserci sempre stati.

Contents

Chapter 1

An Introduction to Brain Networks

1.1 Introduction

Consisting of trillions of neurons and synaptic connections, the *brain* is one of the best-known complex systems in nature, and its associated pathologies, such as dementia, schizophrenia, or Alzheimer's, are disabling diseases for which no cure is known today and on which many scholars are working.
Usually, in the collective imagination, one imagines the brain as a large network, and in fact, this is exactly how it is represented in applications in the field of computational neuroscience. For experts in the field, very important is what is called the *connectome*, i.e. a matrix that represents all the connections between neurons through what are synapses but not only, in fact, the term connectome is often also used to identify the system of connections that exists between the macro areas that make up the brain.
Today, thanks to technological and scientific developments, we have been able to understand a lot about how the brain works on a cellular level and the interactions that exist between the various parts of the brain. Of particular importance is what is known as *complex network science*: in this context, using the mathematical formalism of what is known as *graph theory*, provides a

whole series of fundamental and effective tools for the study of brain networks. This chapter will provide a brief introduction to how network science and graph theory can help to model, estimate, and simulate the topology of brain networks, and what the different approaches to analysing brain networks are, using as a main source of knowledge the book by Fornito et al. (2016) to which reference is made for further details.

1.2 Main Concepts and History of Graph Theory

Graph theory is often used to study and understand many real systems found in nature. The first application of it was by Leonhard Euler with the famous Konigsberg Bridge problem, in which he used a graph to represent the various areas of the city, to try to find a path that entered the city centre and passed once and only once overall the city's bridges. Over the years, the concept of a graph was extended to include networks, i.e. graphs whose nodes and/or edges have attributes, which may be quantitative or qualitative.
Formally, a graph is a set $G = (V, E)$, where V is the set of nodes, while E is the set of edges existing between the nodes. Indicating an edge between generic nodes i and j with the pair (i, j), this pair belongs to the set E only if a direct connection exists between the two nodes. From the graph, it is possible to define the so-called adjacency matrix A, of dimension $n \times n$ where n identifies the number of nodes. This matrix is a binary matrix and the generic element equal to 1 if and only if a direct edge exists between the two nodes. In the case where the graph is undirected, the matrix A will be symmetric, whereas in the case where the graph is directed, the matrix A will be non-symmetric.
What graph theory and network analysis provide are tools for studying the topology of the graph/network, i.e. those characteristics that underlie the organization and formation process of edges between nodes. The first important statistical tool was provided in the work of Erdos and Rényi (1984), who introduced a generative model, also known as the *Exponential model*, with

the aim of capturing the topological characteristics of the network and generating random graphs, through a constant probability p of observing or not observing an edge between two nodes. Subsequently, other scholars have made fundamental contributions to network science, such as the work of Watts and Strogatz (1998) in which they develop another generative model for random networks having some specific topological characteristics, such as a large clustering coefficient and small mean path length, which later gives rise to the discovery of so-called *small world networks*, whose properties are observed in many real-world networks, from social networks to gene networks. It was precisely this work that was one of the first points of contact between neuronal connectomics and modern network science because the network studied represented the neuronal connectivity of the worm *C. Elegans*. Another important contribution was made by Barabási and Albert (1999), developing a generative model, named after them, capable of generating so-called *scale-free networks*, i.e. networks in which the degree distribution follows a *power law* distribution. Such networks are characterized by the fact that most nodes have a higher degree than the average degree forming *hubs*, i.e. nodes with high degrees that are assumed to play a key role within the network. The scale-free characteristic has been observed in many real-world complex networks, and there may be several processes that lead to the emergence of this property, such as the process of *preferential attachment*, i.e. the tendency of nodes to connect to other nodes with similar or higher degrees. Finally, a final contribution to the study of complex network topologies has been made by the work of Girvan and Newman (2002), which emphasizes the presence of *modules/groups* or *communities* of nodes within the network, i.e. groups of nodes that are highly connected internally and loosely connected to each other, and provide a measure of how well the network is divided into groups through *modularity*.

1.3 Brain Networks and Connectivity

Brain networks are complex networks, but at the same time, they are dynamic systems concerning time and space. Dynamics and network topology influence each other. Think, for example, of the distance between neurons: it is very unlikely that two neurons far apart have a synaptic connection; therefore, spatial dynamics in this case plays a fundamental role in determining the topology of the network, as it leads the network to preserve the wiring cost.
The analysis of brain networks takes place at both the microscale and macroscale. The former aims at analysing connectivity at the cellular level, while the latter aims at analysing how cortical areas connect to each other. The latter approach has been very important for studying mental illnesses.

1.3.1 Connectivity at the Microscale

The study of neuronal connectivity at the microscale aims to study the topology of networks in which nodes represent neurons and edges represent the synaptic connections between neurons. One of the main pioneers in this field was Santiago Ramón y Cajal, who reconciled microscale connectivity with the principles of brain dynamics, providing the first topological characteristics of brain networks, including minimizing the cost of wiring synaptic connections due to the spatial dislocation of nodes, as mentioned above, and minimizing the delay in the propagation of information between neurons.
The study of neuronal connectivity at the microscale has been made possible by tract-tracing data. The first important work was that of White et al. (1986), in which they succeeded in reconstructing the entire neuronal connectivity of the *C. Elegans* worm. At the beginning of the 21st century, thanks to the availability of tract-tracing data on cats and monkeys, several discoveries were made about the properties that characterize the brain networks of these animals; these networks were found to have 'small-world' properties, i.e. a short path length and a high clustering coefficient, the same properties that the neuronal networks of C.Elegans were found to have, through the work of Watts

and Strogatz (1998). Recently, however, scientific research has focused on simulating neuronal connectivity on a large scale. For example, **?** proposed in his work a theoretical framework based on Erdos Reny's Exponential model to try to capture and generalize the topological properties of neuronal connectivity to both understand the brain's information processing mechanisms and simulate large-scale neuronal connectivity networks with topological characteristics identical to the observed networks.

1.3.2 Connectivity at the Macroscale

The study of macroscale connectivity concerns how the various areas of the brain are interconnected, and aims to explain a possible correlation between symptoms of brain diseases and pathological lesions in the brain.
The study of macroscale connectivity for the study of mental illnesses and disorders, developed at the beginning of the 20th century, was quickly put on the back burner due to both the methodological weakness of the models available at the time and the poor quality of the data available to researchers. In fact, a substantial difference in the study of connectivity on the microscale was that on the latter, researchers had very accurate data on neurons and neural microcircuits. The idea behind the study of psychic disorders by means of macroscale network models was that all psychic aspects of the human mind (e.g., understanding, emotion, and so on) were related to the functions of particular brain areas and that therefore psychic disorders and diseases affected single, specific areas of the brain. Subsequently, this approach was superseded; many papers showed that in fact many psychological disorders affected more areas of the brain, as opposed to single, specific areas as previously thought, and thus various large-scale network models such as the one shown in Figure 1.1 emerged.
With the availability of better quality data, the analysis of psychological and brain disorders using network models has become another important area of research on the human brain.

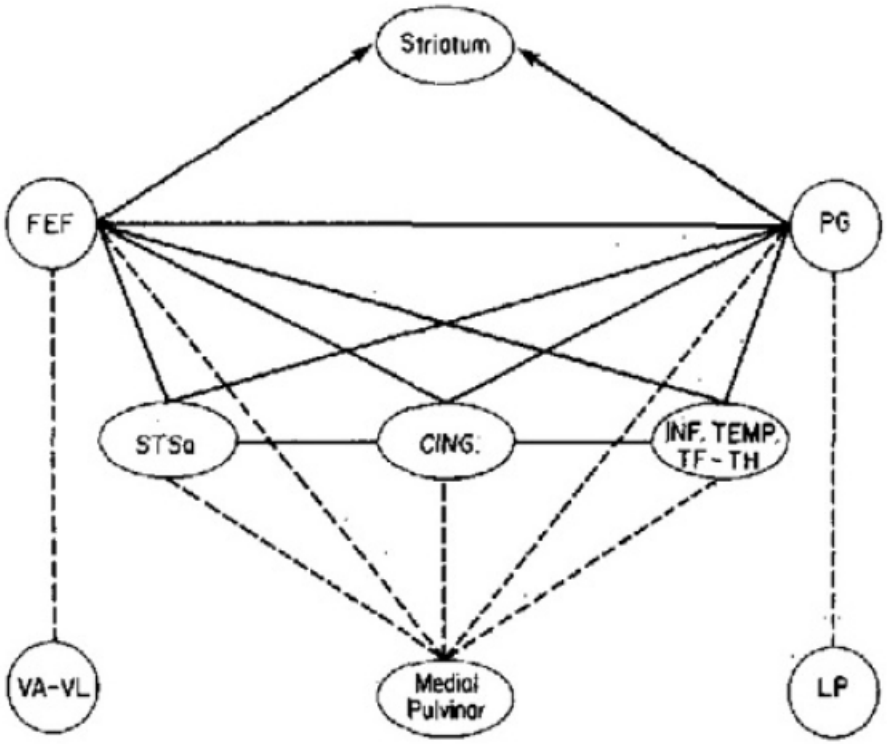

Figure 1.1: Brain graphs based on clinicopathological correlations (spatial attention), from Fornito et al. (2016)

1.3.3 Neuroimaging and Human Brain

Another line of research was interested in how the dynamics of the brain system related to different areas of the brain: this is called *functional connectivity*. In particular, one tries to understand whether two brain areas are functionally connected by analyzing the correlation between two time series of neurophysiological signals, which refer to the two brain areas, obtained through various neuroimaging techniques, such as magnetic resonance imaging or electroencephalography and so on. The data used are therefore called *functional MRI and/or M/EEG data*. Easy to obtain, such data have opened the door to the study of human brain functional networks. If there is coherence between two sets of signals, the corresponding areas of the brain are represented by nodes

and connected to each other via an edge, indicating that they are functionally connected. This leads to the construction of the large-scale functional network; examples of these networks are shown in Figure 1.2. Analysis of human brain networks using MRI data has shown that such networks have the properties of all natural complex systems, showing the small-world property and a hierarchical module structure. The main difference between human brain functional

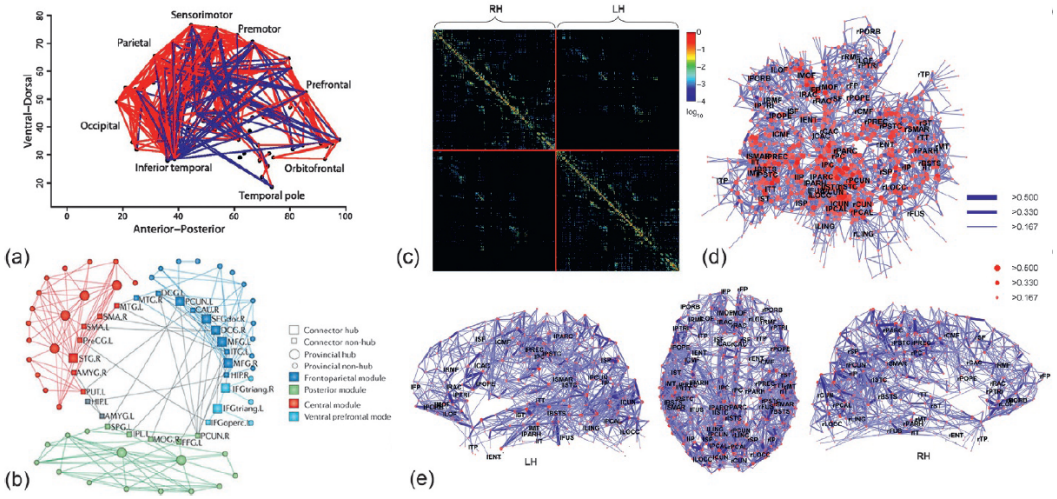

Figure 1.2: Brain graph from human magnetic resonance (MRI), from Fornito et al. (2016).

networks and microscale brain networks is that the latter are constructed from the connectivity between neurons, which constitutes the connectome, and thus is anatomical connectivity that has very little change over time, whereas brain functional networks are networks whose connectivity changes a lot over time and does not necessarily imply a corresponding anatomical connection between different areas of the brain.

Various contributions have been made by the MRI connectome analysis of the human being. For example, it has been possible to understand the modular topology of these networks in relation to the cognitive functions of the human being: each module, comprising different areas of the brain, is specialized in certain cognitive functions such as perception, feelings, and so on. Moreover,

MRI connectome analysis also has some clinical implications in diagnosing neurodegenerative diseases. Indeed, certain topological features of the network, such as the presence of hubs, are closely related to certain physiological situations underlying the manifestations of diseases such as Alzheimer's or dementia.
In conclusion, the analysis of functional connectome topology may provide important biomarkers for early diagnosis and prediction of clinical outcomes in neurology and psychiatry.

1.4 Are Graph Theory and Connectomics Useful?

Graphs are simple models and the mathematics of graphs is both rigorous and, at the same time, simple. It follows that these models are easily accessible to various scientists, such as neuroscientists, physiologists, and so on. Many key concepts of connectomics can be represented graphically and illustrated by analogy with well-known complex systems. A graph cannot be used to model every single detail of the brain, but it can provide answers as to how the structure of the brain network constraints functioning, what the general organizational principles of brain networks are, and it allows us to understand what developmental processes can give rise to networks that look and function like the brain. Furthermore, as mentioned earlier, graph theory can provide clinical support for the diagnosis of mental and neurological diseases. The generalisability of graph theory is another advantage; in fact, graphs have already been used to model a wide range of complex systems. In neuroscience, the applicability of graph theory to all kinds of neuroimaging and neurophysiological data allows us to examine the same topological and spatial properties of brain networks in a wide range of species, scales, and modalities. One can conclude by saying that graph theory and network science are very useful for studying connectomics, however, as we will also see later, the use of the such methodology in this field is severely limited by the available technology required to

conduct these types of analyses.

Chapter 2

Probabilistic Models for Networks

In this section, a general introduction to statistical models for networks will be given, with a particular focus on the model used for the network in question, the *Exponential Random Graph Models*, specifying its formulation and providing details on how it is possible to model different network structures, both for undirected and directed networks, on the model fitting process and on the methodology that allows us to evaluate its goodness of fit.

2.1 Why a Model for Networks?

Several measures have been defined which allow to describe an observed network (such as density, degree, clustering coefficient and so on), but the advantage of using a statistical model are many:

- for complex network, statistical model allows us to capture the link generator process in the network, it allow us to take into account a certain variability that we are not able to model, and most importantly, it allows us to estimate the parameters of the observed data generating model (and the corresponding uncertainty) or estimate the distribution of the data under a precise model specification;
- statistical model allows us to make inference on a specified network sub-

structure, modeled by one or more parameters, and to understand if this substructure is present by change or not. We can build hypotheses about the process that generate these substructure and test it;

- since different processes can generate very similar network structures, a statistical model allows us to model the different processes and understand what is the process from which the observed structure in the network comes from.

2.2 Exponential Random Graph Models

Exponential Random Graph Models (*ERGM*), also known as p^* model, proposed by Frank and Strauss (1986), are a class of models sufficiently large to allow the modeling of both dense and sparse random networks, with specific characteristics in terms of in- and out-degree and, at the same time, characterized by a complex structure of probabilistic dependence.
The aim of *ERGM* is to identify the process of link creation. In this sense, each link between node i and j is assumed to be a *random variable* A_{ij}, which is equal to 1 if there is an edge between node i and j, 0 otherwise. We denoted by $\mathbf{A}$ the adjacency matrix (can be symmetric for undirected graph or asymmetric for directed graph) that contains these random variables, and we denote by $\mathbf{a}$ the adjacency matrix related to the observed graph. In other words, the observed network is a realization from a set of possible networks; the range of possible networks, and the corresponding probability of occurrence under the model, is represented by a probability distribution on the set of all possible graphs with the same number of nodes. In the simplest specification, an *ERGM* assumes:

$$P_{\boldsymbol{\theta}}\left(\boldsymbol{A}=\boldsymbol{a}|\boldsymbol{X}=\boldsymbol{x}\right)=k\left(\boldsymbol{\theta}^{-1}\right)\exp\left(\boldsymbol{\theta}^{T}g\left(\boldsymbol{s}\left(\boldsymbol{a}\right),\boldsymbol{x}\right)\right), \tag{2.1}$$

where $k\left(\boldsymbol{\theta}\right)=\sum_{\boldsymbol{a}}\exp\left(\boldsymbol{\theta}^{T}g\left(\boldsymbol{s}\left(\boldsymbol{a}\right),\boldsymbol{x}\right)\right)$ is the normalization constant and $\mathbf{X}$ denote the vector of *exogenous* explanatory variables (in this contest, they

could be morphological characteristics of neurons, or their spatial location). The vector $g(\boldsymbol{s}(\boldsymbol{a}), \boldsymbol{x})$ contains the *endogenous* statistics with the topological structure of the network and the link formation process. Finally, the vector $\boldsymbol{\theta}$ is the vector of parameters for both exogenous and endogenous effect.
The *endogenous* statistics allow us to specified a *dipendence structure* between edges in the network, relaxing the hypothesis of independence in link formation. In fact, without endogenous statistics the model assumes independence in the link formation. Several endogenous statistics will be described that allow us to specify different models.

2.3 Dependence Assumptions and Models

Thanks to the Hammersley-Clifford theorem, it can be shown that well-specified dependence hypotheses imply a particular class of models (Besag, 1974). Here the most important dependence assumptions and the resulting models will be discussed to, all belonging the ERGM family.

2.3.1 Bernoulli Graphs

Bernoulli random graph distributions are obtained when there is the assumption of *independence* in the process of creating edges, whereby connections between nodes occur randomly according to a fixed probability α. From (2.1), assuming that it has no endogenous effects, the model takes the following form:

$$P(\boldsymbol{A} = \boldsymbol{a}) = \frac{1}{k} \exp\left(\sum_{ij} \theta_{ij} a_{ij}\right). \tag{2.2}$$

In this case, the network statistic $g(a_{ij}) = a_{ij}$ tells us if the edge between node i and j is present or not. By assuming *homogeneity*, i.e. there is a fixed

probability for all possible edges across the network, than $\theta_{ij} = \theta$, hence

$$P(\boldsymbol{A} = \boldsymbol{a}) = \frac{1}{k} \exp\left(\theta \sum_{ij} a_{ij}\right), \tag{2.3}$$

where $\sum_{ij} a_{ij}$ is the number of edges in the observed network and θ, which is called *edge* or *density* parameter, is related to the probability of observing an edge between the two nodes:

$$P(A_{ih} = 1) = \frac{\exp(\theta)}{(1 + \exp(\theta))} = \alpha, \forall i, j \tag{2.4}$$

2.3.2 Dyadic Models

For directed networks, a more complicated assumption is the *dyadic independence* assumption: dyads, i.e. the edges between a pair of nodes, are independent of one another. With homogeneity assumption, the model takes the following form:

$$P(\boldsymbol{A} = \boldsymbol{a}) = \frac{1}{k} \exp\left(\theta \sum_{ij} a_{ij} + \rho \sum_{ij} a_{ij} a_{ji}\right). \tag{2.5}$$

Since in directed network $a_{ij} \neq a_{ji}$, then $\sum_{ij} a_{ij} a_{ji}$ is the number of reciprocal edges in $\boldsymbol{a}$ and is called *reciprocity* statistic. So with this independence assumption we have two types of network configuration in the model: *single edges* and *reciprocated edges*.
In case of undirected networks, Bernoulli graph and Dyadic model are identical, and ρ in (2.5) is irrelevant.

2.3.3 Markov Random Graph Models

Bernoulli and Dyadic independence are unrealistic assumptions in many cases. Frank and Strauss (1986) introduced a new concept of dependence, called *Markov dependence*, which gave rise to the so-called *Markov graph*. Frank and

Strauss provide the following definition: "A graph is said to be Markov graph if only incident dyads can be conditionally dependent". In other words, the edge between nodes i and j is assumed to be *conditional independent* given the values of all other connections in the network. Many endogenous dependencies can be captured as *Markov graph statistics*. In their paper, Frank and Strauss (1986) considered only three Markov graph statistics: the number of edges, the *k-star* and the number of triangles. Over the years, the list of Markov statistics that can be used in models has been expanded. Here we will deal with the most important ones and some of the configuration graph they capture are shown in the figure 2.1.

k-star

k-star statistics tries to capture nodes-based network dependencies. It is a family of statistics, including the one-star statistic (the number of edges), the two-star, the three-star and so on, and takes the name *star* because of the star shape that the sub-configuration of nodes examined takes on as the order k of the k-star statistic increases. The k-star examines how other nodes connect to a specific node; For an undirected network, this structure might allow us to capture the *preferential attachment* process, that is, the tendency for nodes to receive new edges in proportion to their current degree. In particular, a positive coefficient related to a two-star statistic provides evidence for preferential attachment.
The k-star statistic takes the following form:

$$h_S(a) = \sum_{i \leqslant j \leqslant n} \binom{a_{i+}}{k}, \tag{2.6}$$

and in particular, what they do is count the number of two-stars, three-stars and so forth, depending on the order k being considered.
The k-star statistics can be extended for directed networks, getting the *in-k-star* and *out-k-star* statistics. They reflect the number of incoming and outgoing edges to a specific node, respectively, and they are related, in the

social network field, to the concept of *popularity* and *sociality* respectively.

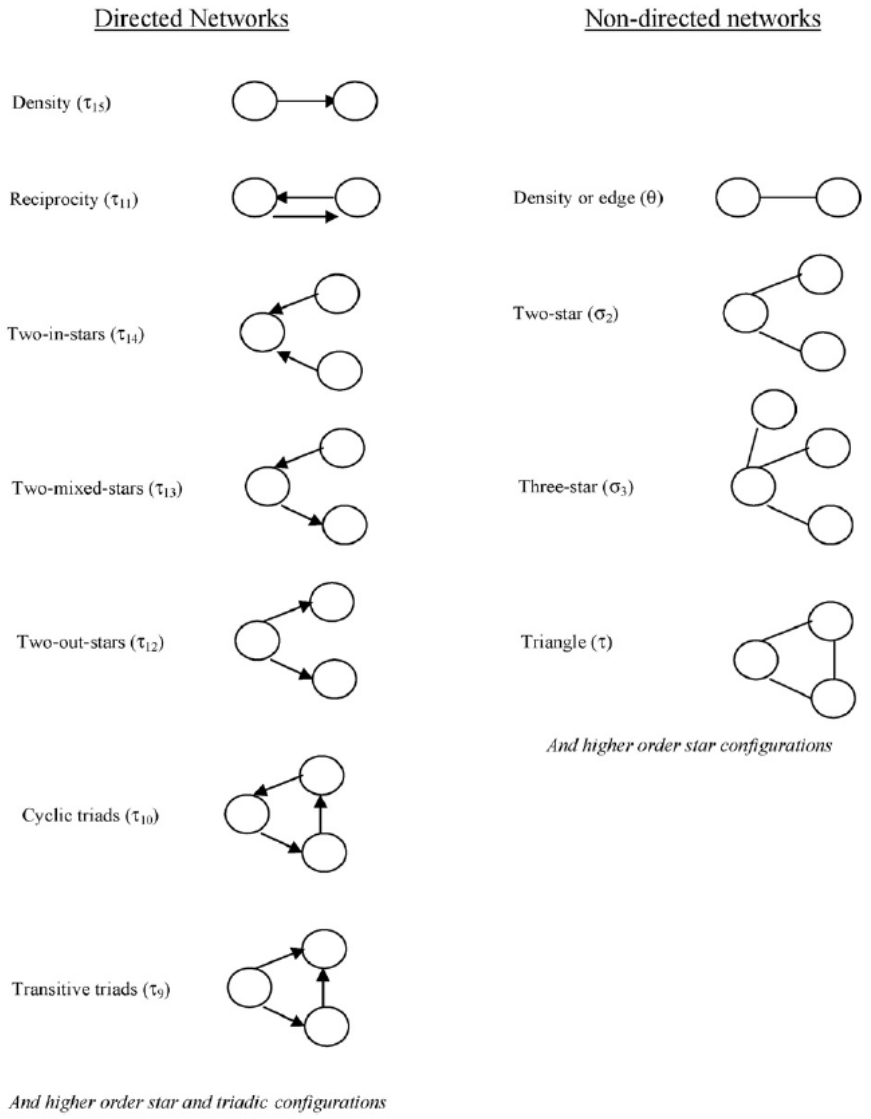

Figure 2.1: Some configuration for Markov graph random models, from Robins et al. (2007)

Reciprocity and Asymmetry

Reciprocity and *Asymmetry* are endogenous effects based on a single specific dyad or pair of nodes. As seen above, reciprocity is the most prominent dyad-effect. The mere use of reciprocity in the model, together with density, gives rise to the models described in the previous paragraph.
Another interesting dyad-effect is the *Antireciprocity*, or *Asymmetry*. Asymmetric dyads occurs when there is a directed edge from node i to node j but there is not an edge from node j to node i. The asymmetry statistic can be obtained as a modification of reciprocity statistic:

$$h_A(a) = \sum_{i \leq j} a_{ij} (1 - a_{ij}), \tag{2.7}$$

in which the product $a_{ij}(1 - a_{ij}) = 1$ if and only if an edge exists between node i and node j but not between node j and node i.

Triadic and Higher-Order Endogenous Effect

The higher order endogenous effect are related on subgraph with three or more nodes. The most common and simple higher-order effect is the *Triangle*. A network has a triangle when node i has an edge to nodes j and k, and those nodes are connected to one another. For undirected networks, this triadic configuration is also called *Closed Triad*, and in Social Science it captures the concept of "the friend of a friend is a friend". A Closed Triad statistic takes the following form:

$$h_T(a) = \sum_{i \leq j \leq h} a_{ij} a_{ik} a_{jk}, \tag{2.8}$$

in which the ordering of connections does matter.
For directed networks, there are different closed-triad configuration. The most prominent configurations are the *Cyclic Triple* and the *Transitive Triple*. The first one occurs in a network when node i has an edge to node j, which has an edge to node k, which has an edge to node i and it can be interpreted as the return of an edge that passes from a third vertex, The second one occurs when

node i has an edge to nodes j and k while node j also has an edge to node k and can be interpreted as node i co-supporting node j and k, or as node k being co-supported by nodes i and j (is related to the concept of *transitivity*). For directed transitivity, the corresponding statistic takes the form of (2.8), in which the order of the terms in the summation is relevant, and it could be shifted to be cyclical by changing $a_{ij}a_{ik}a_{jk}$ to $a_{ij}a_{jk}a_{ki}$.

Dependence Structures with Node-Level Variables

It is possible to introduce node-level variables (or node attributes) into the Markov graph model, and in a more general *ERGM*. These variables could be binary, polytomous and continuous and are assumed to be an exogenous effect of network connections. The introduction of these variables, together with the statistics seen above, lead to the specification of the model in the form seen in the equation (2.1).

2.4 Estimation

For *ERGMs*, *Maximul Likelihood Estimation* (MLE) is very interesting. Since *ERGM* has a canonical exponential family form, MLE has ideal properties: the concavity of the likelihood function implies that the parameter estimates are the best possible estimate; the statistics used inside the model are sufficient statistics, which implies that we do not need any other information to estimate the parameters; the MLE, if it exist, is asymptotically normally distributed. However, the MLE is computationally expensive and it is infeasible for all but smallest networks due to the computation of the normalization constant. Furthermore, the flexibility and complexity in the functional form of the *ERGM* can give rise to *degeneracy problem* (which will be dealt with later) which can also manifest as an estimation problem. Here, we will discuss the two prominent estimation methods for *ERGM*, which allow us to approximate MLE.

2.4.1 Monte Carlo Markov Chain MLE

This method, denoted as *MCMC-MLE*, was developed by Geyer and Thompson (1992); Snijders (2002). The idea behind is that a sample of networks is extracted from the distribution specified by *ERGM* defined by the parameter values obtained in the previous iteration of estimation. Given initial values for parameters, in each iteration the current estimate are used to draw a sample of networks and then updating estimates to maximize the approximated likelihood function. Usually, the initial values of parameters are obtained using *Maximum Pseudo-Likelihood Estimation* (which will be dealt with later) and this does not require the direct computation of the normalization constant. The number of networks to be sampled in each iteration is typically set to at least several thousand. To create a Markov chain, all the entries of the adjacency matrix are considered as random variables that can take value 0 or 1. By switching the value from 0 to 1 or vice versa (that is, by adding or removing an edge), using *Metropolis-Hastings* algorithm (or *Gibbs Sampling* in some software), a sequence of graph will be created in which each graph only depends upon the previous graph.
In other words, the likelihood of the sampled graph in step s is calculated; if it is greater than the likelihood associated with the graph in step $s-1$, the sampled graph will become the new graph in the next iteration. The algorithm runs until it reaches convergence, that is, until the values of the parameters or likelihood differ greatly between one iteration and another of the algorithm. In the initial phase it may be necessary to use *burn-in* to eliminate part of the chain if the initial values are poor. Usually, a maximum number of iterations is specified, which very often is set at 20. In practice, it is good practice to evaluate the convergence of estimates, and possibly increase the number of iterations, and / or the *MCMC* sample size if it is too small, as this sample may not accurately capture the network distribution.

Algorithm 1 MCMC-MLE with Metropolis Hasting

1: Let $\boldsymbol{A}_0$ be the adjacency matrix of the observed network
2: $s = 0$
3: Let $\boldsymbol{\theta}^0$ the vector of initial parameter values
4: **repeat**
5: $s = s + 1$
6: Drawn $\tilde{\boldsymbol{A}}$ from ERGM($\boldsymbol{A}, \boldsymbol{\theta}^{s-1}$) with the same number of nodes and edges, using Metropolis-Hasting
7: Evaluate the likelihood $p^s_{\boldsymbol{\theta}}$
8: Using Hasting Ratio to accept the sampled network: $min\left\{1, \frac{p^s_{\boldsymbol{\theta}}}{p^{s-1}_{\boldsymbol{\theta}}}\right\}$
9: **until** a convergence criterion is met

2.4.2 Maximum Pseudo-Likelihood Estimation

Maximum Pseudo-Likelihood Estimation (MPLE) for *ERGMs* was introduced by Strauss and Ikeda (1990). Let A_{ij} denote the $ij-th$ entry of the adjacency matrix, the join likelihood is replaced by the conditional form, or *pseduolikelihood*, defined as the product of the conditional probability of each entry of the adjacency matrix given the rest of the network:

$$\pi_{ij}(\theta) = P(A_{ij} = 1 | A_{-ij}, \theta) = \frac{1}{[1 + \exp(-\theta' \delta_{ij}(\boldsymbol{x}(\boldsymbol{A})))]}, \tag{2.9}$$

where A_{-ij} denote the network except the ij^{th}-element and $\delta_{ij}(\boldsymbol{x}(\boldsymbol{A}))$ denote the vector of *change statistics*: the vector of changing values of the networks statistics when A_{ij} change from 0 to 1, or vice versa, given the rest of the network. By using a hill-climbing algorithm, the pseudolikelihood will be maximize:

$$\arg\max_{\theta} \sum_{\langle ij \rangle} \ln \left[(\pi_{ij}(\theta))^{A_{ij}} (1 - \pi_{ij}(\theta))^{1 - A_{ij}} \right], \tag{2.10}$$

where $\langle ij \rangle$ denotes all pairs of nodes.
This method does not require the construction of a Markov chain, but it has some inferential disadvantages. First of all, this method is like a logistic regression problem without independent assumption between observations; this

implies that the estimate are biased and the standard errors are approximate. The Wald statistic to test the statistical significance of the parameters should be used with caution, and furthermore, we can not assume that the corresponding deviance is asymptotically Chi-squared distributed. However, the MPLE has been shown to be consistent, approximating the MLE in distribution as the network size increases.

2.5 Degeneracy problem: Curved ERGM

MCMC-MLE methods can lead to the problem of *degeneracy*. The degeneracy is configured as a case in which only sparse or high-density networks have a high probability mass. These methods aim to simulate the probability distribution of the network, adding edges between one iteration and the other. By specifying an *ERGM* using higher-order subgraph counts, a new edge that is added during the estimation process contributes to potentially greatly increase the number of higher-order configurations. For example, an additional edge can turn a 2-star configuration into a triangle, or it can become a 3-star and so on. Considering that this could occur for each edge that is added step by step, they have a multiplicative effect on the number of subgraph configurations, increasing the parameter relating to the corresponding statistic of the subgraph configuration concerned. This leads to having networks with high probability mass.
Several authors have developed statistics that allow us not to fall into the problem of degeneracy. Of all, the first and most important contribution was made by Snijders et al. (2006), who developed weighted versions of some statistics responsible for the degeneracy of the model. Here, statistics for undirected and directed networks proposed by Snijders et al. (2006) will be discussed.

2.5.1 Geometrically Weighted Degrees and Related Functions

Since in a Markov graph model we can use any arbitrary linear function of the k-star counts, S_k), for $k = 1, ..., n-1$, which are independent polynomials of the nodes degree; then, the k-star counts can be seen as a linear combination of polynomials of degree $n-1$. Therefore, any function of degree distribution can be used in the linear predictor, remaining in the Markov graph models family. When a parameter related to a k-star count has a positive sign, then there is the risk of running into the problem of degeneracy, as we would have a high probability for hyper-connected graphs, i.e. with very high nodes degrees. Based on this, Snijders et al. (2006) proposed using a decreasing weights for high degrees, getting the following statistic called *geometrically weighted degrees*:

$$u_{\alpha}^{(d)}(y) = \sum_{k=0}^{n-1} e^{-\alpha k} d_k(y) = \sum_{i=1}^{n} e^{-\alpha y_{i+}}, \tag{2.11}$$

where y_{i+} are node degrees, $d_k(y)$ is the number of nodes with degree k and $\alpha \geq 0$ is called *degree weighting parameter*. In this way, α makes the effect of nodes with high degree attenuate when it takes on high values, while it gives greater weight to graphs with high degrees when it tends to zero The resulting model is a standard ERGM with the *geometrically decreasing degree distribution assumption*, that is, an ERGM with the use of degrees as statistics, for $k = 1, ...n-1$, with the following constraints on the parameters:

$$\theta_k = e^{-\alpha k}. \tag{2.12}$$

Since the degree distribution is a function of k-star counts, an equivalent model can be obtained using the k-star statistics, defining the following new statistic, called *geometric alternating k-star*:

$$u_{\lambda}^{(s)}(y) = S_2 - \frac{S_3}{\lambda} + \frac{S_4}{\lambda^2} - ... + (-1)^{n-2} \frac{S_{n-1}}{\lambda^{n-3}} = \sum_{k=2}^{n-1} (-1)^k \frac{S_k}{\lambda^{k-2}}, \tag{2.13}$$

in which the weights have alternating signs. Then, when we have graphs with increasingly high degrees, the contribution from extra k-stars is balanced by the contribution from extra $(k+1)$-star. In formula (2.13) $\lambda = \frac{e^{\alpha}}{e^{\alpha}-1} \geq 1$. The model obtained using this statistic, is like standard ERGM with k-star statistic, with the following parameter constraints in the corresponding parameter:

$$\theta_k = \frac{-\theta_{k-1}}{\lambda}. \tag{2.14}$$

The use of one of these statistics transform the *ERGM* in a class of model called *curved ERGM*.
For directed network, *geometrically weighted out-degrees* and *geometrically weighted in-degrees* can be obtained as the following:

$$u_{\alpha}^{(od)}(y) = \sum_{k=0}^{n-1} e^{-\alpha k} d_k^{(out)}(y) = \sum_{i=1}^{n} e^{-\alpha y_{i+}}, \tag{2.15}$$

$$u_{\alpha}^{(id)}(y) = \sum_{k=0}^{n-1} e^{-\alpha k} d_k^{(in)}(y) = \sum_{i=1}^{n} e^{-\alpha y_{i+}}. \tag{2.16}$$

from which the corresponding alternating in-k-star and alternating out-k-star can be obtained.

2.5.2 Transitivity and Alternating k-Triangles

As discussed above, transitivity is captured by the model by entering the triangle count as a statistic. In his work, Snijders et al. (2006) emphasises that the problem of degeneration in Markov models can arise when the triangle count is entered in the model but the network has incomplete substructures, incomplete "cliques", much larger which in turn contain many triangles. He deduces that the Markov dependence is too simple and modelling transitivity with just the triangle counts is too simplistic. Therefore, Snijders et al. (2006) propose a new statistic based on a more general independence concept, called *partial conditional independence* (CD).

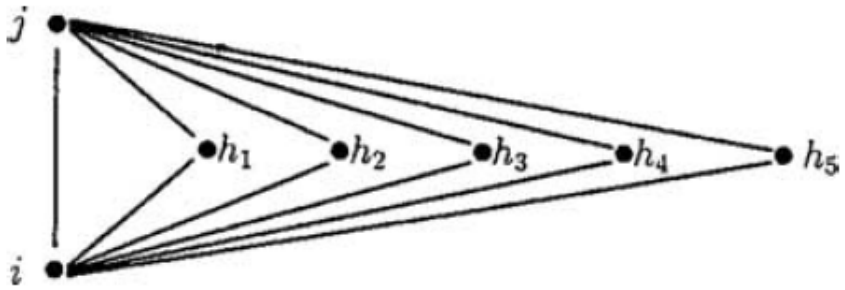

Figure 2.2: 5-triangle, from Snijders et al. (2006).

The partial conditional independence proposed by Snijders et al. (2006) says that: "two edge indicator A_{iv} and A_{uj} are conditionally dependent, give the rest of the graph, only if one of the two following conditions is satisfied:

1. They share a vertex, that is $\{i, v\} \cap \{u, j\} \neq \emptyset$ (the usual Markov independence assumptio);

2. $a_{iu} = a_{vj} = 1$, that is if the edges existed they would be part of a four-cycle."

Building on this new dependency assumption, Snijders et al. (2006) provides a new statistic capable of capturing transitivity, based on more complex structures than triangles, called *k-triangles*. For undirected networks, a k-triangle with base defined by the edge between node i and j is defined by the presence of at least k other nodes adjacent to both i and j. An example of this structure is represented in Figure 2.2; it can be seen as a combination of k individual triangle that share the same edge base. As the Figure 2.2 shows, a k-triangle including all the k-triangles with less k: in the example, a 5-triangle necessarily comprises four 4-triangles, four times the number of 4-triangles and so on.
The number of k-triangles in the network is given by:

$$T_k = \sum_{i<j} a_{ij} \binom{L_{2ij}}{k} \, (fork \geq 2) \,, \tag{2.17}$$

while the number of triangles, called also 1-triangles, is given by:

$$T_1 = \frac{1}{3}\sum_{i<j} a_{ij} L_{2ij}, \tag{2.18}$$

where $L_{2ij} = \sum_{h\neq i,j} a_{ij}a_{hj}$ is the number of two-path connecting nodes i and j; If there is an edge between nodes i and j, than, $k = L_{2ij}$ represent the maximum order of k-triangles with base (i, j).
The following k-triangle counts statistics proposed by Snijders et al. (2006) is a sufficient statistics used in *ERGM*, with weights that have alternating signs and are geometrically decreasing and which satisfies partial conditional independence assumption:

$$\begin{aligned} u_\lambda^{(t)}(a) = 3T_1 - \frac{T_2}{\lambda} + \frac{T_3}{\lambda^2} - ... + (-1)^{n-3}\frac{T_{n-2}}{\lambda^{n-3}} = \\ = \sum_{i<j} a_{ij} \sum_{k=1}^{n-2} \left(\frac{-1}{\lambda}\right)^{k-1} \binom{L_{2ij}}{k} = \\ = \lambda \sum_{i<j} a_{ij} \left\{1 - \left(1 - \frac{1}{\lambda}\right)^{L_{2ij}}\right\}, \end{aligned} \tag{2.19}$$

where $\lambda = \frac{e^\alpha}{(e^\alpha - 1)}$. The resulting model can be viewed as a standard *ERGM* with triangle counts statistic with the following parameter constraint in the corresponding parameters: $\tau_k = -\frac{t_{k-1}}{\lambda}$, for $(k \geq 3)$.
For directed networks, in which the corresponding configuration is showed in Figure 2.4, the alternating transitive k-triangles statistic is defined as

$$u_\lambda^{(t)}(y) = \lambda \sum_{i,j} y_{ij} \left\{1 - \left(1 - \frac{1}{\lambda}\right)^{L_{2ij}}\right\}. \tag{2.20}$$

2.5.3 Alternating Independent Two-Paths

In his work, Snijders et al. (2006) proposes a third statistic to capture transitivity, with the idea of modelling not the k alternating triangles, but the sides that would form the k alternating triangles, if there were an edge between

node i and j that creates the base of the triangles, introducing the concept of *k-independent two-paths*, each of which can be seen as a four-cycle, the combination of which produces the sides of the k-triangle configuration. These

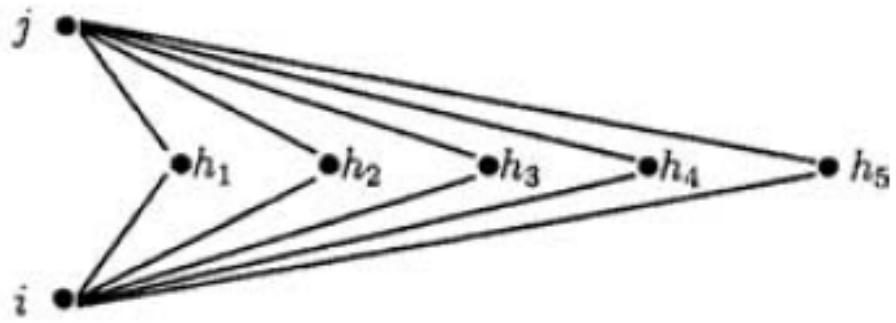

Figure 2.3: Example of five-independent two-paths, from Snijders et al. (2006).

configuration is showed in Figure 2.3. Their number is given by:

$$U_k = \sum_{i<j} \binom{L_{2ij}}{k} (\text{for} k \neq 2), \tag{2.21}$$

$$U_2 = \frac{1}{2} \sum_{i<j} \binom{L_{2ij}}{2}. \tag{2.22}$$

The corresponding statistic, called *alternating independent two-paths*, has two equivalent expression, given by:

$$u_\lambda^p(a) = U_1 - \frac{2}{\lambda} U_2 + \sum_{k=3}^{n-2} \left(\frac{-1}{\lambda}\right)^{k-1} U_k = \tag{2.23}$$

$$= \lambda \sum_{i<j} \left\{ 1 - \left(1 - \frac{1}{\lambda}\right)^{L_{2ij}} \right\}, \tag{2.24}$$

$$u_\lambda^p(a) = \lambda \binom{n}{2} - \sum_{i<j} e^{-\alpha L_{2ij}}, \tag{2.25}$$

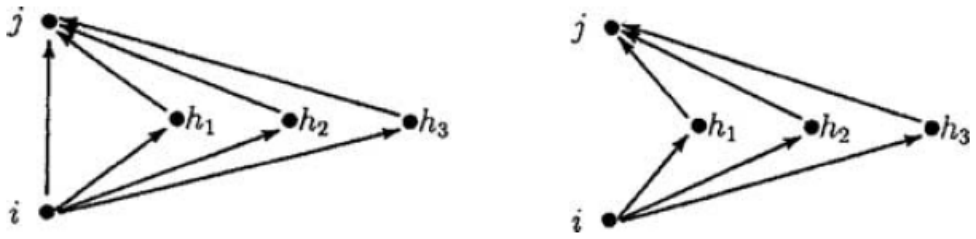

Figure 2.4: Transitive three-triangle (left) and three-independent two-paths (right), from Snijders et al. (2006).

where (2.24) has alternating weights for the counts of independent two-paths, while (2.25) has geometrically decreasing weights for the counts of pairs with given numbers of shared partners. In both expression $\lambda = \frac{e^{\alpha}}{(e^{\alpha}-1)}$. The resulting model can be viewed as a standard *ERGM* with two-paths counts statistic with the following parameter constraint in the corresponding parameters: $v_k = -\frac{v_{k-1}}{\lambda}$.

For directed networks, in which the corresponding configuration is showed in Figure 2.4, the alternating independent two-paths statistic is defined as

$$u_{\lambda}^{p}(a) = \lambda \sum_{i,j} \left\{ 1 - \left(1 - \frac{1}{\lambda} \right)^{L_{2ij}} \right\}. \tag{2.26}$$

2.6 Goodness of Fit Diagnostics

The evaluation of goodness of fit in the *ERGMs* is quite different from the checks that are made to evaluate goodness of fit in classical statistical models, however it is quite simple. While in second case, the evaluation of the goodness of fit is based on *residuals*, whose simplest form is given by the difference between observed and predicted values, in *ERGMs* the goodness of fit is evaluated on the basis of how much the observed network differs, and it is an outlier, with respect to the distribution of networks simulated by the estimated *ERGM*. Assuming that the model has captured the generating pro-

cess of the network, by simulating a certain number of networks it is verified whether the topological characteristics of the observed network are similar to those simulated; characteristics that are captured by various summary statistics: in the case of direct networks, in-degree and out-degree, dyad / edgewise and geodesic distance are used. The boxplot is the graphic tool used to represent the results so that they can be interpreted.
In order to provide a practical example of how the evaluation of goodness of fit is carried out in an *ERGM*, an example is proposed using the *Grey's Anatomy* hookup network, discussed in Cranmer et al. (2020). It is an indirect network with 44 knots and 46 edges. The model specified, for the purpose of showing the example, is a model with only edges and degree equal to 1. As can be seen from the Figure 2.5, the evaluation of the goodness of fit is done on topological features such as the degree distribution, the proportion of edge-wise shared partners and the proportion of dyads forming the minimum geodesic distance, for different values. Each boxplot represents the distribution of a given value of the corresponding topological feature, obtained by simulating 100 networks from the estimated model, while the black curve shows the trend of the corresponding topological features in the observed network. One can see how well the model captures the process of connection formation between nodes within the network, as the topological characteristics of the observed network are in line with the medians of each distribution obtained from the simulations.

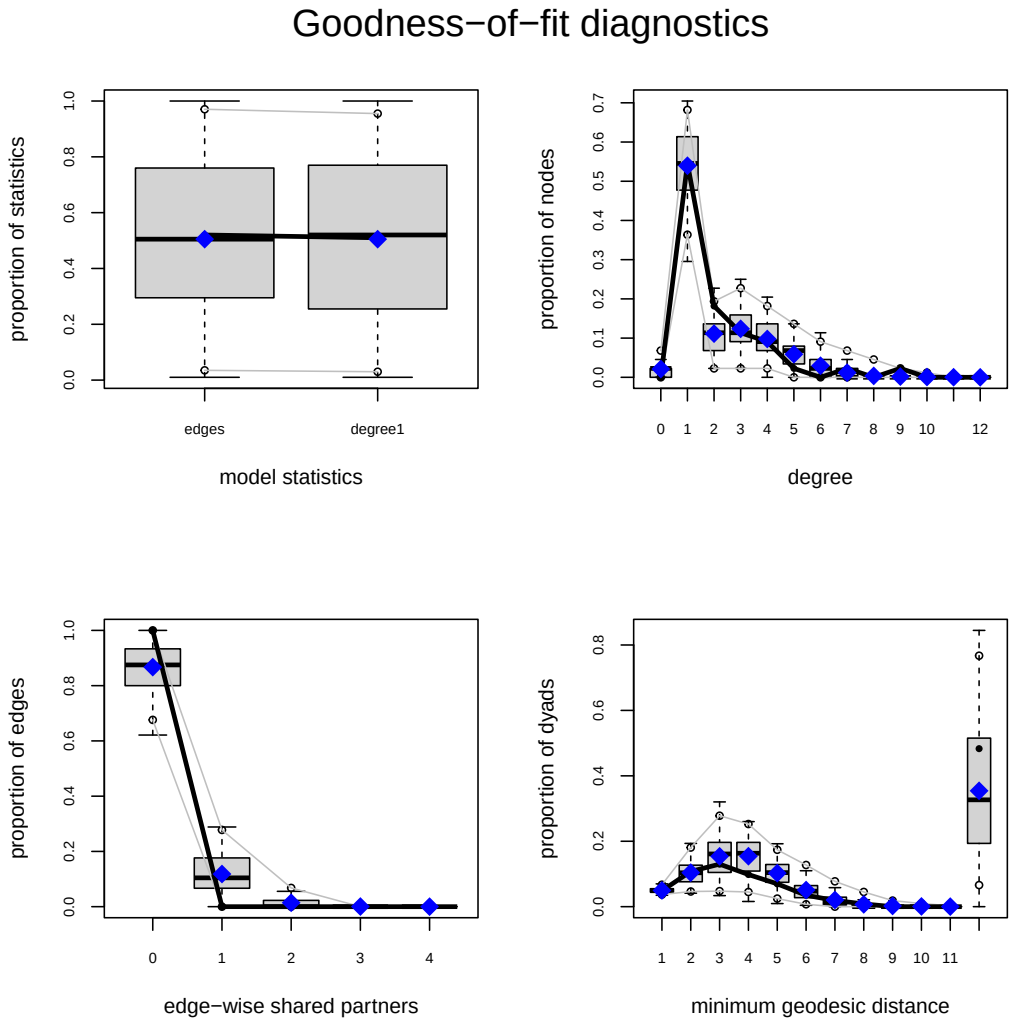

Figure 2.5: Example of goodness of fit in *ERGM*.

Chapter 3

Analysis and Applications

3.1 Data Description and Goal of the Analysis

The data used in this work are relate to a population of 31346 neurons obtained from the digital reconstruction of the microcircuitry of the somatosensory cortex of young rats, using a technique developed by Markram et al. (2015), which allowed the connectivity matrix (or adjacency matrix) to be obtained. In particular, using laboratory data extracted from a portion of the brain, the technique, made it possible to reconstruct the connectivity matrix by modelling the behaviour of both neurons and synapses of 55 layer-specific morphological and 207 morphoelectrical neurons are used (Figure 3.1).
Additional informations have been considered, such as:

- neuron locations in 3d space (for single neuron);
- number of afferent connections per neuron;
- number of efferent connections per neuron;
- number of afferent synapses (multiple synapses per connection possible);
- number of efferent synapses.

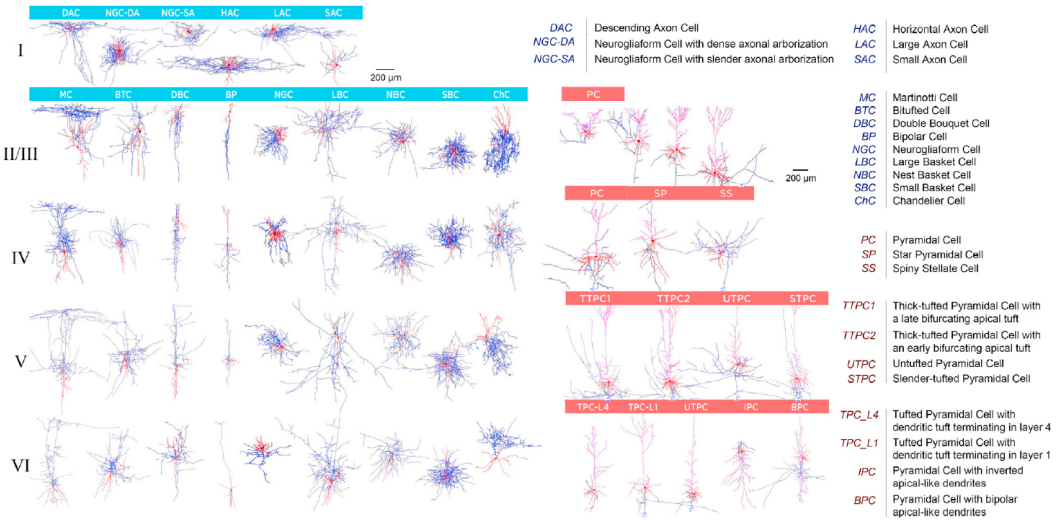

Figure 3.1: 55 layer-specific neocortical neuronal morphologies, from Markram et al. (2015).

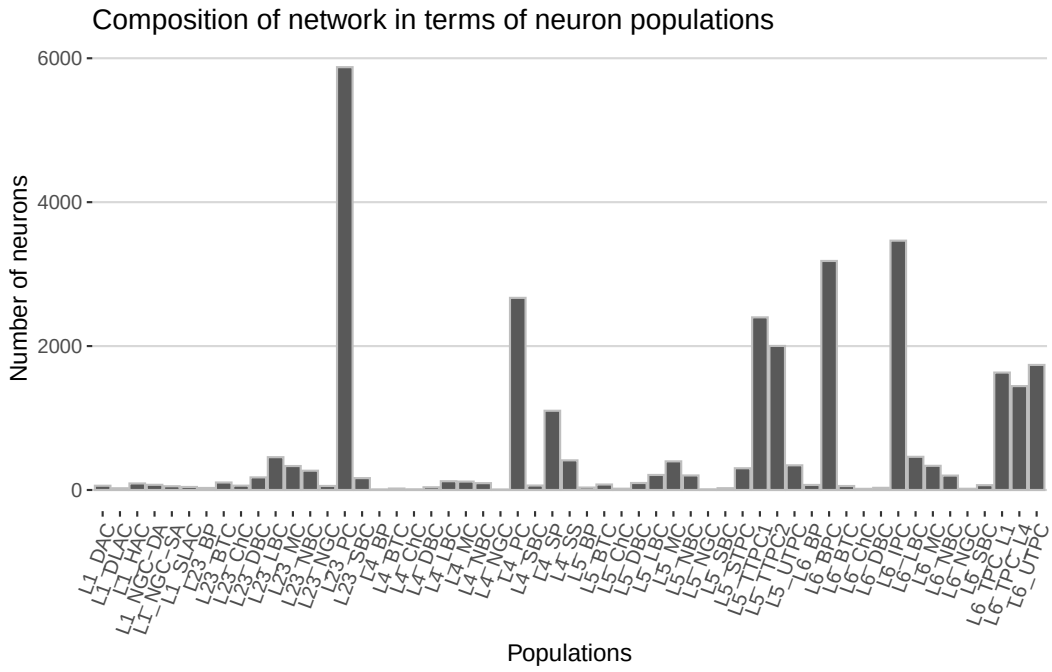

Figure 3.2: Composition of the network in terms of neuron populations.

Figure 3.2 shows that of the 55 populations of neurons considered, only 9 are predominant and only 5 of these exceed two hundred neurons: L23-PC with about 6000 neurons, i.e. about 19% of the total number of neurons, L6-IPC with about 3500 neurons, i.e. about 11% of the total number of neurons, L6-BPC with about 3200 neurons, i.e. about 10% of the total number of neurons, L4-PC with about 2700 neurons, i.e. 8.5% of the total neurons, and L5-TTPC1 with about 2400 neurons, i.e. about 7.6% of the total neurons.
The goal of the analysis is part of a larger goal of the *Human Brain Project* research project, that is to allow a deeper knowledge in the brain functionalities on the base of neuroscience, computing, and brain-related medicine. One of the project's goals could be achieved by simulating the neuronal connectivity in all its details through mechanisms that eliminate synaptic connections and neurons based on how likely it is to observe certain connections and neurons, in order to make the connectivity itself computationally simulable. With this information in mind, this analysis aims to provide a model capable of simulating brain networks with topological characteristics able to reproduce those observed experimentally.

3.2 Explorative Analysis

In this section we provide an explorative analysis of the network to understand its main topological characteristics. In particular, we will focus on degree distributions, reciprocity measure and other indices, useful for the model specification.

3.2.1 Degree Distribution

Degree Distribution derives from the concept of *Degree* of a node. In an undirected network, the degree of the node i, denoted by k_i, is the number of edges the node has to other nodes. For directed network, a node will have *in-degree*, k_i^{in}, and *out-degree*, k_j^{out}, which are the number of incoming and outgoing edges, respectively. In the context of microscale brain networks, the

in-degree of a node i is the number of synaptic connections that the i-th neuron receives from other neurons, i.e. the number of afferent synapses, whereas the out-degree is the number of synaptic connections that the i-th neuron has towards other neurons, i.e. the number of efferent synapses. Degree, in-degree and out-degree can be written in terms of adjacency matrix. Formally, the degree of node i in an undirected network is given by:

$$k_i = \sum_{j=1}^{n} A_{ij}, \tag{3.1}$$

while for directed network in-degree and out-degree are given by:

$$k_i^{in} = \sum_{j=1}^{n} A_{ij}, \qquad k_j^{out} = \sum_{i=1}^{n} A_{ij}. \tag{3.2}$$

The *degree distribution* p_k is defined as the fraction of nodes in the network with degree k:

$$p_k = \frac{n_k}{n}, \tag{3.3}$$

where n_k denotes the number of nodes with degree k and represents the probability that a node in the network has a degree equal to k. In a similar way, it is possible to derive the in- and out-degree distribution in directed network. The degree distribution can be see as *log-log scale*, and important information can be deduced from it. If the logarithm of p_k is a linear function of the logarithm of k:

$$log\,(p_k) = -\alpha log\,(k) + c, \tag{3.4}$$

then we shall say that the degree distribution follow a *power law* distribution. Networks with power law degree distribution are said to be *scale-free networks* and show interesting features.

Figure 3.3 shows the overall in- and out-degree distribution for the analysed network. A 50% of the nodes have at most a in-degree equal to 231 and a out-degree equal to 218, while the average in- and out-degree is equal to 253.9. The network has many hubs but, in any case, it does not have the characteristics

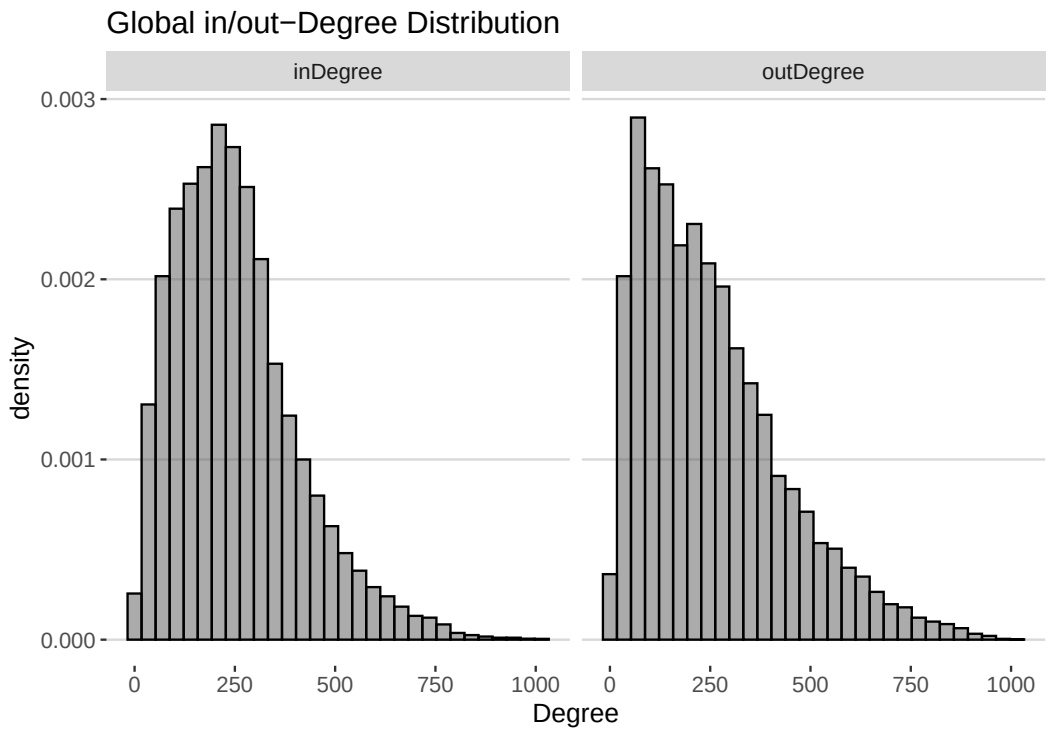

Figure 3.3: Global In/Out-Degree Distribution of the observed network.

of a power law, and it is well known in the literature, as was also pointed out in the most recent work by Giacopelli et al. (2020). This can also be seen graphically in Figure 3.5 and Figure 3.4, which shows how, on a logarithmic scale, the in- and out-degree distribution does not follow a straight line.
It is important to point out that a knowledge of the degree distribution gives important information about the network but, in most cases, it does not give us information about the complete structure, therefore additional measures are needed to fully describe it.

3.2.2 Density

Density is a measure that provides information on the composition of the network in terms of the number of edges compared to a network with the maximum possible number of edges, the latter being easily calculated through

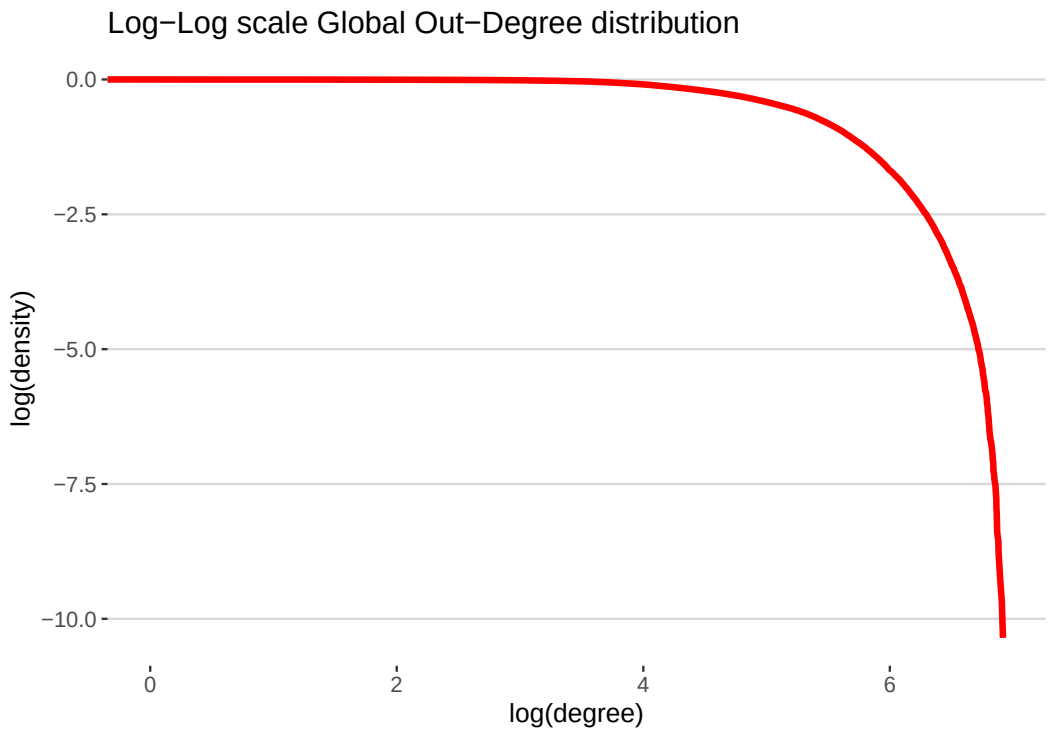

Figure 3.4: Global Out-Degree Distribution on log-log scale.

the binomial coefficient in the case of undirected networks. So, density ρ can be computed as:

$$\rho = \frac{E - E_{min}}{E_{max} - E_{min}} \tag{3.5}$$

where E is the number of edges in the network, $E_{min} = n - 1$ and E_{max} are the minimum and maximum number of edges in a connected network with n nodes, respectively. In the case of undirected networks, $E_{max} = \binom{n}{2}$; In directed graphs, the maximum number of edges depends on whether there are self-connecting nodes or not: in the first case, the maximum number of edges is given by $E_{max} = N(N-1)$, while in the second case by $E_{max} = n^2$.
The density ρ lies in the interval $[0, 1]$. For large n, the more the density tends to zero, the more the network is said to be *sparse*; Conversely, the more the density tends to one, the more the network is said to be *dense*. In the analyzed network $\rho = 0.0081$, which indicates that the network is very sparse, and this

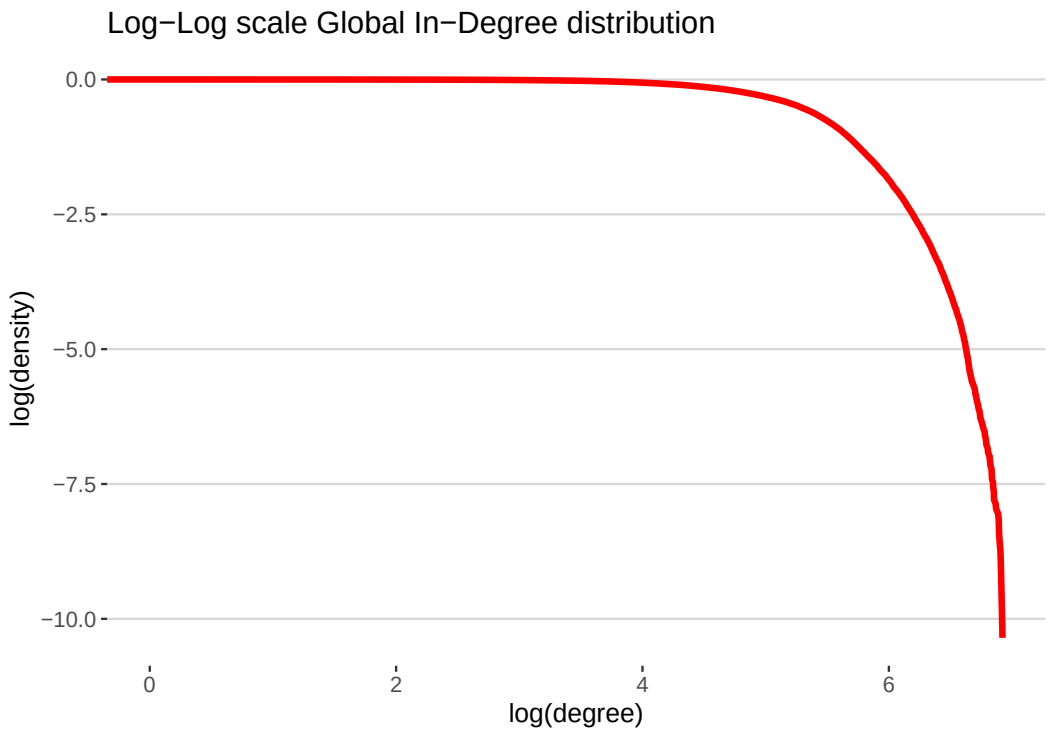

Figure 3.5: Global In-Degree Distribution on log-log scale.

will be taking into account for the choice of the model.

3.2.3 Shortest Paths and Diameter

A *shortest path* in a network, also called *geodesic path*, is a concept that refers to a pair of nodes and can be defined as the shortest path between all paths between two nodes, i.e. the one with the fewest edges, if there is a path between the pair of nodes under consideration. In particular, we focused on the average path length in a graph, calculating the shortest paths between all pairs of vertices and averaging them, which is 2.47. Another measure taken into consideration is the *diameter*, which is given by the longest shortest path among all possible shortest paths existing between two nodes. In the analysed network, diameter is equal to 7.

3.2.4 Reciprocity

The *Reciprocity* is a concept that comes into play only in direct networks and concerns pairs of nodes. Specifically, it is a measure of the fraction of node pairs that have edges to each other (i.e. a situation where node i has an edge to node j, and vice versa), hence the probability that such a structure is observed in the network. Reciprocity r is defined as the fraction of edges that are reciprocated:

$$r = \frac{1}{m} \sum_{ij} A_{ij} A_{ji} = \frac{1}{m} Tr\mathbf{A}^2, \tag{3.6}$$

where $A_{ij}A_{ji}$ is the product of two entries of adjacency matrix, which is equal to 1 if there is an edge from i to j, and vice versa; m is the total number of directed edges in the network.
In the analysed network, the reciprocity is equal to 0.024, which means that it is unlikely to observe a reciprocal situation in pairs of nodes.

3.2.5 Modularity and Community Detection

Since the network is very large, to be able to apply the model, the estimation and goodness of fit evaluation was made on a community of the network, obtained by community detection. Subsequently, the goodness of fit on the other communities surveyed was also evaluated.
There are several algorithms which allows us to do community detection, but most of them are for undirected network. Moreover, most of them doesn't work with large network and give bad results. For the analyzed network the *Leading Eigenvector method* proposed by Newman (2006) was used in order to get community structure. This method is grounded on *Modularity maximization*. It is a method used for undirected networks, therefore in order to obtain a subdivision of the nodes into communities, the network was treated as undirected.

For undirected networks, *Modularity* is denoted by Q and it is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{g_i g_j}, \tag{3.7}$$

where m is the total number of edges, A_{ij} is the generic entries of the adjacency matrix, k_i and k_j are the degrees of node i and j respectively, g_i and g_j are the community of node i and j respectively, and they are integer taking values from 1 to G, with G the total number of communities. Finally, $\delta_{g_i g_j}$ is the Kronecker delta. Modularity quantifies the level of non-randomness in the placement of edges in the network by comparing the total number of edges between nodes in the same community, given by $\frac{1}{2}\sum_{ij} A_{ij}\delta_{g_i g_j}$, with the expected number of edges between pairs of nodes within the same community, given by $\frac{k_i k_j}{2m}\delta_{g_i g_j}$. The constant $\frac{1}{2}$ compensates for the fact that each pair of nodes i, j is counted twice in the second term in brackets. Finally, the whole is divided by m to work with fractions. This quantity takes on positive values if there are more edges between nodes within the same community than we would expect by chance. All community detection algorithms based on Modularity aim precisely at maximising this quantity.
The heart of the *Leading Eigenvector method* is the *Modularity Matrix*. Denoted by $P_{ij} = \frac{k_i k_j}{2m}$ the probability that there is an edge between vertices i and j in a random network in which the degrees of all vertices are the same as in the input graph, and $\delta_{g_i g_j} = \frac{1}{2}(s_i s_j + 1)$. Modularity can be written in the following manner

$$Q = \frac{1}{4m} \sum_{ij} \left[A_{ij} - P_{ij} \right] (s_i s_j + 1) = \frac{1}{4m} \sum_{ij} \left[A_{ij} - P_{ij} \right] s_i s_j \tag{3.8}$$

or in matrix form

$$Q = \frac{1}{4m} \mathbf{s}^T \mathbf{B} \mathbf{s} \tag{3.9}$$

where $\mathbf{B}$ is called *Modularity Matrix* with generic entries $B_{ij} = A_{ij} - P_{ij}$. Without going into mathematical formalism, and focusing only on the simplest case, i.e. a situation in which one has a network with two communities, the method

works by calculating the eigenvalues and eigenvectors of the modularity matrix. Then the eigenvector corresponding to the largest positive eigenvalue is selected, and according to the sign of each element of the eigenvector (each element of the eigenvector corresponds to a node) the nodes are separated into two communities. It follows that if there is uniformity of signs in the eigenvector elements, the network has no community structure.
This method allowed us to find a structure of five communities with a corresponding Modularity value of 0.31. Table 3.1 not only shows the number of nodes and edges for each community but also the statistics used to evaluate certain topological characteristics of the overall network. As can be seen, all five communities, despite their different sizes, especially in terms of a number of edges, present a very similar level of sparsity among themselves and similar to that of the entire network. The same conclusion can be reached regarding reciprocity and the average shortest path.

Table 3.1: Characteristics of the five communities.

Community	**N.Nodes**	**N.Edges**	ρ	r	**Av. Shortest Path**
1°	8146	1167172	0.017	0.022	2.424
2°	7660	1114867	0.019	0.043	2.279
3°	7246	1024737	0.020	0.024	2.395
4°	4671	529649	0.024	0.023	2.402
5°	3623	399342	0.030	0.049	2.194

The model was fitted on the fifth community, the smallest in terms of the number of nodes and edges. In terms of community structure concerning populations of neurons, as Figure 3.6 shows, the fifth community is representative of only 29 out of 55 populations of neurons and, in particular, there are two predominant populations: L6-IPC with 969 neurons, i.e. about 27% of the neurons in the entire community, and L6-BPC with 855 neurons, i.e. about 22%. As Figure 3.8 shows, a similar structure is also observed for the second community, both in terms of the populations represented and the distribution of neurons within them. From Figure 3.7, it can be seen that the first community is the largest in terms of the number of neurons and edges, and, together

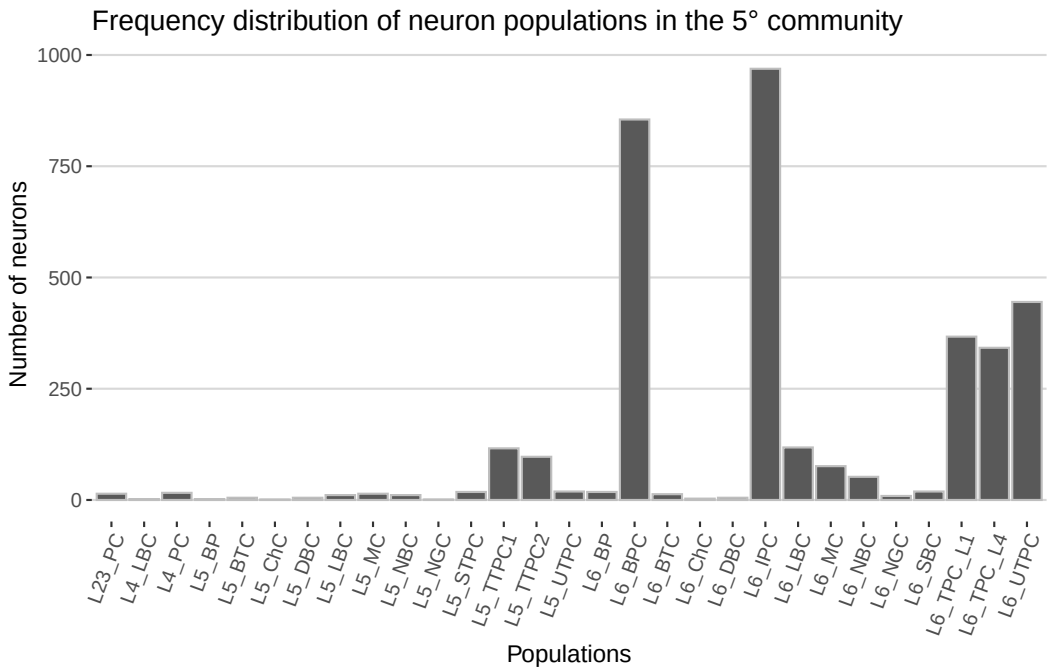

Figure 3.6: Structure of fifth community in terms of neuron populations.

with the third and fourth communities, is representative of almost all populations. In particular, the largest population is L23-PC, with 2351 neurons, i.e. about 29% of the total number of neurons contained in the first community. A similar structure is observed for the third and fourth communities, as shown in Figures 3.9 and Figure 3.10.

3.3 Model Results

In this section, we will show the main results of the model used to reach the goal of our analysis. However, it should be noted that, given the high computational complexity of the model estimation algorithm and the considerable size of the community 5 on which the model has been applied, it has not been possible to specify complex models in terms of graph configuration statistics for being able to better explain the mechanism for generating the network edges. The analysis was carried out on a PC with Windows OS, hav-

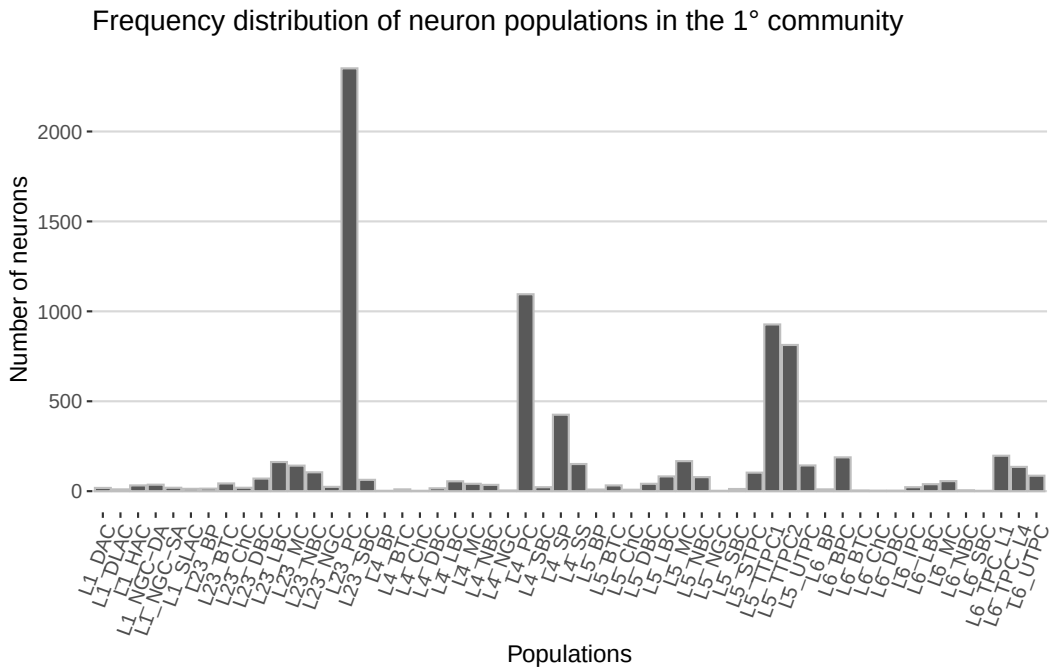

Figure 3.7: Structure of first community in terms of neuron populations.

ing the following characteristics: CPU INTEL(R) CORE(TM)i7-8550U with maximum frequency 4GHz; 8GB of RAM DDR4. For some models, where the required RAM exceeded that available of the used PC, the free Google Compute Engine service via Google Colab was used, which offers up to 13 GB of RAM.

The specification of the models that were estimated is given in Table 3.2. Model M0 is the model with only counting edges; this model is the baseline model from which we started, is the equivalent of the Erdős-Rényi model, and is the basis of the model developed by Giacopelli et al. (2020). Models M1, M2, and M3 only take exogenous covariates into account. In particular, only one exogenous covariate was taken into account, the distance matrix, and the square of it, of the neurons, whose individual distance was calculated as the Euclidean distance from the coordinates of the nodes. The reason why this covariate was chosen is related to the assumption that the further apart two neurons are, the less likely it is that an edge exists between them. This is be-

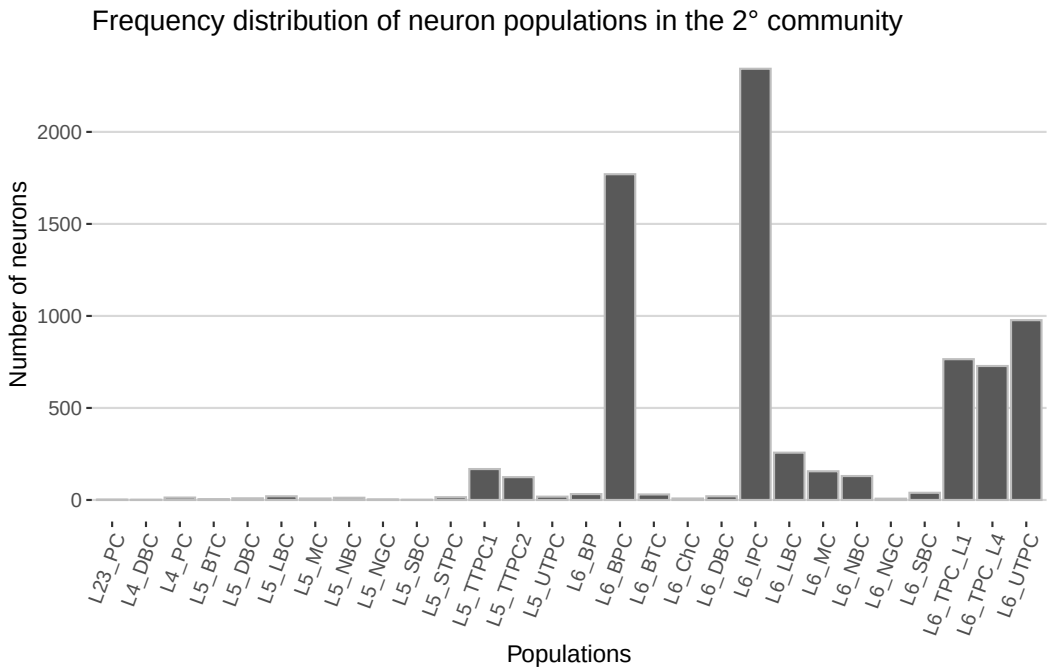

Figure 3.8: Structure of second community in terms of neuron populations.

cause it is already known in the literature that brain networks minimize wiring cost, and the Euclidean distance can be used as a proxy for wiring cost. The subsequent models only take exogenous statistics into account, except for the M5 model. In particular, the exogenous statistics that have been taken into account are inherent to specific graph configurations based on both pairs of nodes, such as reciprocity, and high-order configurations, such as in-k-star and out-k-star, cyclic triads, and transitive triads. In detail, the M4 model only considers reciprocity (also referred to as mutuality in Table 1), while the M5 model also considers distance. Model M6 takes into account more exogenous statistics such as reciprocity and in-/out-2-star, while model M7 only takes into account in-/out-k-star statistics of order 2 and up to order 4. Finally, the M8 model not only takes into account the statistics just mentioned but also triplets. The interpretation of the parameters for this type of network is very difficult, as these models were mainly designed to analyse social networks.
Model selection was made by three measures of goodness of fit, i.e. AIC, AICc

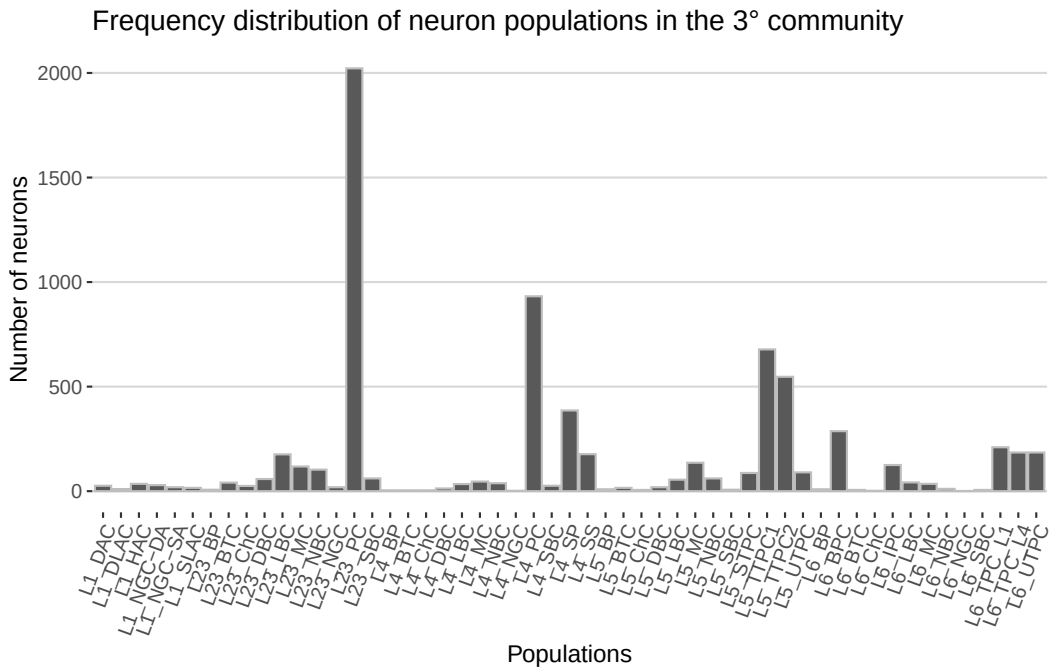

Figure 3.9: Structure of third community in terms of neuron populations.

and BIC, which also allow for the comparison of non-nested models. Table 3.3 shows the estimated models, their model comparison measures, the estimation times (in second) and on which machine the estimate was made; For all the models, MPLE methods was used to estimate them, since for large networks MPLE and MCMC-MLE gives similar estimates. The best model in which distance is included is the model with the Euclidean distance and the square of it, as it has the smallest AIC, AICc and BIC values. From a computational point of view, the estimation algorithm took less than 2 minutes to estimate the model, however it was not estimated locally as it required a lot of memory resources. From model results, shown in Table 3.4, all model estimates are highly significant at the 5% level. The *edges* parameter refers to the edge count and represents the "intercept" of the model, so it is not of useful interpretation. The negative sign is due to the fact that the network is very sparse. The distance parameter, although significant, on an exponential scale is equal to 0.99, very close to unity, so there is a negative effect on distance,

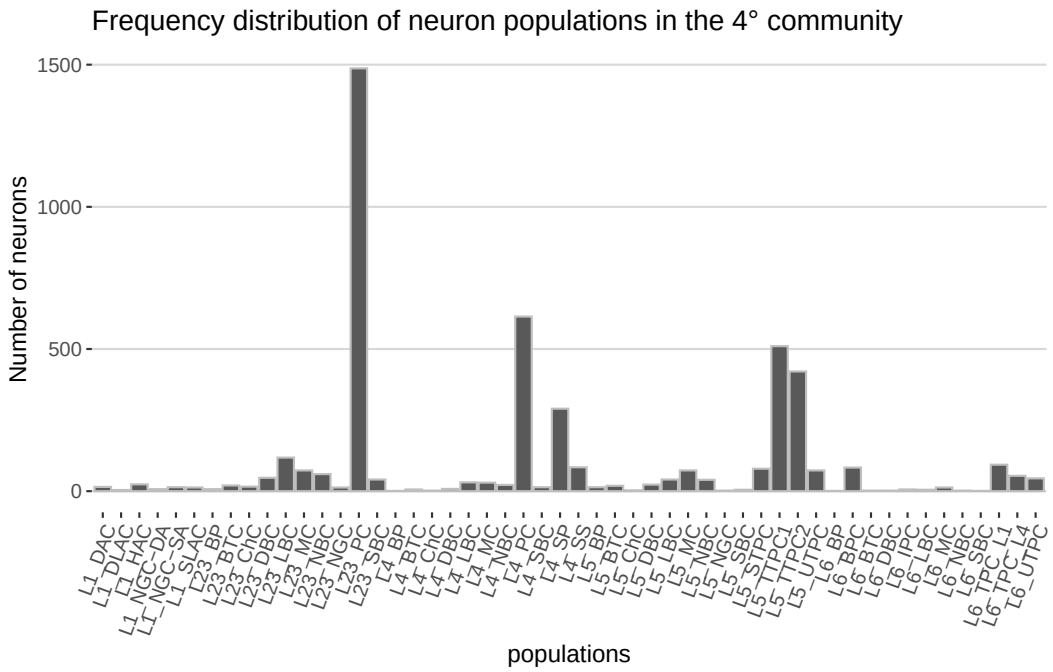

Figure 3.10: Structure of fourth community in terms of neuron populations.

as previously guessed, such that it is 1.01 times more likely not to observe an edge between two neurons given a unit increase in the distance between them. This result was to be expected as Euclidean distance can be used as a proxy for wiring cost, and it is known in the literature that brain networks are also characterized by minimising the wiring cost of synaptic connections. The distance squared, although significant, on an exponential scale is equal to 1; this means that on a quadratic scale there is no effect on the probability of edge formation. The best model among those estimated turns out to be the one in which reciprocity, in-2-star, out-2-star, transitive and cyclic triplets are specified, as it is the one with the lowest values of AIC, AICc and BIC. From a computational point of view, the algorithm took about 2 hours to estimate the model and required so much memory resources that it was not possible to estimate it locally. From model results, shown in Table 3.5, all model estimates are highly significant at the 5% level. The highest magnitude parameter is the reciprocity parameter, indicating that there is evidence that connections

Table 3.2: Specification of estimated models.

Model	**Specification**
M0	$edges$
M1	$edges + distance$
M2	$edges + distance^2$
M3	$edges + distance + distance^2$
M4	$edges + mutual$
M5	$edges + distance + mutual$
M6	$edges + mutual + istar(2) + ostar(2)$
M7	$edges + mutual + istar(2) + ostar(2) + istar(3) + ostar(3) + istar(4) + ostar(4)$
M8	$edges + mutual + istar(2) + ostar(2) + ctriple + ttriple$

Table 3.3: Estimated models with associated base information.

Model	**AIC**	**AICc**	**BIC**	**Time**	**Local**
M0	3575625	3575625	3575639	11.05	Yes
M1	3431173	3431173	3431202	74.313	Yes
M2	3481760	3481760	3481788	76.080	Yes
M3	3413172	3413172	3413215	98.354	No
M4	3571208	3571210	3571237	21.30	Yes
M5	3430539	3430539	3430582	93.590	No
M6	3356053	3356053	3356111	25.65	Yes
M7	3332040	3332040	3332156	79.01	Yes
M8	3196231	3196231	3196318	5596.706	No

between pairs of neurons tend to appear in substructures where they exchange information with each other, conditional on the other features in the model. From a neurophysiological point of view, as discussed above, this result should be in accordance with the well back.propagation phenomenon of the signal through the neuron itself, that makes the reactivation of the adjacent neuron plausible.

The in-2-star and out-2-star parameters can be interpreted as the popularity and sociability of neurons. However, although these parameters are statistically significant, on an exponential scale they are equal to 1, which indicates that these substructures do not have a large effect in explaining the process

Table 3.4: Model results with distance and distance squared.

Parameter	**Estimate**	**SE**	**exp(Estimate)**	**z-test**	**Pvalue**
edges	-2.061e+00	3.684e-03	0.1274	-559.6	<1e-04
distance	-6.859e-03	2.033e-05	0.9931	-337.4	<1e-04
$distance^2$	3.749e-06	1.835e-08	1	204.4	<1e-04

Table 3.5: Results of the best estimated model.

Parameter	**Estimate**	**SE**	**exp(Estimate)**	**z-test**	**Pvalue**
edges	-4.525e+00	5.528e-03	0.0108	-818.634	<1e-04
mutual	1.119e-01	8.053e-03	1.1184	13.895	<1e-04
*istar*2	2.065e-04	2.948e-05	1.0002	7.002	<1e-04
*ostar*2	1.262e-04	3.462e-05	1.0001	3.645	0.000268
ctriple	-6.613e-02	4.705e-04	0.9360	-140.552	<1e-04
ttriple	7.549e-02	1.936e-04	1.0785	389.987	<1e-04

of connection formation between the neurons in the network under investigation, conditional on the other features in the model. It is worth of mention, that similar results were obtained in model M7 having also higher-order k-star statistics. From a neurophysiological point of view, these situation should means that all the information in that portion of the network, have to necessarily pass through a single central neuron, to be processed and transferred to the neighbours ones. This would be a huge problem from a functional point of view, because if the cardinal neuron would stop to function (for example as a result of pathological conditions or little injuries) all the brain region would risk to be compromised. Thus, such statistics might be useful when trying to understand the mechanism of formation of synaptic connections in 'sick' networks. Additionally, we could also speculate on the sign of the parameters; following the reasoning above, one would expect the corresponding parameters of these statistics to assume a negative sign. In reality, the values are so close to zero that it is reasonable to think that the estimates obtained are biased. The parameter for cyclic triplets is of negative sign but does not have a strong magnitude, so it is implausible that cyclic connections between neurons are formed, conditional on the other features in the model. Indeed, such a connec-

tion, identifying a situation in which three neurons exchange information with each other, in a loop like structure, is not functional from a neurophysiological point of view.
Finally, the parameter relating to transitive triplets is of positive sign but small in magnitude; one can conclude by saying that, conditional on all the other features of the model, connections between neurons tend to appear forming closures. From a neurophysiological point of view, this substructure makes sense because it represents a situation in which a neuron receives signals from two neurons.
The assessment of the goodness of fit of the latter model was made concerning how the model simulated networks whose in-degree and out-degree distributions were as similar as possible to those of the observed network (in particular, for visualization problems we stopped at analysing distributions up to degree 200) simulating from it one hundred networks. A model with a good fit is such that the probability value for each individual degree is equal to the median of the degree distribution obtained from the simulated networks. From both Figure 3.11 and Figure 3.12, it can be seen that there are structural deviations of the probability value of the degree for the observed network, represented by the red curve, from what are the medians of the distributions of each degree, but nevertheless the trend is captured. This is an indication that the model does not have a good fit, i.e. it is poorly specified, so it would be necessary to introduce additional statistics that have a greater effect in explaining the mechanism of synaptic connection formation. In fact, it is also opportune to consider the fact that the signal propagation mechanism within the network does not occur on a limited number of nodes, so it makes sense to add higher-order statistics to the model that capture sub-configurations formed by many more nodes. I would like to emphasise the fact that such attempts have been made, but the specified models have not achieved convergence because they require more hardware resources than are available to perform the analysis.

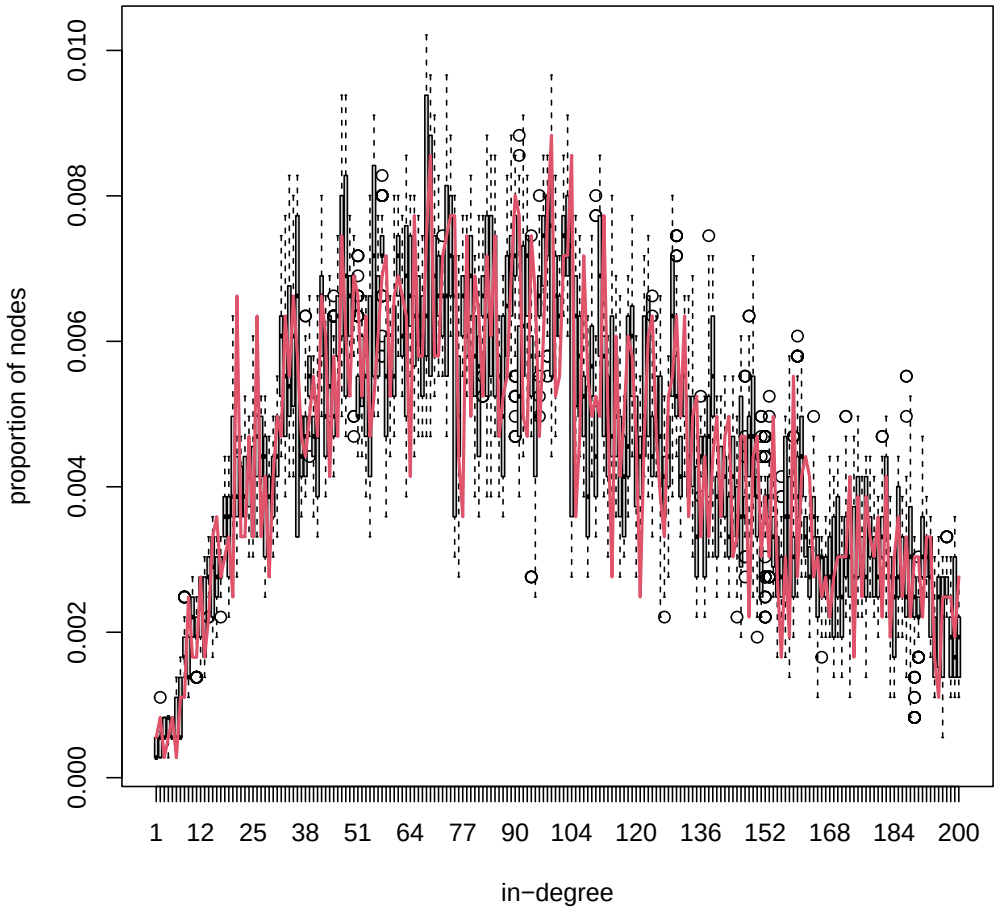

Figure 3.11: Goodness of fit on in-degree distribution.

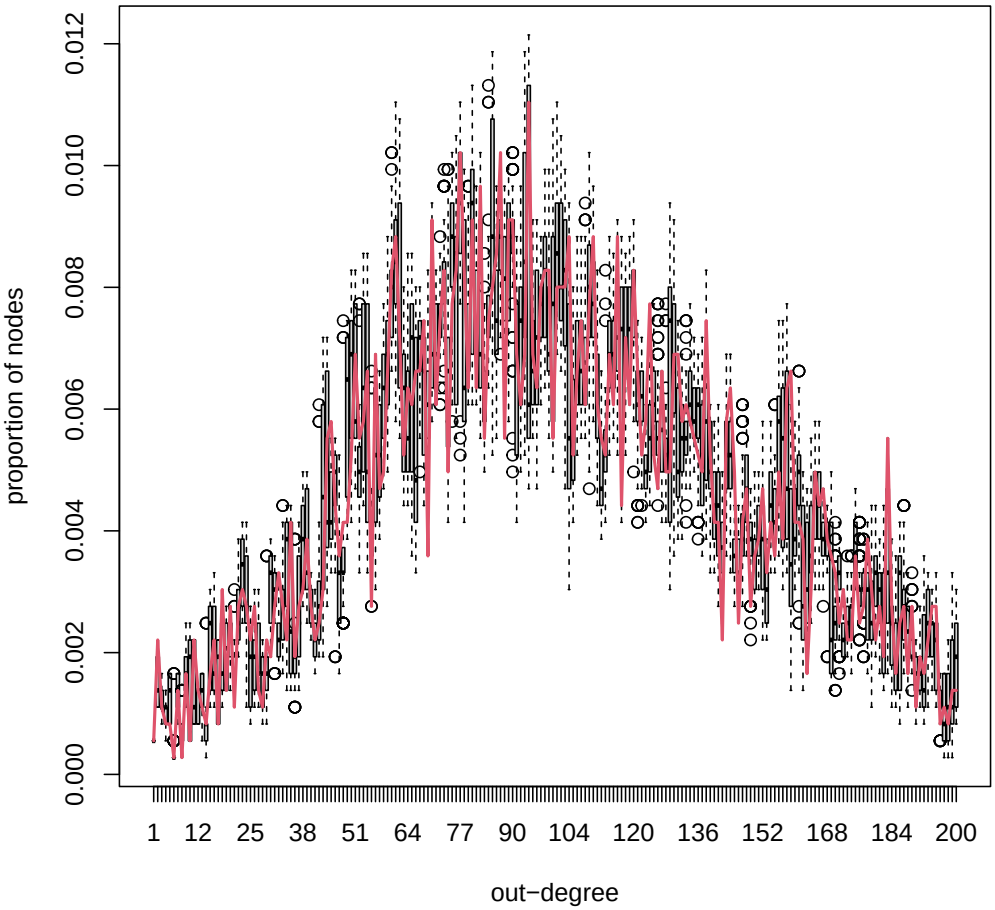

Figure 3.12: Goodness of fit on out-degree distribution.

Chapter 4

Conclusions

We have seen how the analysed network has characteristics that make it completely different from the networks that are analysed and studied by Network Analysis. Certainly, the first major difference that can be seen is the size of the network, which is considerably large. This had a great impact on the whole analysis. The size of the network made it virtually impossible to apply *ERGM*, so it was necessary to search for communities within it in order to make the model applicable. The size of the network also had an impact on the choice of algorithm to detect communities: due to the scarce presence in the literature of algorithms that allowed for the detection of communities in direct networks, and to the size of the network that invalidated the use of such algorithms, an approach based on the modularity constructed for undirected networks was used; this constitutes an initial limitation to the analysis that was carried out. Secondly, although *ERGMs* are widely used models capable of providing information on the topological characteristics underlying the formation of connections between nodes in a network, this is not the case for the network under investigation. Despite the fact that they have been applied to a larger network community, considerable computational problems have been encountered, which have had a great impact on the specification of the model. In fact, the statistics that were used were not sufficient to explain and grasp the mechanism of the creation of synaptic connections between neurons. By

the very nature of the network, it is reasonable to assume that the mechanism of information propagation between neurons, and thus the connectome within a brain network, involves a large number of neurons and it is reductive to model this through statistics looking at a maximum of three neurons, such as cyclic and transitive triplets. Furthermore, the estimation of this model for the community in question was also a challenge from a technological point of view: the memory resources required of the machine by these models are considerable, and it is also due to the amount of RAM available to the machines used that more complex models could not be specified.

This actually opens the door to new research possibilities on various aspects. First of all, the development of a community detection algorithm for high-dimensional direct networks. Secondly, brain network-specific exogenous statistics that take into account the characteristics of such networks; this is because such models were developed to analyse social networks and the interpretation of model parameters is not straightforward for brain networks. A third line of research could concern the fitting algorithms used for such models, in order to make them estimable in a reasonably short time.

Bibliography

A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, Oct. 1999. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.286.5439.509.

J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2): 192–236, 1974. ISSN 0035-9246.

S. J. Cranmer, B. A. Desmarais, and J. W. Morgan. Inferential Network Analysis. https://www.cambridge.org/highereducation/books/inferential-network-analysis/A7797D36A24647AA1F900CE7EF694C7E, Nov. 2020.

P. L. Erdos and A. Rényi. On the evolution of random graphs. *Transactions of the American Mathematical Society*, 286:257–257, 1984.

A. Fornito, E. T. Bullmore, and A. Zalesky. *Fundamentals of Brain Network Analysis*. Elsevier, 2016. ISBN 978-0-12-407908-3. doi: 10.1016/C2012-0-06036-X.

O. Frank and D. Strauss. Markov Graphs. *Journal of the American Statistical Association*, 81(395):832–842, Sept. 1986. ISSN 0162-1459. doi: 10.1080/01621459.1986.10478342.

C. J. Geyer and E. A. Thompson. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):657–699, 1992. ISSN 0035-9246.

G. Giacopelli, M. Migliore, and D. Tegolo. Graph-theoretical derivation of brain structural connectivity. *Applied Mathematics and Computation*, 377: 125150, July 2020. ISSN 00963003. doi: 10.1016/j.amc.2020.125150.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.122653799.

H. Markram, E. Muller, S. Ramaswamy, M. W. Reimann, M. Abdellah, C. A. Sanchez, A. Ailamaki, L. Alonso-Nanclares, N. Antille, S. Arsever, G. A. A. Kahou, T. K. Berger, A. Bilgili, N. Buncic, A. Chalimourda, G. Chindemi, J.-D. Courcol, F. Delalondre, V. Delattre, S. Druckmann, R. Dumusc, J. Dynes, S. Eilemann, E. Gal, M. E. Gevaert, J.-P. Ghobril, A. Gidon, J. W. Graham, A. Gupta, V. Haenel, E. Hay, T. Heinis, J. B. Hernando, M. Hines, L. Kanari, D. Keller, J. Kenyon, G. Khazen, Y. Kim, J. G. King, Z. Kisvarday, P. Kumbhar, S. Lasserre, J.-V. Le Bé, B. R. Magalhães, A. Merchán-Pérez, J. Meystre, B. R. Morrice, J. Muller, A. Muñoz-Céspedes, S. Muralidhar, K. Muthurasa, D. Nachbaur, T. H. Newton, M. Nolte, A. Ovcharenko, J. Palacios, L. Pastor, R. Perin, R. Ranjan, I. Riachi, J.-R. Rodríguez, J. L. Riquelme, C. Rössert, K. Sfyrakis, Y. Shi, J. C. Shillcock, G. Silberberg, R. Silva, F. Tauheed, M. Telefont, M. Toledo-Rodriguez, T. Tränkler, W. Van Geit, J. V. Díaz, R. Walker, Y. Wang, S. M. Zaninetta, J. DeFelipe, S. L. Hill, I. Segev, and F. Schürmann. Reconstruction and Simulation of Neocortical Microcircuitry. *Cell*, 163(2):456–492, Oct. 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.09.029.

M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, Sept. 2006. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.74.036104.

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29 (2):173–191, May 2007. ISSN 03788733. doi: 10.1016/j.socnet.2006.08.002.

T. Snijders. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 2, 2002. ISSN 1529-1227.

T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New Specifications for Exponential Random Graph Models. *Sociological Methodology*, 36(1):99–153, 2006. ISSN 1467-9531. doi: 10.1111/j.1467-9531.2006.00176.x.

D. Strauss and M. Ikeda. Pseudolikelihood Estimation for Social Networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990. ISSN 0162-1459. doi: 10.2307/2289546.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998. ISSN 0028-0836, 1476-4687. doi: 10.1038/30918.

J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 314(1165):1–340, Nov. 1986. ISSN 0962-8436. doi: 10.1098/rstb.1986.0056.